

Big Data and Social Web Research Methods

An updated and extended version of the book:
Introduction to Webometrics. Includes chapters to
appear in a forthcoming social web methods book.
This updated book was previously called:
Webometrics and social web research methods

5/23/2018
University of Wolverhampton
Mike Thelwall

Contents

1. Introduction and big data theory [new].....	6
Information-Centred Research: A big data theory	7
Overview	7
2. (L1) Studying networks and links [new].....	8
Introduction	8
Terminology	9
Gathering network data on individuals in social web sites	10
Gathering network data	13
Summary	14
Creating networks from collections of web sites.....	14
Summary and further reading	18
Visualising networks	18
Non-network diagrams to represent networks.....	20
Summary and further reading	22
Measuring networks with Social Network Analysis (SNA).....	23
Metrics for nodes.....	23
Metrics for whole networks.....	24
Summary and recommended reading	25
Final thoughts	26
3. (L2) Link Analysis 27	
Background: Link counts as a type of information	27
Types of webometric link analysis	28
Link impact assessments	28
Alternative sources of link data: Alexa and link databases	30
Alternative link counting methods	30
Case study: Links to ZigZagMag.com	31
Content analysis of links	31
Link relationship mapping.....	32
Case studies	34
Co-link relationship mapping.....	37
Link differences between sectors – An information science application	38
Summary	39
4. (L3) Web Impact Assessment	40
Web impact assessment via web mentions.....	41
Bespoke Web citation indexes.....	43
Content analysis	45
Category choices	45
Sampling methods	46
Example.....	46
Validity.....	47
URL analysis of the spread of results.....	48
Web citation analysis – An information science application	49
Advanced web impact studies	50
Summary	50
5. (L4) Web Network Thick Description [new].....	52
Introduction	52
Web Network Thick Descriptions: A rationale	52
Methods overview	53
Pure WNTD.....	53
Hybrid WNTD.....	54
Technical requirements for a web network thick description project.....	55
Examples of possible WNTD Studies	56
Example 1: Twitter networks	57

Network of pre-existing key topic actors in Twitter.....	57
Topic-based Twitter users network	57
Example: Influential Twitter climate accounts.....	58
Example 2: Web link or mention networks	61
Example: US mathematics schools	62
Summary	63
6. (T1) Analysing comments and sentiment [new]	65
Introduction	65
Gathering comments to analyse.....	66
Units of analysis	66
Basic sampling strategies	66
Sampling topics via keywords.....	67
Storage.....	68
Graphs and simple statistics	69
Summary	71
Content analysis	71
Planning a content analysis	72
Conducting a content analysis	74
Reporting the results.....	74
Summary and recommended reading	75
Grounded theory.....	75
Summary and recommended reading	76
Sentiment analysis.....	77
Summary	79
Final thoughts	79
7. (T2) Blog and Twitter Searching [updated throughout].....	81
Micro/blog search engines [modified].....	81
Date-specific searches	82
Trend detection [modified].....	83
Checking trend detection results	85
Limitations of micro/blog data	86
Advanced blog analysis techniques.....	87
Advanced microblog analysis techniques	88
Summary	88
8. (T3) Web Texts Thick Description [new]	90
Introduction	90
Stage 1: Constructing and testing queries	90
Stage 2: Post-hoc spam, irrelevant content, and duplicate filtering	91
Stage 2a Irrelevant content	91
Stage 2b Spam.....	92
Stage 2c Duplicate filtering.....	92
Stage 3a: Content analysis (all projects).....	93
Stage 3b: Relative word frequency analyses compared to general texts (all except small projects).....	93
Stage 3c: Relative word frequency analyses for collections of texts within the topic (all except small projects).....	94
Stage 3d: Time series for keywords for large collections of texts (large projects collected for at least a week).....	94
Stage 3e: Time series scanning for large collections of texts (large projects collected for at least a week)	95
Stage 3f: Summary and thick description: Integrating the analyses (all projects).....	95
Example 1: UK Tweets about the Egyptian revolution.....	96
Example 2: Autism.....	99
Summary	102

9. (S1) Automatic Search Engine Searches: Webometric Analyst [Updated throughout]	103
Introduction to Webometric Analyst	103
Webometric Analyst Web Impact Reports	104
Web Impact Reports – classic interface example	105
Webometric Analyst Link Impact Reports	107
Link Impact Reports – classic interface example	108
Webometric Analyst for network diagrams.....	108
Rearranging, saving and printing network diagrams	109
Network diagram – classic interface example	110
Co-link network diagrams	111
Webometric Analyst additional features	111
10. (S2) Web Crawling: SocSciBot	112
Web crawlers	112
Overview of SocSciBot	113
Network diagrams of sets of web sites with SocSciBot	113
Other uses for web crawls	118
11. (A1) Search Engines and Data Reliability	119
Search engine architecture.....	119
Duplicate and near-duplicate elimination.....	120
Comparing different search engines.....	121
Research into search engine results	121
Modelling the web’s link structure.....	122
12. (A2) Tracking User Actions Online.....	125
Single site Web analytics and Web server log file analysis	125
Multiple site Web analytics	126
Search engine log file analysis	127
13. (A3) Advanced Techniques	128
Query splitting.....	128
Virtual memetics	129
Web Issue Analysis	130
Data mining social network sites.....	131
Social network analysis and small worlds.....	132
Folksonomy tagging	133
API programming and mashups	133
14. Summary and Future Directions [modified]	135
15. Glossary	135
16. References	136

1. Introduction and big data theory [new]

The web and mobile communication technologies are embedded in the lives and work of a substantial proportion of the world's population. This creates new challenges to understand the continual changes to life and work made possible by each new technology. These changes may take the form of adapting existing practices, such as tweeting instead of phoning to arrange to meet a friend, or of creating new behaviours, such as publically listing friends or acquaintances within social network sites. The web in particular has made all kinds of information easier to access and has encouraged people to make information publically available that they previously would not have thought to do. Both of these are huge advantages for researchers and students because the information needed to study a phenomenon might be freely available online, saving time in the data collection phase of a project and making larger scale studies possible. This book describes ways to exploit this free public information with a focus on two types: text and links. It focuses on big data methods from the information science field of webometrics that do not require an extensive computer science background to carry out.

The simplest way to investigate a web phenomenon is probably to read what the participants have written, look at their pictures and videos, or listen to their recordings, as appropriate. For example, reading the comments on a YouTube video may give insights into the audience reaction to the video or the issues within it. This book describes how to systematically sample such texts as well as formal methods to investigate them, such as content analysis. The book also describes more complex methods to investigate texts, such as with word frequency analysis or time series analysis, some of which require the use of supporting software, but the focus is on using simple methods that can give transparently meaningful results and are not difficult to master.

One thing that the web has given us on an unparalleled big data scale is information about the connections between documents, in the form of hyperlinks, and about the connections between people, such as in the form of lists of Friend or subscriber lists in social web sites. Analyses of these links can help to shed light on patterns of interactions between groups of people, documents or websites in ways that may not be apparent from examining individual web pages. This network analysis truth has long been recognised in the fields of social network analysis and citation analysis but the web has made it possible to translate this theory into practice on a larger scale than before. For instance, it might only be possible to find the most influential people in a social network site by identifying those with the most subscribers or those whose Friend connections span two influential communities. This book describes what kinds of link data can be gathered from the web and a range of standard types of methods that can be used to analyse it.

The purpose of this book is to describe, in general terms, methods that are appropriate for analysing text and links from a wide variety of different kinds of sites and social web services on the web. It uses big data because the sources of data described, from the web itself to Twitter, are huge although the methods described in the book do not need large scale computing resources or computer programming skills. The book should be read in conjunction with standard generic social sciences methods text books, such as those for content analysis and social network analysis. Whilst dedicated textbooks on these methods will give them more comprehensive and wide-ranging treatments, this book fills a gap by discussing issues that are specific to the web and by suggesting combinations of methods that make sense for web analyses. For someone researching the web, it would make sense to start with this book and then supplement it with dedicated methods textbooks for the chosen methods.

Although this book is about the web and gives examples of methods applied to specific websites, the methods described are intended to be generic so that as new sites emerge and old ones fade the methods will still apply in a similar way. In parallel with this, with the exception of two illustrative chapters, this book does not give specific instructions for any computer program to help with an analysis or any website to investigate but such instructions are provided on the web and will be adapted over time as the web itself changes.

One of the aims of the book is to make it as easy as possible for social scientists to conduct interesting and appropriate web research without spending a lot of time mastering complex methods or software. The methods therefore tend to be relatively simple and have high face validity so that the results are easy to interpret. Because of its focus on methods, the book does not introduce specific theories for analysing the web and does not recommend topics to investigate because these depend upon your disciplinary background: you can study what you want and for whatever reasons you want and this book will help you to know how to investigate it.

Information-Centred Research: A big data theory

Sometimes the web produces new sources of data that appear to be useful for research but it is not clear which topics they will be most useful for. The Information-Centred Research (ICR) theory argues that researchers can investigate new information sources to identify which types of research topics it is most suitable for (Thelwall & Wouters, 2005; Thelwall, Wouters, & Fry, 2008). This is in contrast to traditional research direction, which is to start with a topic-specific hypothesis and then assess it using data with the assumption that it is suitable. The ICR approach can be more efficient if the potential uses of a new data source is unclear from initial examinations (see also: <http://cybermetrics.wlv.ac.uk/icr.html>). Some of the techniques in this book can be used for this exploratory goal.

Overview

The chapters are organised into several coherent groups. The first group (L1-L4) focuses on text analysis methods and finishes with a specific recommended method for analysing texts, web text thick description. The second group of chapters (T1-T3) focuses on link analysis methods and finishes with a specific recommended method for analysing sets of links, link network thick description. In both of these chapters there are some overlaps between the methods described and the purpose of this overlap is to reinforce the key methods and issues. The next group of chapters (S1-S2) describes two specific computer programs for gathering and analysing texts or links for web research. The purpose of these chapters is to illustrate how specific programs can be used to help web research rather than to teach their use. Finally, the last group of chapters (A1-A3) gives some additional background information that is relevant to data gathering for web research, such as how search engines work, and describes some advanced web research techniques that may be useful in special cases.

2. (L1) Studying networks and links [new]

- Key example 1: Natural-looking makeup tutorial.
- Key example 2: Coming out videos.

This chapter is about studying networks of things, including people, texts, and websites. This can give holistic insights into phenomena that are only make sense as networks or that are easier to understand in some way when viewed as a network. One key example for this chapter is a popular video from YouTube called *Natural Looking Makeup Tutorial*, which attracts about two million views per year. This is a mundane rather than controversial issue since the goal of the video is to show you “how to achieve a natural simple look with makeup”. Nevertheless, it is interesting precisely because it is relatively mundane and has still become highly popular in the competitive environment of YouTube. The video was produced by an American make-up artist called Michelle Phan in 2007.

Another key example for this chapter is a coming out video from YouTube. Many young people come out by making a YouTube video, perhaps because the online environment seems to be relatively safe. Also, the author of a video has time to think through what they want to say and present themselves in a manner that they are happy with. The people that comment on this video form a network since some of them know each other. Studying the network of commenters can give insights into whether the comments are individual disconnected points or if they reflect a community responding to the video.

Introduction

The network perspective is important for understanding the social web because at its heart are interactions within and between groups of people. For instance, to understand the dynamics of an online political discussion it would not be enough to examine each participant separately because the overall pattern of interactions may also be important. Perhaps one central person interacts with everyone else, or perhaps two opposing groups interact primarily within themselves or against the other group. Examining how a collection of people, organisations or web sites interact with each other as a whole rather than as a collection of individuals is the network perspective. Sometimes, just drawing a picture of the network can immediately reveal useful facts about it. In other words, drawing a network can make new types of pattern visible that were previously hidden. There is also a field, Social Network Analysis (SNA) that has developed measures of aspects of networks that can reveal socially-relevant information about them.

A little bit of background about networks in the social sciences will help to set the context of this chapter. Although networks have been recognised and examined in mathematics for a long time (e.g., graph theory), their widespread recognition in the social sciences stemmed from anthropologists studying small, self-contained communities. To understand how information reaches everyone in a small, isolated tribe, their complete communication network must be known. For instance, if someone drew a network illustrating who talks to whom the network could be expected to reveal the answers to questions like the following: if one person is told some news, will everyone eventually find out, assuming that friends always pass on news to all their friends? Which person should be told some news to make sure everyone gets it quickly? Are any people critical to connect separate otherwise unconnected groups? The above questions could be addressed with interviews but the people concerned might not know the answers even though they would be clear from a network diagram. Figure 2.1 illustrates a small network consisting of just six people, only some of which talk to each other. The network can be seen to be in two disconnected halves and this has implications for information sharing. For instance, if some important information was told to Mike then Farida, Li and Gaynor could be expected to find out this information, either directly from Mike (Farida) or indirectly via Farida (Li and Gaynor). In contrast, Jeevan and Geeta will not find out the information, assuming that they do not have other means to get it, because their sub-network has no connections to Mike’s sub-network.

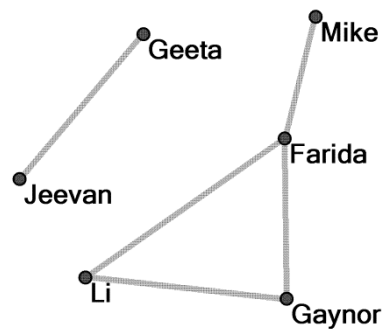


Figure 2.1. A small social network in which a line between two people represents that they talk and share information.

The network perspective can also be used to study large groups. Perhaps most famous is Stanley Milgram's six degrees of separation social science network experiment in which he claimed that any two humans selected at random would probably find that they were connected to each other via at most six connections such that everyone in the chain was an acquaintance of the person before or after them. For groups that are too large to draw networks of, it can still be useful to find out properties of the network as a whole or individual people in the network. Perhaps the most useful whole network property is the number of different disconnected sub-networks it splits into. For instance Figure 2.1 splits into two sub-networks. Similarly, the most important property of individuals in a network is the number of connections that they have. For example, in Figure 2.1, Mike has one connection but Farida has three. Sometimes the most important people in a network are those with the most connections, so this is useful information to know.

This chapter describes methods to gather data on and analyse networks of individuals in the social web or networks of social web sites (e.g., blogs) in the following main subsections:

- Gathering network data on individuals in social web sites.
- Creating networks from collections of social web sites.
- Visualising networks.
- Measuring networks with Social Network Analysis (SNA).

Terminology

Some special terminology is often used when discussing networks and network diagrams. The *nodes* in a network are the entities, such as people, organisations, web pages or web sites. The connections between nodes are sometimes called *edges* if they have no particular direction. In contrast, the connections between nodes are known as *arrows* if they have a direction. There are two types of networks: *directed networks* contain nodes and arrows and *undirected networks* contain nodes and edges.

An example of a directed network is a set of blogs where the nodes are the individual blogs and there is an arrow from blog A to blog B if there is a hyperlink from A to B. An example of an undirected network is a set of members of a particular SNS where the nodes are the individual people and there is an edge between two people if they are Friends in the network. This is an undirected relationship (i.e., edges) because it is bidirectional and equal. Figure 2.1a illustrates both types of network.

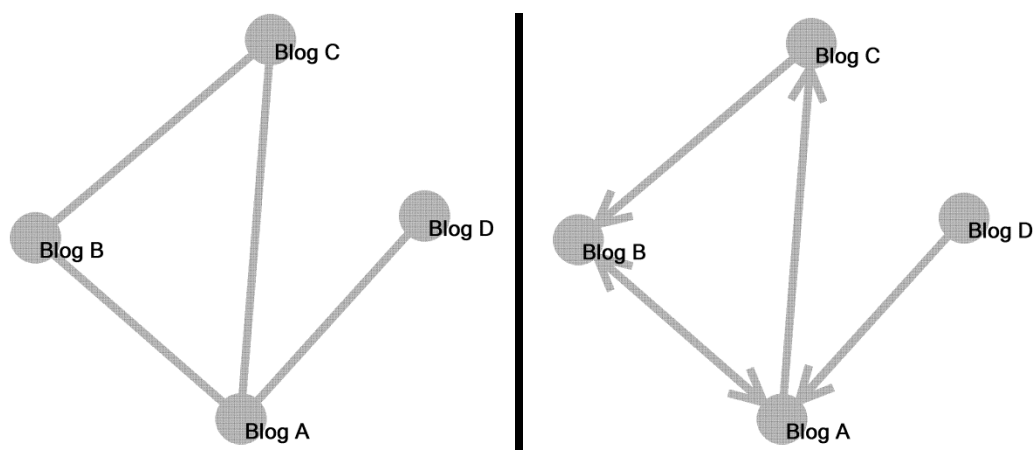


Figure 2.1a. The network on the left is an undirected network. The nodes in the network represent individual blogs and the arrows between the nodes represent the hyperlinks between the blogs. For example, it can be seen that Blog A links to Blog C but Blog C does not link to Blog D. The network on the right is a directed network. The nodes again represent blogs but the edges between the blogs represent the bloggers being offline acquaintances or friends – a bidirectional relationship.

Gathering network data on individuals in social web sites

The section “Choosing a sample of users for quantitative data gathering” gives useful information on sampling individuals in social web sites but the situation is different when gathering information on networks. The most important difference is that it is not possible to apply the network perspective to a random sample of members of a social web site. This is because network statistics only make sense if the network data is complete and this isn’t possible with a random sample. Hence, when studying a network of individuals in a social web site it is important to pick a collection that is manageable in the sense that information can be gathered on all members of the network. For example, the network might be one of the following.

- All people that comment on a given YouTube video.
- All of the registered Friends of a member of a social web site.
- All members of a particular group in a social web site.

An important limitation is that the group should be small enough to be manageable – and this normally means a maximum of 50 members. It is difficult to draw understandable networks for more than about 50 people and network data often has to be collected manually, which may take too long for more than 50 people, especially since the data will need to include connections between the members of the network.

Once the group has been identified then the people are the nodes in the network to be analysed. The second stage is to decide upon the type of connection between people that will be used for the network. A connection is any type of relationship between two people. This would normally be a simple and obvious relationship, such as a Friend connection. It is also possible and sometimes useful to define a more indirect relationship based upon a shared property or activity. For example, two people could be defined as connected in a network if they had both commented on the same blog post or video or if they had at least one Friend in common. This kind of indirect network connection may be useful to study if it is relevant to the research question. If the research question relates to multiple modes of connections between people then it can also be useful to create multiple networks, each with the same nodes (people) but with different definitions for the connections.

Example The figures below are networks built from the commenters on a popular YouTube coming out video (top) and on a newer and less visited YouTube coming out video (bottom).

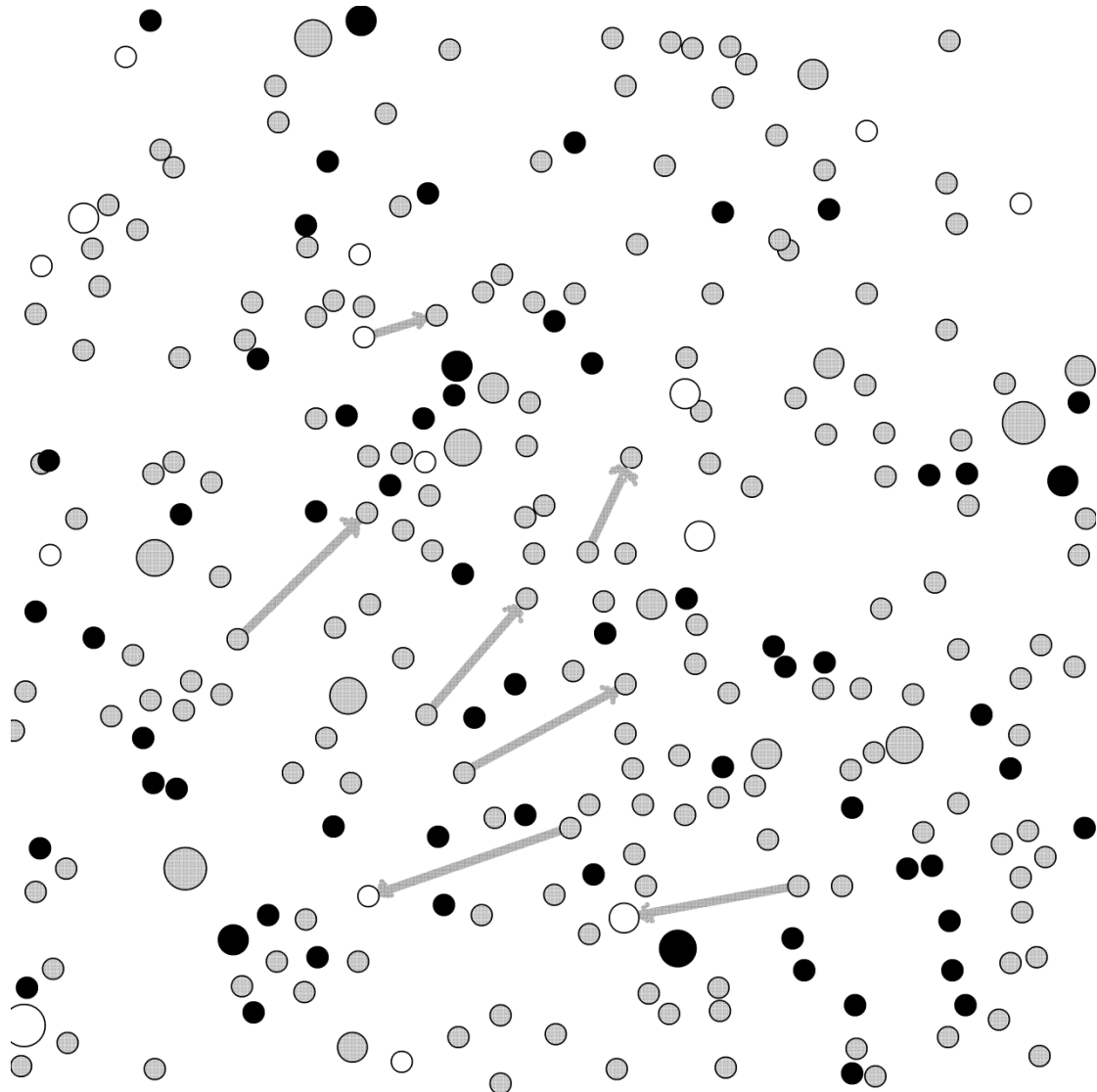


Figure 2.2a. This is a network of the YouTube subscriber connections between commenters on a single YouTube coming out video. Circles represent commenters, with their sizes being proportional to the number of comments made. Black circles are commenters declaring a male gender and grey circles are commenters declaring a female gender. Lines between commenters indicate a subscriber connection. The diagram shows that subscriber connections are very sparse in this set of comments, suggesting that either the commenters are generally unknown to each other or that they choose not to subscribe to each other in YouTube.

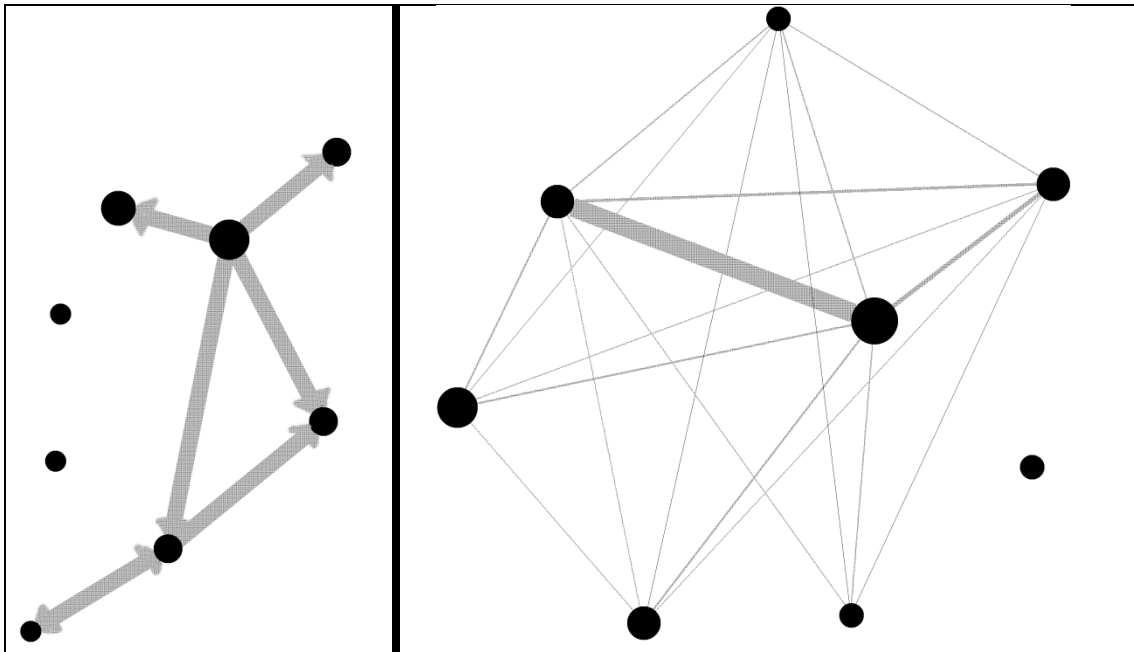


Figure 2.2b. There are two networks above for the same *coming out* video, one with just a few commenters. The network on the left is for subscriptions. On this video, however, most of the commenters subscribe to at least one other commenter and so this network is qualitatively different from the one above. It may be that these are people that know the video author offline and are part of his circle of friends. The network on the right is the same as the one on the left except that a line is drawn between two nodes that have at least one *common subscription connection* in YouTube. There are many more of these indirect connections, producing a much denser network. The subscriptions in common may be a popular channel that both have subscribed to and so the lines may reflect a shared interest or opinion rather than having an acquaintance in common. From this perspective, only one person is unconnected to the rest. This could be a random visitor or someone that does not use the YouTube subscriptions. In the diagram on the right, line widths are proportional to the number of subscriptions in common and it is clear that two of the commenters have many more subscriptions in common than do the others.

The diagrams in this section illustrate three different kinds of social web connection that can be used to build networks: Subscription, subscriptions in common, and replying to someone. Subscriptions in common is a symmetric relationship and so the networks drawn do not have arrows on the lines. In contrast, the reply network is asymmetric: if a person replies to a comment by another then the other does not necessarily reply back. Subscribing is also asymmetrical: if a person subscribes to another person's YouTube channel then the subscribed person will not necessarily reciprocate. Many other relationships could also be used to form networks in the social web outside YouTube, including the following.

- Membership of the same group.
- Hyperlinking between profile pages or blogs.
- Commenting on another person's SNS profile page, image or video.

Example The figure below is a small part of a network created by comments on the *Natural Makeup Tutorial* YouTube video. In this network, each circle (node) represents a person that commented on the video, with the size of the circle being larger for people that made more comments. Arrows between circles indicate when one person replied to the comment of another person. Numbers near circles represent commenters' ages, if given. Darker nodes are for males and lighter nodes are for females. White nodes are for commenters with undeclared

genders. The whole network is too large to show but it has 180 circles (i.e., commenters) and most of them pair up with just one other commenter. Hence, most “conversations” about this video are dialogs. Although it would be time consuming to spot by reading through all the comments themselves, there are a few larger dialogs that include several people, such as the one shown. Interestingly, the three largest dialogs, like the one shown, are based around a male commenter making derogatory or offensive comments about the video or the video producer and triggering a reaction from the predominantly female commenters.

Note that the network here is not a friendship network or even an acquaintanceship network (only two of the 180 commenters were registered Friends in YouTube) and so it is only a network of exchanging comments on this particular video.

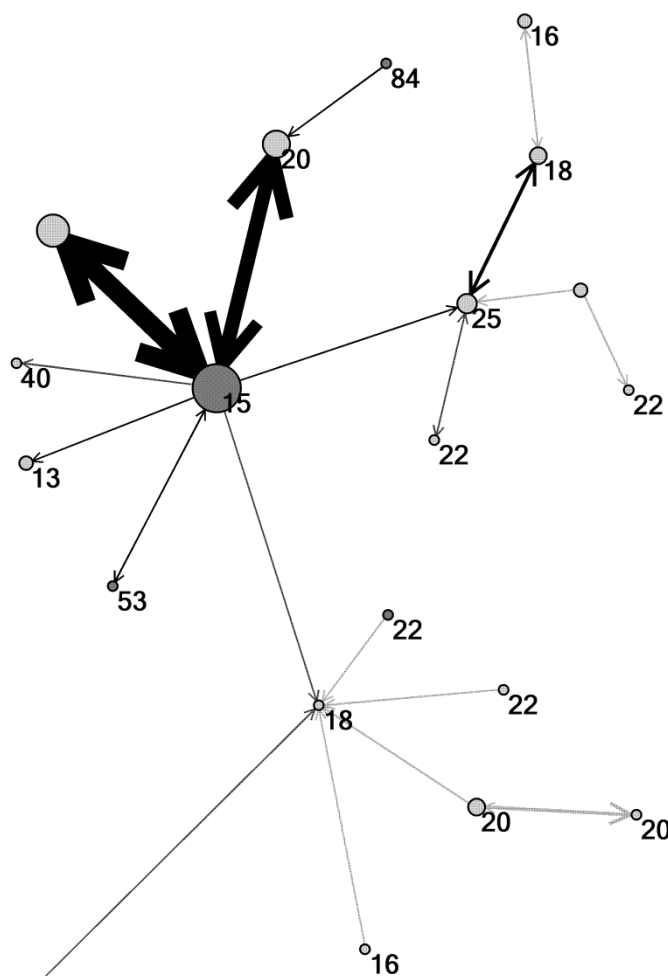


Figure 2.3. This YouTube reply network is a section of the network of commenters that reply to other comments on the YouTube *Natural Looking Makeup Tutorial*. The section shows part of the largest connected sub-network within the whole network. Nodes represent commenters and their sizes are proportional to the number of comments posted to the video. Other node properties are user-declared: the three darker nodes are males (aged 15, 84 and 22 here) and the lighter nodes female. Numbers attached to nodes are declared ages. Darker arrows indicate predominantly negative sentiment expressed in the comments, and arrow widths are proportional to the number of replies between the commenters.

Gathering network data

Once the sample of people and the type of network connection have been selected, then the network data needs to be collected. Often this has to be done manually and this can be done in the following way.

- Create a new spreadsheet.
- Label the first row *and* the first column of the spreadsheet with the names of the network members (nodes).
- Record a “1” in the appropriate row and column of the spreadsheet if person A is connected to person B. Find out whether the connection exists in each case by examining the connection in the social site.
- Copy the spreadsheet data to the software used to run the analysis or draw the network diagram manually or using a drawing program.

If there are a lot of people but only a few connections in the network then instead of the table approach described above, it may be more efficient to list the pairs of people that are connected in the network. This can be done in a spreadsheet by using two columns, one for the first person in each connection and one for the second person.

In some cases the data can be collected automatically instead. Please see the book web site for more information about this. For example, at the time of writing this book, YouTube commenter networks could be automatically constructed with the *Webometric Analyst* software and Facebook Friend networks for a person’s own Friends could be connected by a program written by Bernie Hogan.

Summary

- Networks should not be too large – a maximum of 50 members is a practical limit.
- Networks should be complete sets of members of a particular group or people with a particular property and not random samples of larger networks.
- Networks can be collected automatically from some social web sources but must be manually collected from others.
- The connections between people (or nodes) in a network do not have to represent friendship but can also reflect other things, including hyperlinks, replies to messages, and having friends in common.
- See the book web site for information about gathering network data automatically.

Creating networks from collections of web sites

It sometimes makes sense to study collections of web sites rather than collections of individuals within a single social web site. Here are some examples.

- *Blogs* A project might investigate a set of blogs that focus on a particular topic or issue, such as politics within a country, transgender life, slow food, marathon running or makeup styles. The blogs here are all web sites and they can form a network with the blogs being the nodes and their hyperlinks forming the connections between them.
- *Organisational web sites or pages* A collection of web sites or pages of organisations that are related to a particular theme may form an interesting project. Although organisational web sites and the links between them is a Web 1.0 type of topic, the web sites may embed social web components, such as blogs, or functionality, such as Facebook like buttons. It may also be that some of the web sites are simply pages within a social network site, such as a group page in Facebook.

Example One project studied e-science organisations in South Korea via links between their web sites. After collecting and analysing the link network created the project concluded, “The findings suggest that some e-science terms, including cyberinfrastructure, have prominent Korean web presences. At the same time, Korean government-sponsored national e-science centers and their affiliates do not have a strong web presence and do not actively participate in the hyperlink network that connects e-science-related institutional websites. Instead, they have a closed network among themselves. This result stems from the institutional dynamics within Korea’s public and private e-science research communities” (Park, 2010).

A project studying a collection of web pages or web sites from the network perspective will need to systematically find out all the connections between them. These connections can be

found manually in some cases and automatically in others. The first decision that such a project must take concerns the kinds of connections to include. Although hyperlinks are often used, the decision should be based upon the research objectives and the options, if any, for types of links to collect. For instance, suppose that twenty prominent technology bloggers were being analysed. Blogroll links between them would probably indicate endorsement because endorsement is the function of blogrolls. In contrast, links within individual blog posts would suggest one blogger reading and commenting upon another's post. Such links could represent connected content rather than endorsement. A connection like this could be part of a debate and hence would not be an endorsement. A study about relationships between bloggers might therefore only count blogroll links whereas a study of the flow of content in the blogosphere might count only blog post links. In contrast, a study of the awareness of the bloggers of each other might use both types of links.

Example A classic blog link analysis project is a study of US politicians' blogs in the 2004 election. Based upon the links between bloggers, they found a clear divide between the conservative and liberal bloggers, with each set tending to link far more to blogs with the same political orientation than to blogs with an opposing orientation. This is an important finding because it suggests that these bloggers are not primarily using their blogs to engage in discussions with bloggers with opposing views. The authors also found "differences in the behavior of liberal and conservative blogs, with conservative blogs linking to each other more frequently and in a denser pattern" (Adamic & Glance, 2005).

The simplest but often most time-consuming way to gather data is to manually check all relevant web pages to identify appropriate links, recording the source and target of each link found for later processing. This is only practical if there are not too many web pages to be checked – perhaps not more than a few hundred. Hence it will not work for collections of web sites unless the web sites all only have a few pages.

The process of manually identifying links can be difficult because links can display anchor text that does not name the target web site. For instance, "Click [here](#) for a related story" would be unhelpful. To identify links in such cases, a simple trick is to move the mouse pointer over the link and check the browser status bar at the bottom of the screen for the URL of the link. All links with an unclear target should be checked using this or an equivalent technique. Specific URLs in links can also be searched for by using a browser's "view source" menu option to see the underlying code of a page (i.e., its HTML tags) and then by using the browser's "Find" function to search for the URL within this code.

Hyperlinks can also be automatically identified by web crawlers. A web crawler is a computer program that can be pointed at a collection of web pages or sites and can identify the hyperlinks within them. If a web crawler can be used then this can save a lot of manual work identifying the hyperlinks but it also requires some time to learn how to use the crawler. Academic web crawlers typically cannot cope with large web sites with hundreds of thousands of pages but are suitable for collections of small web sites. If a project studies only large web sites then a compromise would be to use a web crawler but to restrict the number of pages per site that the crawler fetches.

It is possible for small web sites with fewer than 50,000 pages using the free academic web crawler SocSciBot, which is designed for social science link analyses. This program can be fed with a list of home pages of the web sites of interest and then it will attempt to find and download all the web pages in all of the web sites. SocSciBot also identifies all hyperlinks between the web sites in the list and even creates a network data set from this information (for instructions, see <http://socscibot.wlv.ac.uk>). A disadvantage of using a web crawler is that it is impossible to specify which kinds of links should be studied; the only option is to count all links. A second disadvantage is that the web crawler might not be able to find all pages in all web sites. It is likely to have particular problems with complex web sites using a lot of Java, Flash or JavaScript. As a result, web sites found not hosting any links should be double checked manually. In addition, the above limitations should be discussed when presenting the results.

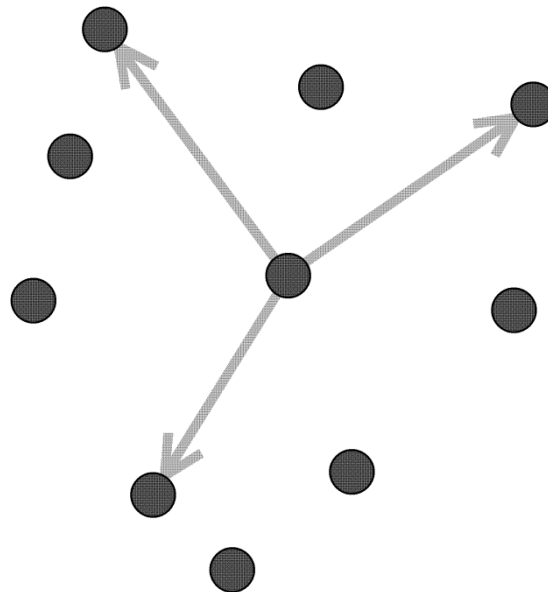


Figure 2.4. This network was constructed using a web crawler to crawl the web sites of 10 UK transgender blogs. The web crawler automatically identified the links between the blogs and this information was used to construct the network diagram below. The names of the blogs have been removed for privacy reasons. Only one blog in the set linked to any other blogs in the set and this linked to three of the other blogs.

The link networks produced by SocSciBot can be graphed with SocSciBot's network diagram functions and can be exported into other programs for social network analysis calculations (see book web site and the social network analysis section for more information).

Advanced types of web site network connections Some studies use types of links that are not hyperlinks because it is easier to gather data on them. Two examples are *title mentions* and *URL citations*. A title mention is a mention of the name or title of an organisation, person, or object in a web page. Title mentions can form implicit inter-document connections. For instance, if a page in the BBC web site mentions then this is an implicit connection with the CNN web site. Hence a study of connections between organisations or web sites could use title mentions to identify such connections instead of hyperlinks.

This is an advantage because title mentions between organisations can be found using some free web searching software much more quickly than with a web crawler and can be found even in web sites that are too large to use a crawler on. The reason is that title mentions can be found with web searches in commercial search engines and at the time that this book went to press these searches could be submitted automatically to search engines. A disadvantage of title mentions is that they can work poorly for ambiguous names. To find whether Web site A contains any title mentions of web site B the following query can be submitted to a major search engine.

```
site:A "[name of web site B]"
```

For example, to search the BBC web site for mentions of high profile blogger Belle de Jour, the following query could be submitted. In this case, the results would have to be manually filtered to remove mentions of the film with the same name.

```
site:www.bbc.co.uk "Belle de Jour"
```

An odd alternative to both title mentions and hyperlinks is the URL citation. This is the mention of the URL of one web site in a web page from another web site. URL citations can also be calculated using search engine queries, and are recommended if title mentions do not work – for example if they are too ambiguous. To find whether Web site A contains URL citations of web site B the following query can be submitted to a major search engine.

```
site:A "[URL of web site B omitting the initial http://]"
```


For example, to search The Independent newspaper web site for mentions of high profile blogger Belle de Jour, the following query could be submitted. In this case, the results would not have to be manually filtered. The disadvantage of URL citation searches is that it is relatively rare for URL citations to be used instead of links and so this method is likely to miss most connections between web sites.

```
site:www.independent.co.uk "belledejour-uk.blogspot.com"
```

Belle de Jour: Diary of a London Call Girl. Read more at Belledejour-uk.blogspot.com

Figure 2.5. This extract from a web page shows a title mention of Belle de Jour (at the start) and an URL citation for her blog (at the end).

Another advanced type of inter-site connection is the *co-inlink*. This is an indirect connection between two web sites that can exist even if the two sites are not directly connected. A co-inlink is created between two web sites B and C whenever a third web site A links to both of them, as shown in Figure 2.6. A less commonly used type of link is the co-outlink, which is also shown in Figure 2.6. If a pair of web sites has a lot of co-inlinks then this means that many other web sites link to both. A common cause of this happening is that the two web sites are similar in some way, for example because they cover the same topic or because the site owners are geographically close together. Network diagrams built from co-inlinks are normally denser than diagrams built from links and can reveal different structures in the set of web sites investigated. In a co-inlink diagram, only the co-inlinks would be shown and not the direct links. Hence, a co-inlink network diagram for the six nodes in Figure 2.6 would show only one connection: an edge between B and C to represent the fact that they are co-inlinked.

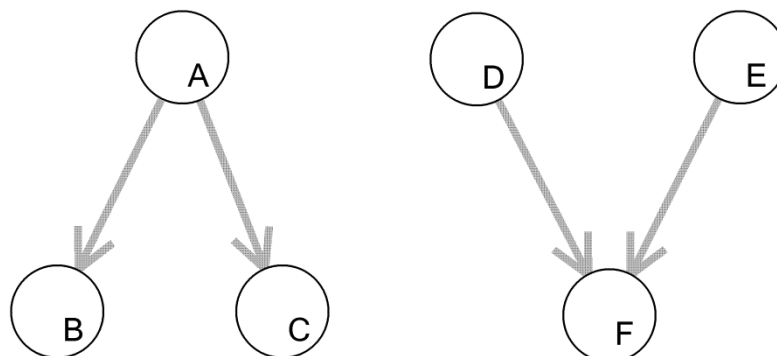


Figure 2.6. In this network B and C are co-inlinked by A. Also, D and E co-outlink to F.

Co-inlink networks can be searched for directly using search engines or can be obtained using appropriate software, as described in the book web site. Below are searches for co-inlinks based upon URL citations and title mentions.

To calculate title mention co-inlinks between web sites B and C, the following query can be entered in a search engine.

```
"[name of web site B]" "[name of web site C]" -site:B -site:C
```

For example, to search for title mention co-inlinks between CNN and the BBC the query could be the following. It returns a list of pages from outside the BBC and CNN web sites that mention both the BBC and CNN.

```
"BBC" "CNN" -site:www.bbc.co.uk -site:www.cnn.com
```

Similarly, to search for URL citation co-inlinks between CNN and the BBC the query could be the following. It returns a list of pages from outside the BBC and CNN web sites that mention the web sites of both the BBC and CNN.

```
"www.bbc.co.uk" "www.cnn.com" -site:www.bbc.co.uk
-site:www.cnn.com
```

A very important extra step is to test and evaluate the search engine queries to make sure that they are giving you what you think they should be. This involves the following stages:

1. Identification of query syntax and formulation of queries (e.g., as above). For this, the URL of the search engine help page listing the syntax should be reported. In a few cases the syntax is not reported by search engines and so another source should be reported instead (e.g. a recent journal article or blog post).
2. Description of queries used. This includes both the queries themselves and the types of web page that the queries are intended to match (e.g., the query site: wlv.ac.uk matches all pages with domain names ending in wlv.ac.uk).
3. Testing the queries. This involves taking a sample of the queries (e.g., 10) and running them in the search engine and checking that the results are correct in the sense described in (2). A sample of 10 results should be enough for the check. If many results are incorrect then the query syntax is incorrect and must be fixed. For example, if you think that the query `linkdomain:wlv.ac.uk` matches pages containing a link to the University of Wolverhampton website then the above process will show you that it does not.
4. Repeat 1-3 above for every search engine that you use because they do not follow the same rules.
5. Write up the results of 1-4 above for any student projects because this is an important part of your work. For a research paper, a few sentences summarising the results are enough.

Summary and further reading

- Analysing collections of blogs, web sites or web pages around a theme can reveal the patterns of connections between them and give insights into online communication around the topic investigated.
- Hyperlinks are normally used to identify connections between web sites in a network, but there are alternatives that are suitable for larger web sites.
- Link data can be gathered manually and recorded in a spreadsheet or document, or automatically with a web crawler like *SocSciBot*.

For more information, see:

- The book web site for information about other ways of gathering web link data automatically.
- Rogers, R. (2006). *Information politics on the Web*. Massachusetts: MIT Press.
- Thelwall, M. (2009). *Introduction to webometrics: Quantitative web research for the social sciences*. New York: Morgan & Claypool.

Visualising networks

A simple diagram of a network can be very informative. It can reveal basic information at a glance, such as which nodes have the most connections, the density of the connections between the nodes, which nodes are connected and whether the network is a single connected whole or whether it splits up into disconnected sub-networks. Once information has been collected about the nodes and connections within a network, it can be drawn manually or entered into a program that can draw networks. At the time of writing, three such programs were Pajek, UCINET and NodeXL, although programs that collect network data can also sometimes directly visualise the networks.

Network diagrams are drawn by *Pajek*, *UCINET*, *NodeXL* and other similar programs in different ways because there is no single best way to draw a network. Perhaps the key issue is the overall layout of the network: i.e., where to position all the nodes. Some programs have the default setting to arrange the nodes in a circle and then draw the arrows or edges between them. This circular arrangement can obscure patterns in the connections, however, such as groups of interconnecting nodes. It is usually better to ask the program to automatically position the nodes so that interconnected nodes tend to be close together. Two common network layout algorithms that do this are Fruchterman-Reingold and Kamada-Kawai, which are built into some network drawing programs. There is not one best algorithm to use and so

it is reasonable to try different ones out and select the one that gives the clearest result. This is likely to be the diagram that best reveals patterns in the data. It is acceptable to experiment with different network layouts because all different layouts contain the same information – the nodes and the connections between them – and the purpose of the arrangement of the nodes is just to help the viewer to identify patterns.

Once the network diagram has been produced then a visual inspection should reveal some basic information such as: which nodes are in the centre of the network and which are at the edge; whether the network splits into disconnected components (i.e., groups of nodes that do not interlink); whether there are “disconnected” nodes that do not connect to any other nodes; and whether there are any important “bridging” nodes that help to connect different clusters of nodes. These facets can also be identified using metrics, as discussed below, but it is helpful and more intuitive to see them in a diagram. Network diagrams work best if the nodes are clearly labelled and colour coded by type, if relevant. This is impossible for networks with more than about 50 nodes because the diagram will be too cluttered. For such large networks a network diagram is much less useful but may still reveal overall patterns such as how disconnected the network is. Finally, note that the exact position of any node in a drawn network is irrelevant; the significant information is the arrows or edge between nodes.

When producing a network diagram, it is often useful to code additional information into the nodes of the network. For example, the sizes of the nodes can be chosen to be proportional to a property of a web site, such as the number of pages in it or the number of links to or from it. The colours and/or the shapes of the nodes can be used to represent a category that is relevant to the nodes, such as the gender or location of the owner or the topic of the node. For example, left wing blog nodes might be coloured red and right wing blog nodes coloured blue. Alternatively, in a network of SNS members, black nodes might represent males and grey nodes might represent females. Using codes in this way allows the network to contain extra visual information and might help to spot patterns in the network relating to the categories used. Labelling each node with the name of the person or web site that it represents will also help interpretation, but this can produce very cluttered diagrams and so it is not always possible or helpful.

Networks with connection strengths Many networks have connections that are “binary” in the sense that a connection either exists or it doesn’t. An example is the social network site Friend connection: either two people are Friends or not, and in many SNSs there is no such thing as the strength of a Friendship. In contrast, some connections have a natural strength value. An example is a network of blogs where the arrows between the nodes correspond to hyperlinks. Since there may be multiple hyperlinks between the same pair of blogs, each arrow could be associated with a strength value: the number of hyperlinks.

Connection strength can be represented in a network diagram by the width of the arrow or edge that represents it. For instance, the width of an arrow from node A to node B may be set to be proportional to the number of hyperlinks from blog A to blog B. Figure 2.3 is an example of a network with varying connection strengths. In most of the other diagrams in this chapter, all the connections have the same width. If connection widths are set to be proportional to connection strengths then connections with too weak a strength might not be visible at all if the arrow was too thin.

An alternative way of dealing with connection strength is to choose a cut-off value so that all connections that are weaker than the cut-off are not drawn. The remaining connections could be drawn with equal width or with widths varying according to connection strength. This method is particularly useful if there are many connections in the network because it allows the visualisation to illustrate only the strongest connections. Any patterns in the strong connections should be more easily visible if the weaker ones are removed. Unfortunately, it seems that there is not a best method to calculate the cut-off point and so it is hard to defend whatever choice is made. A reasonable approach is to pick a round number (e.g. 10,100) as the cut-off point and to choose this round number so that the patterns in the network diagram are as clear as possible. Although fewer connections make a diagram less cluttered, if too

many connections are removed then patterns may disappear and so a balance must be struck in the choice of cut-off point.

Non-network diagrams to represent networks

This section describes a method for creating a picture of a network without drawing the network connections. The method, known as multi-dimensional scaling (MDS), is a statistical technique that works by positioning the network nodes so that interconnected nodes tend to be close together. The resulting picture can show trends and clusters more clearly than a network diagram in some cases. The method is particularly useful when network diagrams are ineffective because there are too many nodes (e.g., more than 50) or if there are too many connections so that the network diagram is cluttered. Nevertheless, MDS maps are also difficult to read if there are too many nodes on them because of the practical difficulty of finding space in the diagram to label the nodes clearly. MDS relies on connections between the nodes and works best when there are a lot of these connections and these connections have strengths. It often works better for co-inlink connections than for direct link connections because there are more of the former and they have strength values. MDS does not work on nodes that have no connections, so these should be taken out of a network before it is loaded into an MDS program.

Although MDS is a statistical technique, it is not necessary to understand how it works in order to use it. In general terms the algorithm reads in the network data and then searches for coordinates for the nodes so that when the nodes are plotted on a diagram with these coordinates then interconnected nodes tend to be close together and non-connected nodes tend to be far apart. Because of the nature of the algorithm, there is no guarantee that the closest nodes are the most connected but nodes should *tend* to be connected more frequently to close nodes than to distant nodes in the diagram.

Making MDS diagrams An MDS diagram can be produced with statistical software such as SPSS and R, or with the social network analysis programs like UCINET. First, the network data must be fed into the chosen program and then the MDS option can be located and started. There are many different variants of MDS but the main decision is whether to use “metric” or “non-metric” MDS. For a weighted network metric MDS should normally be selected (indicating that the numbers should be related to distances between the nodes). For a binary network non-metric MDS should be selected (indicating that the 1s and 0s are categories). In either case the network data should be registered as a matrix of similarities rather than dissimilarities. This is because higher numbers indicate more “similar” nodes (i.e., nodes that should be positioned closer together). Most programs have an option to draw the MDS picture produced and this should be used together with the option to label the nodes in the picture. Some MDS programs do not draw a picture but only give a set of coordinates for the nodes to be used to draw the picture in another program. Figure 2.7 shows an MDS diagram created by SPSS. Compare this to the network diagram below it created with the same data. The network diagram is significantly more informative but MDS diagrams can be helpful when there are too many nodes or connections for a network diagram to be viable.

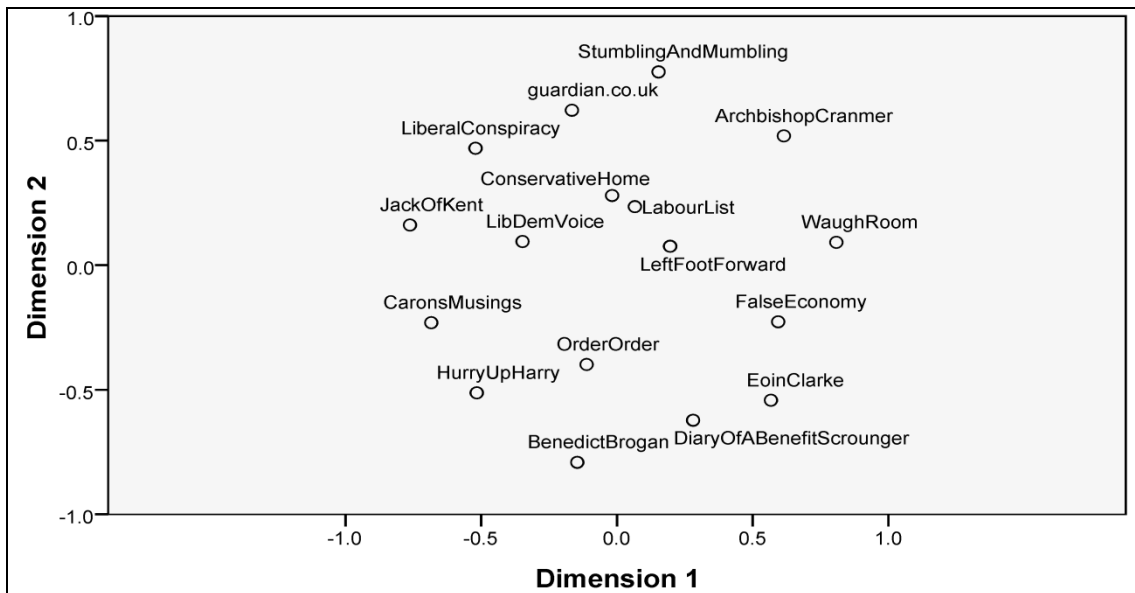


Figure 2.7 An MDS diagram of a selection of UK political bloggers based upon links between them.



Figure 2.8. A network diagram of a selection of UK political bloggers based upon links between them. This diagram uses the same data as Figure 2.7. The network diagram is clearer for this data since the connections can be seen but an MDS diagram can be clearer for large networks with too many connections to be individually distinguished.

Interpreting an MDS diagram means looking for a pattern in it. There are two types of pattern: clusters and trends.

Clusters are clear groupings of nodes close together in the picture. This indicates that the nodes within a cluster tend to be connected to each other more than to nodes outside that cluster. Once a cluster has been identified it needs to be interpreted. This means finding a characteristic that is common to almost all nodes in the cluster and is rare for nodes outside the cluster. For example, if the nodes are blogs then this characteristic might be blog topic, blogger affiliation or blogger geographic location. The common characteristic can be used to label the cluster on the diagram and as a basis for a discussion of the results. Clusters are not always clear-cut and can be an odd shape or partly overlapping with other clusters and so some flexibility is needed when identifying them. In any given MDS diagram there may be no clusters, one cluster or more than one cluster. It is also likely that some nodes do not fit into any of the clusters identified.

The other pattern to look for is a *trend*. This is a tendency for a characteristic of the nodes to be reflected in their overall position in the diagram. For instance, in a network of Friend connections for people within the UK, the more northerly people might tend to be on the left whereas the more southerly people might tend to be on the right. Alternatively, the older people might tend to be towards the bottom and the younger people towards the top of the diagram. Essentially, a trend is any systematic tendency for a node property to vary in a line across the MDS diagram, whether horizontally, vertically or diagonally. Unless the trend is associated with a theory-driven hypothesis, research question or objective then looking for it is a trial-and-error process of testing different properties.

MDS stress values In addition to the MDS picture or node coordinates, a program carrying out MDS will also calculate a *stress* value. This is a single number representing the extent to which the MDS diagram accurately reflects the input network data. A rule of thumb for the stress value (developed in 1964 by Kruskal) a value of 0.20 indicates a poor fit, 0.05 indicates a good fit and 0.00 is a perfect fit (Bartholomew, Steele, Moustaki, & Galbraith, 2008, p. 65). In practice, a stress value of 0.15 or higher is often used to indicate that an MDS diagram is too poor to be useful.

Summary and further reading

- Network maps and MDS diagrams are most useful for exploring networks rather than for testing theory-driven hypotheses. They may reveal patterns in the network and give an overview of the network structure, including the presence of well-connected nodes or separate groups of connected nodes.
- Node size, colour and shape can be used to encode additional information into the diagram that may help to reveal patterns.
- Software can help to produce network diagrams.
- MDS maps position network nodes on a diagram so that interconnecting nodes tend to be close together. They are most suitable when there are a lot of connections between the nodes.
- MDS maps can be examined to identify clusters of nodes or overall trends in nodes. Both may be related to a particular property of the nodes.

For more information, see:

- Nooy, W. d., Mrvar, A., & Batagelj, V. (2005). *Exploratory social network analysis with Pajek*. Cambridge: Cambridge University Press.
- Bartholomew, D. J., Steele, F., Moustaki, I., & Galbraith, J. I. (2008). *Analysis of multivariate social science data (2 ed.)*. Boca Raton, FL: Chapman & Hall/CRC. (Chapter 3 on MDS).
- Hansen, D., Shneiderman, B., & Smith, M. A. (2010). *Analyzing social media networks with NodeXL: Insights from a connected world*. San Francisco, CA: Morgan Kaufmann.

At the time of writing, network drawing programs included: Issue Crawler, SocSciBot and Webometric Analyst (for hyperlink data for collections of web sites); and Gephi, NodeXL, Pajek, and UCINET for any type of network.

Measuring networks with Social Network Analysis (SNA)

Social Network Analysis (SNA) is both a research field and a set of primarily quantitative methods for analysing networks. As discussed in the introduction, SNA grew in recognition that some social phenomena, such as the transmission of information by word of mouth, could only be understood by looking at whole networks. SNA methods have also been used to study the transmission of diseases, the flow of information on the internet and many other things. At the heart of SNA is a collection of metrics for networks. These metrics are formulae or techniques to measure some aspect of a network or its nodes.

The section on visualising networks is important because network visualisations can reveal patterns and can give an overview of a network. The patterns identified visually can sometimes be checked and quantified with SNA metrics, and SNA metrics can also sometimes identify information and patterns that are not evident from a visual inspection of a network. Many of the most common SNA metrics are quite simple and can be identified quite easily from a network diagram. Recall from the above section that a network consists of a set of nodes (e.g., web sites or people) and the connections between the nodes (e.g., hyperlinks or friendship). These connections are called arrows if they have a direction and are called edges if they are undirected. The connections can be binary (i.e., they exist or they don't) or weighted (i.e., each connection has a number associated with it that indicates its strength in some sense).

Metrics for nodes

SNA includes hundreds of metrics for networks but this section introduces a few of the most useful ones. Although each metric is calculated by a formula or other method, SNA software is used to conduct the calculations in practice. It is important to understand what the results mean, however. This subsection discusses metrics for individual nodes and the next subsection discusses whole network metrics.

Probably the most used metric is *degree centrality*. For an undirected network (i.e., edges rather than arrows between nodes) the degree centrality of a node is the number of other nodes that are connected to it by edges. In Figure 2.9, for example, the degree centralities of the nodes are: A 2, B 2, C 4, D 2, E 1, F 3. Nodes with the highest degree centrality are the most connected and are often the most important. A common simple method to gain useful information with the degree centrality metric from a large network is to list the 10 or 20 nodes with the highest degree centrality and comment on why they are important. Nodes with degree centrality zero are disconnected from the rest of the network and so this metric can also be used to identify disconnected nodes.

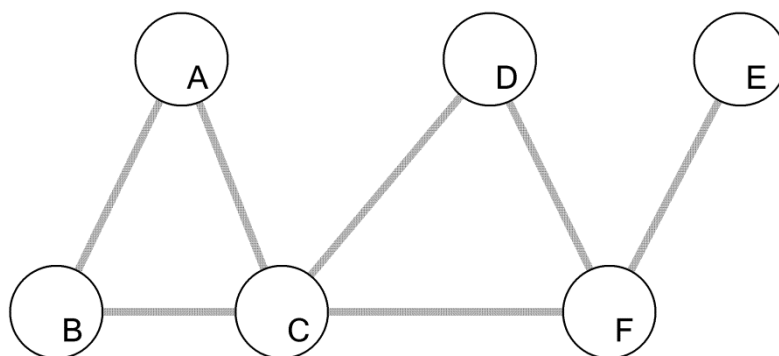


Figure 2.9. This is an undirected network of six nodes and seven edges between them.

Directed networks have two types of degree centrality. The *indegree centrality* of a node is the number of arrows pointing to it. Similarly, the *outdegree centrality* of a node is the

number of arrows originating from it. In Figure 2.10, for example, the indegree centralities of the nodes are: A 0, B 1, C 3, D 0, E 0, F 2. The outdegree centralities of these nodes are: A 2, B 0, C 0, D 2, E 1, F 1. From the perspective of indegree centrality, node C is the most important since it has the highest value. In contrast, the joint most important nodes from the perspective of outdegree centrality are A and D. In a directed network the isolated or disconnected nodes are those that have an indegree centrality of 0 and an outdegree centrality of 0.

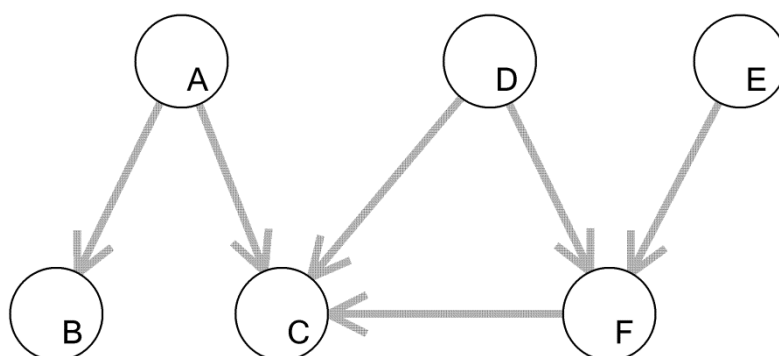


Figure 2.10. This is a directed network of six nodes and six arrows between them.

Indegree and outdegree centralities will have differing significances depending on the particular nodes and arrows. Nevertheless, indegree centrality is often seen as more important, particularly in web networks. This is because a node that is pointed to by many other nodes seems to be singled out as important or at least highly visible. For instance, if the nodes are blogs and the arrows are hyperlinks between blogs then blogs with the highest indegree centrality are the most linked to. These seem likely to be the best known blogs and even the most authoritative blogs. In contrast the blogs with the highest outdegree centrality are those that link to the most other blogs. These might be from the most widely read bloggers but the blogs may not themselves be read much.

A more complex measure than degree centrality is *betweenness centrality*. This measures the importance of a node in a network to the transmission of information through the network and is also relevant to the way in which different clusters of nodes connect with each other. A node with a high betweenness centrality is important to a network in the sense that if it is removed from the network then either the network will split into separate components or information will take longer to flow through the network. The definition of betweenness centrality for a node A is that it is the probability that if two other nodes B and C are chosen at random from the network then A lies on the shortest path between B and C. In Figure 2.9, Node C has a high betweenness centrality because it lies on all paths between either A or B and any of D, E and F. In contrast, node E has a betweenness centrality of zero because it does not lie on the shortest path between any pair of nodes.

Usually nodes with high betweenness centrality also have high degree centrality and so this extra metric does not add much to what is already known about the importance of nodes in the network. In the few cases where there is a difference then the nature of the nodes with high betweenness centrality but not high degree centrality can be an important research finding. For instance, a network analysis of political blogs might find that the blogs tend to cluster by political parties. The most interesting blogs in this context might be the ones that connect to multiple political parties as these are unusual and may help information to flow between the opposing camps. This example also illustrates that although the betweenness centrality is designed for information flow it is also helpful to identify nodes that connect separate clusters even if no information flow is assumed.

Metrics for whole networks

Some metrics calculate a single number from an entire network. Such a metric can be useful to compare different networks. The main metric is *density*, which is the proportion of

theoretically possible connections that exist in the network. For instance, in Figure 2.9, there are 7 connections but if every node was connected to every other node then the total number of connections would be 15. Hence the density of the network is $7/15 = 0.467$. If one network is denser than another then this indicates that it has a higher proportion of the maximum number of connections for the network. It is not reasonable to compare the densities of networks that are greatly different in size, however. Large social networks are naturally less dense because there is a limit to the number of connections that a person normally has. For two similar-sized networks it can be useful to identify differences in density because this indicates a systematic difference in behaviour. For instance if the network of left wing bloggers is denser than the network of right wing bloggers then this suggests different underlying communication behaviours or strategies that further investigations might uncover.

If only one network is available for analysis then it may be possible to split it into separate parts and compare their densities. For a network of political bloggers the split could be by political affiliation, as mentioned above. As another example, for a social network the split could be along gender or geographic lines. Networks can also be split into disconnected components, if these exist.

A common task for a network is to identify the most important connected groups of nodes. There are two main SNA strategies to decompose a network into groups: clique calculations and cluster analysis.

A *p-clique* in a network is a set of p nodes that all connect to each other. This is a perfect group in the sense that within the group all possible connections exist. SNA software can be used to list all p -cliques in a network for any value of p . A simple p -clique analysis strategy is to use trial and error to identify the largest p such that p -cliques exist and then list and comment on all the p -cliques as important groups in the network from the perspective of connectedness. If there is only one maximum size p -clique then p could be reduced to identify more to comment on. In Figure 2.9 there are two 3-cliques: A, B, C and C, D, F. Within both of these sets of nodes, all nodes connect to all other nodes. There are no 4-cliques, however, because if any subset of 4 nodes is selected from Figure 2.9 then not all of the nodes will connect to all of the other nodes.

p -cliques have two disadvantages. First, they can overlap with each other, which makes the results awkward to interpret in some cases. Second, the condition that all nodes in the clique connect to all other nodes is restrictive in the sense that some important coherent groups would fail this clique test because one or more possible connections are absent.

An alternative to the clique approach is clustering. A network can be split into clusters of nodes, so that the nodes in each cluster tend to connect to each other more than to nodes outside the cluster. Switching from cliques to clusters has the advantage that the restrictive clique condition is removed. The main practical problem with clustering is that there are many different ways to do it, even with SNA software, and so it is not straightforward to decide which set of clusters to use. It is possible to decide to some extent in advance how many clusters should be found with some algorithms and this may help the decision if a particular number of clusters is expected. In other situations a reasonable approach is to try different clusterings and use the one that gives the clearest clusters. When reporting the results it is important that this strategy is described and justified. It is also critical that the clusters found are not described as *the* clusters in the network but as a way of clustering the network. This description helps to guard against overstating the importance of the clusters.

Summary and recommended reading

- SNA analyses networks as a whole and can identify patterns in the network that are not evident from studying the individual parts of the network separately.
- SNA has a number of centrality metrics for the nodes in a network that can be used to identify the most and least important nodes.
- SNA whole network metrics, such as density, allow different networks to be compared.
- The metrics can be calculated with SNA software.

Social Network Analysis has so many metrics that textbooks about it may be too full of information for the casual user to read comprehensively. As a result, it is recommended to read them selectively to extract information mainly about the metrics to be used.

- Hansen, D., Shneiderman, B., & Smith, M. A. (2010). *Analyzing social media networks with NodeXL: Insights from a connected world*. San Francisco, CA: Morgan Kaufmann.
- Nooy, W. d., Mrvar, A., & Batagelj, V. (2005). *Exploratory social network analysis with Pajek*. Cambridge: Cambridge University Press.
- Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications*. Cambridge, NY: Cambridge University Press.

Final thoughts

This chapter has covered the network approach to studying the social web. Most of the methods discussed in this section considerably predate the social web and are applicable in many different contexts but are particularly useful in the social web because the data for them can be collected easily or automatically. In contrast, previous researchers sometimes had to interview large numbers of people in order to collect enough data to build a network from. The methods in this chapter will be helped by specialist software for drawing networks and these are listed and discussed on the book's web site.

3. (L2) Link Analysis

This chapter introduces a range of link analysis techniques and explains why links can be a useful source of data for social scientists. Using links between web pages turns the web itself into a big data source since the web contains hundreds of billions of hyperlinks. Fortunately, it is not necessary to copy the entire web to analyse links within it. Instead, relevant sections of the web can be downloaded by web crawlers, or information about links can be extracted from the indexes of commercial search engines like Bing through appropriate queries.

Background: Link counts as a type of information

Although hyperlinks are designed purely as a navigational aid, to help web users to jump quickly between web pages, they also contain implicit information that can be exploited for research purposes. A hyperlink can sometimes be regarded as an endorsement of the target page, especially if the creator of the link has chosen it to point to a useful or important page. For instance, hyperlinks in a course web site might be chosen to point to web sites with relevant high quality information and to helpful resources such as an online library catalogue. As a result of many links being created like this, it seems that the best and most useful pages tend to attract the most hyperlinks. For instance, the Google home page is linked to by at least a hundred million pages, and major news web sites can expect to attract at least a million links whereas the average blog or personal home page could expect to attract few links or none at all because they are simply less useful and important.

The number of web pages linking to a given web page or web site is called its *inlink count*. The above discussion essentially argues that inlink counts may be a reasonable indicator of the importance of the target site or page. This property of inlink counts is exploited in Google's PageRank algorithm (Brin & Page, 1998), in Kleinberg's similar topic-based search algorithm HITS (Kleinberg, 1999) and also in the ranking algorithms of other search engines. These search ranking algorithms are designed so that the pages listed first in response to any standard keyword search tend to be those that have attracted the most links. If you query a search engine then it will first identify pages that match your search and then list them in an order that it guesses will place the most useful ones first – using hyperlink counts as important evidence for this ordering.

Inlink counts are not a perfect source of evidence, because not all hyperlinks are carefully created by web authors after consideration of which page would be the best link target. Many links between pages in a web site are created primarily for navigation within the site. As a result, these links are ignored or given low importance by search engines and are typically ignored in webometric link analyses. A different issue is that links are sometimes replicated within a web site. For example, a web site may have a link to Google on every page. It would not seem fair to count every such link because all are presumably the result of a single web author's decision to add a Google link everywhere. As a result search engines and webometrics calculations often count a maximum of one link between any pair of web sites. In other words, pairs of linking web *sites* are counted rather than pairs of linking web *pages*. Here a web site is normally either equated with the full domain name of each URL (e.g., www.microsoft.com, www.wlv.ac.uk, www.ucla.edu, www.bbc.co.uk) or with the ending of the URL associated with the domain name owner (e.g., microsoft.com, wlv.ac.uk, ucla.edu, bbc.co.uk).

In addition to links being useful indicators of the value of the target page they often connect pages about the same topic and so are useful as indicators of content similarity (Chakrabarti, Joshi, Punera, & Pennock, 2002). Additionally, links can be created to reflect or acknowledge organisational connections. For example a research group web site may link to the sites of other groups that it collaborates with, the organisations that have funded it and to its parent university and department. Similarly a local government body may link to its parent and subsidiary organisations' web sites and to companies that provide it with services. Finally, links are sometimes created for relatively trivial reasons, such as when an academic's home page contains links related to their hobbies. When carrying out a webometric link analysis, such links often form the unwanted "noise" in the results.

For webometric purposes, URL citations are often used instead of hyperlinks. An URL citation is a different kind of link: the mention of the URL of one page in another page with or without an associated hyperlink. URL citations are counted instead of hyperlinks when gathering data from commercial search engines like Bing or when using Webometric Analyst because these do not provide information about hyperlinks.

Types of webometric link analysis

Partly reflecting the differing types of reason for creating links discussed above, there are two main types of webometric link analysis study: *link impact assessments* and *link relationship mappings*. Link relationship mapping typically illustrates collections of web sites through the links associated with them. It includes link-based organisational relationship mapping and link-based maps of topic areas or organisations intended to reveal content similarity. In practice, these two types often overlap.

Link impact assessments typically start with a collection of web sites and compare the number of web pages or web sites that link to each one. This is essentially the same as the web impact assessment using keyword searches that is discussed in the previous chapter. The purpose of link impact assessment is often similar: to evaluate whether a given web site has a high link-based web impact compared to its peers. Link-based impact can also be used as an indirect indicator of other attributes of the owning organisation. For example, one study has shown that links to U.K. universities correlate highly with their research productivity, suggesting that link impact might also be used as an estimator of research productivity for universities in countries that do not publish research evaluation results (Thelwall, 2001). The Webometrics ranking of universities also uses links as one of its indicators of universities' online presence (Aguillo, Granadino, Ortega, & Prieto, 2006).

Link relationship mapping is normally carried out for web sites rather than web pages and results in some kind of diagram illustrating the relationships between the sites. The goal of the mapping may be exploratory: to get an overview of the web environment of the sites. Alternatively the map may be used to formally or informally test a specific hypothesis, such as whether two groups of sites tend to interlink with each other or whether the interlinking within one collection of sites is denser than the interlinking within another collection of sites. Another purpose can be to identify the overall pattern of interlinking, such as whether there are one or more central sites or whether the linking reflects the geographic locations of the sites (Vaughan & Thelwall, 2005).

The following sections describe the techniques needed for simple link impact assessments and relational link analysis.

Link impact assessments

The basic premise for link impact assessments is that a count of links to a web page or web site is a reasonable indicator of its utility or value. In parallel with citation analysis, the term *web impact* (sometimes shortened to just *impact*) is used to denote that which is represented by link counts. Starting with a collection of web sites, the data for a link impact assessment is the number of hyperlinks to each one from elsewhere on the web. The terminology *site inlink* (or *external inlink*) is used for the links sought: only those from outside of the web site being evaluated (i.e., excluding the internal site navigation links). These link counts used to be easily available from some search engines via the advanced link search commands but these commands have been withdrawn. Instead, URL citations must be used: mentions of the URL of a web page in the text of another. To search for the URL of one page in other pages, any search engine can be used with the query being the relevant URL in quotes and without the initial `http://`. For example the search "news.bbc.co.uk" matches web pages that mention the URL of any page within the BBC News web site news.bbc.co.uk, and the search "webometrics.blogspot.com" matches web pages that mention the URL of any page within this blog. These searches are not perfect for webometrics, however, because they include linking pages both inside and outside of the target site. Fortunately it is possible to modify the search to exclude the unwanted links from pages within the same site (called *site self-links*

and unwanted for the reasons discussed above). These pages can be identified and subsequently eliminated using the *site* advanced search command.

The *site* command matches all pages with the specified domain. Hence submitting the command `site:www.microsoft.com` to Bing or Google returns a list of pages within the `www.microsoft.com` web site. The trick for link searches is to subtract these results from the URL citation results. This can be achieved by adding a minus sign in front of the *site* command. All major search engines interpret the minus sign as an instruction to exclude results matching the subsequent keyword or search command. In order to obtain a list of pages that link to `www.microsoft.com` from outside the `www.microsoft.com` web site the following single search should therefore be issued in Bing or Google:

```
"www.microsoft.com" -site:www.microsoft.com
```

In other words this is an instruction to Bing or Google to identify all pages linking to any page in the `www.microsoft.com` web site, and then to exclude from those results all pages within the `www.microsoft.com` web site and report the remainder. The following two searches use the same pattern.

```
"news.bbc.co.uk" -site:news.bbc.co.uk
```

```
"webometrics.blogspot.com" -site:webometrics.blogspot.com
```

In fact the above searches are still not quite optimal. The reason is that large web sites typically have important subdomains and these should be taken into account when constructing the searches. For instance the `www.microsoft.com` web site may have derivative web sites with domain names like `blog.microsoft.com`, `developer.microsoft.com` and `search.microsoft.com`. Hence to be on the safe side the search should be modified to encompass all these subdomains. As long as this does not generate false matches, the search can be modified by truncating the domain name concerned to the part that sufficiently and uniquely identifies the whole web site (the truncation must be performed at a dot in the domain name though). In the above example this would be `microsoft.com` and so the final and best Bing or Google search for pages that link to the Microsoft web site from other sites is the following:

```
"microsoft.com" -site:microsoft.com
```

Assuming that for the BBC News search the focus is only on BBC News and not other aspects of the BBC then it would not be desirable to change `news.bbc.co.uk` to `bbc.co.uk` for the URL citation part of the command but it would be useful to make this change for the *site* command to remove any internal BBC links. Hence the eventual Bing search would be the following:

```
"news.bbc.co.uk" -site:bbc.co.uk
```

It would also be possible to modify the above search to exclude other BBC owned sites like `www.bbc.com` by subtracting them, as follows:

```
"news.bbc.co.uk" -site:bbc.co.uk -site:bbc.com
```

The `webometrics.blogspot.com` search above does not need to be modified because it has no derivative subdomains. Although it is itself a derivative subdomain of `blogspot.com` it would not make sense to exclude all links from `blogspot.com` (e.g., changing `-site:webometrics.blogspot.com` to `-site:blogspot.com`) because `blogspot.com` blogs are typically owned by different and unrelated individuals and so links between different `blogspot.com` blogs tend not to be in any sense internal site navigation links.

The Google link search is not recommended, however, because it cannot be combined with the other commands like `-site:` and because it reports only a small fraction of the links known to Google.

The following are guidelines for an initial link impact assessment. All the necessary raw data for this is provided by Webometric Analyst reports.

- Compare the overall hit counts to discover which pages or sites have the most impact, notice any particularly weak or strong performers and comment on why they might be weak or strong.

- Identify the most popular *types* of site linking to each page or site (e.g., news, blogs, universities) and comment on why they might be popular.
- Identify *differences* in the most popular *types* of site between the different pages or sites and comment on why the differences might be there. Are any of the differences surprising or glaring omissions? For instance, an academic document not attracting mentions from any universities would be surprising.
- Identify the countries that most often link to each page or site (e.g., from the Top-Level Domains of the results, ignoring .com, .net, .org) and comment on why they might be hosting so many links.
- Identify differences in the most popular countries between the different pages or sites and comment on why the differences might be there. Are any of the differences surprising or glaring omissions? For example, perhaps one site has attracted much attention in the US or another page is the only one not linked to from Mexico.

Finally, please see the following page for information about syntax for link-related searches in the major commercial search engines
<http://cybermetrics.wlv.ac.uk/QueriesForWebometrics.htm>.

Alternative sources of link data: Alexa and link databases

Alexa.com's *Sites Linking In* statistic is an alternative to URL citations for information about the number of links to a web site. For web sites covered by Alexa.com, visiting their page gives this statistic, as gathered by Alexa from information sent by the toolbars of registered users. Clicking on the link gives a list of 100 of the sites linking in (e.g., <http://www.alexacom.com/siteinfo/chicagotribune.com>). This has been shown to be the best available source of evidence about the links to a web site (Vaughan & Yang, 2012). Its main disadvantages are that it can only be used for inlink counts and not for other types of statistic (e.g., co-links or direct link counts) and it is only available for major web sites and not for subdomains – for instance it works for wlv.ac.uk but not for webometrics.wlv.ac.uk.

There are also search engines that track links and give access to link data for members, normally for a fee. An example is Blekko SEO Tools, which delivers data about hyperlinks that is targeted at Search Engine Optimisation (SEO) professionals. These people may try to exploit information about hyperlinks to get a higher ranking in search engines. The same link data could, in theory, be used for research purposes.

Alternative link counting methods

Once searches have been designed to gather the link data, a decision should be made about whether to use the reported hit count estimates on the results pages, the list of matching *URLs* reported or to calculate how many different *sites* are in the list of matching *URLs*. This parallels a similar decision for web impact reports. The simplest statistic to gather is the hit count estimates reported on the first results pages but recall that this can be considerably different from the actual number of matching *URLs* returned. It is not true that the number of *URLs* returned is a more accurate statistic than the hit count estimates, because the former is likely to be a significant underestimate because search engines filter results to eliminate similar pages and too many pages from the same site (see the chapter on search engines).

The best statistic is the number of linking web sites because of the possibility of repeated URL citation links within a site, for example if a link acknowledging collaboration, web site design or funding is placed on every page of an organisation's web site. The number of linking sites can be calculated by manual checking of the number of different domain names represented in the *URLs* of the results. This process can be automated with appropriate software like Webometric Analyst but, whether manual or automatic, the ability to count domains is dependent upon there being fewer matching *URLs* than the maximum returned by a search engine, which is normally 1,000. If there are more results than this then the query splitting technique discussed elsewhere in this book can be used.

Once the URL citation link count statistics have been obtained, they can conveniently be reported in a graph or table for ease of comparison, unless they are used in a specialist

statistical test. It should be remembered and reported that the counts do not report the whole web, just the part of the web covered by the search engine used, and also that the search engine may not reveal all the information that it knows.

Webometric link analysis studies sometimes also use web crawlers or automatic search engine queries to gather data and these are discussed elsewhere in this book.

Case study: Links to ZigZagMag.com

This section presents a brief summary of a web impact evaluation of a web site. This evaluation was carried out to assess the web impact of the BBC World Service Trust's ZigZagMag youth culture magazine web site that was part of its Iranian journalism training project. An overall assessment of the impact of ZigZagMag was made in comparison to the most popular similar web sites in Iran, as identified by ZigZagMag users. Table 3.1 reports the results. The number of pages with URL citation links to each site can be identified by Webometric Analyst Bing queries (e.g., "haftan.com" -site:haftan.com for the first search), using query splitting to gain additional matches beyond the first 1000. The main impact statistic used was the number of web sites linking to each of the 12 sites, counted by domain name. The results showed that the web impact of ZigZagMag was comparable to that of the top similar web sites in Iran and that it was therefore a highly successful site, especially given its relative youth.

Table 3.1. A comparative impact evaluation of ZigZagMag through counts of URL citation links to its web site and to the best-known similar web sites.

Website name and genre	Estimated number of pages linking to site	Est. number of sites (counted by domain name) linking to site	Est. number of TLDs linking to site
Haftan - online portal (haftan.com)	31,300	1,556	22
Balatarin - news sharing site (balatarin.com)	57,400	1,455	28
7Sang - online magazine (7sang.com)	11,900	1,250	32
Meydaan - women's rights site (meydaan.org)	9,940	821	35
Sobhaneh - online portal (sobhaneh.org)	9,320	633	21
40cheragh - weekly magazine (40cheragh.org)	18,700	594	14
ZigZagMag (ZigZagMag.com)	64,300	575	21
Radio Zamaneh (radiozamaneh.info, Dutch-funded)	6,530	468	18
Rang - online magazine (rangmagazine.com)	9,330	346	14
Jadid Media (jadidmedia.com, Dutch-funded)	7,350	251	17
Jadid Media (jadidonline.com, Dutch-funded)	3,270	224	20
ZigZagMag (ZigZagMag.net -blocked in Iran)	48,400	91	14

Content analysis of links

As with web impact evaluations, a content analysis of a random sample of the links is highly desirable in order to be able to interpret the link counts. Without this, the significance of the link counts in Table 3.1 could only be described in very general terms. The classification should be conducted as discussed for web impact assessment but focussing on the context of each link in the page containing it. The content analysis should shed light on why the links were created using categories that are meaningful for the objective of the link impact study. See the content analysis section of the web impact evaluations chapter for more details.

In some cases the content analysis rather than the link counts can be the objective of a link analysis. For example, one study attempted to find out why political web sites used hyperlinks. The findings included that these links were much more often made to like-minded sites than to opposing sites to engage in debate. Links also appeared to be included sometimes in apparent imitation of other web sites (Foot & Schneider, 2002). Table 3.2 summarises the results of another content analysis study. The objective was to identify the types of reasons

why links were created to university home pages, in the belief that some of these links were fairly trivial. In support of this, the categories include some for links that it was argued were unlikely to be useful for the source page visitors, such as a link to a person's previous employer or their degree awarding institution.

Table 3.2. An example of the results of a content analysis. The data is source pages for 100 random links to external UK university home pages (Thelwall, 2003).

Type of page/type of link	Count
General list of links to all university home pages	16
Regional university home page link list	2
Personal bookmarks	2
Subject-based link list	5
Other link lists	6
Personal home page of lecturer	
/ link to degree awarding institution	8
/ link to previous employer	6
/ link to collaborator's institution	3
/ other	3
Collaborative research project page/ link to partner site	17
Other research page	
/ link to collaborator's institution	3
/ link to institution of conference speaker	2
/ link to institution hosting conference	2
/ other	3
Link to home institution of document author e.g. in mirror site	7
Collaborative student support	
/ link to partner institution	6
/ link to institution for access to information	4
Other type of page	5

Link relationship mapping

It is sometimes useful to illustrate the pattern of interlinking within a collection of web sites. The natural form for this is a network diagram with small circles (or nodes) representing web sites and arrows between the circles representing links between them. Starting with a collection of web sites, the data needed for a standard link relationship map is the number of links between each pair of web sites, counted separately for each direction of links. The data can be obtained from Bing or Google using a combination of URL citation and site searches. The search pattern "A" site:B matches pages in web site B that link to web site A. For example, the following search in Bing would return pages in news.bbc.co.uk that contain an URL citation to the www.microsoft.com web site.

```
"www.microsoft.com" site:news.bbc.co.uk
```

This search is similar to the form used for link impact calculations except that there is no minus sign before the *site:* command and the two domain names are always different. To search for links from www.microsoft.com to news.bbc.co.uk the order of the domains is reversed as follows:

```
"news.bbc.co.uk" site:www.microsoft.com
```

As in the case of link impact measures it may be better to replace full domain names with the shortest part uniquely identifying the web site. In the above example news.bbc.co.uk would not be changed but the other domain could be shortened to microsoft.com.

For the link impact searches the recommended counting method is to count pairs of interlinking *sites* rather than pages because of the problem of unwanted replicated links in some sites. This method could be applied to the data gathered as above but since only two domains are involved in each search the result would always be 1 link or 0 links. In some

cases the strength of connections between pairs of sites is important, however, and so link pages can be counted directly from the results or the hit count estimates on the results pages could be used instead.

If many web sites are involved in a study then it could take a long time to calculate the number of links between them due to the number of searches needed. Mathematically, if there are n websites then n^2-n link searches are needed to get all the link counts. For example, 10 sites would need $10^2-10 = 90$ searches but 100 sites would need too many to run by hand: $100^2-100=9,900$. Such searches can be automated for large numbers of sites, however, as discussed in the automatic search engine searches chapter.

The key issue for link relationship mapping is how to graph the results. The following list gives a range of common options, roughly in ascending order of complexity. The list is not exhaustive and more exotic techniques are always possible. If software like Webometric Analyst is used to gather the data then the same software can probably directly produce the visualisation or save the data in a format suitable for importing into Pajek or another network drawing program.

- A *simple network diagram* with circles representing web sites (or pages) and identical arrows representing links between web sites. This kind of illustration could be drawn with any graphical software package, such as Corel Draw or Microsoft Paint, or with a specialist network analysis package, such as Pajek (Holmberg & Thelwall, 2009). In order to make large network diagrams as easy to interpret as possible, it is desirable to position web sites close together if they interlink and further apart if they do not interlink. In addition, the positioning should try to minimise cross-overs between arrows. If a network analysis program like Pajek is used to produce the graph then its network layout algorithms like Kamada-Kawai (1989) or Fruchterman-Reingold (1991) can be used as they partially achieve these goals.
- A *node-positioned diagram* is a network diagram in which the circles representing web sites (often known as nodes in network terminology) are positioned in order to convey information about interlinking through the positioning. This is particularly important for diagrams in which there are many nodes and so the “shorthand” of node positioning helps to reveal linking patterns. Statistical node-positioning techniques, such as the Multi-Dimensional Scaling can also be used to position the nodes (Vaughan & Wu, 2004). In this type of diagram the links between sites are often not drawn, because the position of the nodes conveys the necessary interlinking information.
- A *geographic network diagram* is a network diagram in which the circles representing web sites (i.e., the nodes) are positioned on a geographic map showing their origins in order to highlight geographic linking patterns (Ortega & Aguillo, 2009).

Network diagrams are most suitable for small or medium-sized networks. If there are too many web sites then it can be impossible to see any pattern in the visualisation because there are so many arrows that they all overlap and so the node-positioned diagram is more suitable. For any of the above types of network diagram the appearance can be customised to provide additional information.

- *Variable arrow-widths*: setting the width of the arrows to be proportional to the number of links between the source and target web site is useful if the number of links is an important factor. A threshold minimum number of links for a line to be drawn can be set so that the thinnest lines are omitted. There are alternative ways in which line widths can be calculated, which are especially useful for web sites of significantly varying sizes. For example the widths could be divided by the number of pages in the source web site, the target web site, or both (Thelwall, 2001). This avoids the results being dominated by the largest sites.
- *Variable nodes*: the appearance of nodes can illustrate properties of the associated web sites. For instance, the area of the circles representing the web sites could represent their sizes or inlink counts. Different shapes or colours can also illustrate properties

such as the type of organisation owning the web site or its country of origin (for several examples see: Heimeriks, Hörlesberger, & van den Besselaar, 2003; Ortega & Aguillo, 2008).

In addition to constructing diagrams it is also possible to calculate related statistics from network data, many of which originate or have equivalents from the field of social network analysis (Björneborn, 2006). The simplest statistics are the number of links to each web site and the number of links from each web site. This data can be used to find the web site with the most inlinks and the site with the most outlinks, which may be the key sites in the network. The average number of inlinks and outlinks per site can also be calculated as a descriptive statistic and this can also be used to compare different networks.

Within the field of complex networks there are also many algorithms to calculate statistics from network data (Börner, Sanyal, & Vespignani, 2007; Newman, 2003) and these tend to be most useful for those seeking to model the dynamic forces underlying the creation of a network rather than to understand the network itself. Within computer science there are also many sophisticated network visualisation techniques that advanced users may wish to try (Chen, 2004), as well as methods to automatically identify clusters of similar sites within a network (Flake, Lawrence, Giles, & Coetzee, 2002).

Case studies

This section introduces a few different types of network diagram to illustrate the range of visualisation techniques available. In each case, little information is given about the underlying study because it is the diagram that is of primary interest.

Figure 3.1 is a simple network diagram of interlinking between the top 5 universities in Asian and European nations with nodes positioned by the Kamada-Kawai algorithm. Without a positioning algorithm the patterns in the diagram would have been difficult to detect because the lines would overlap too much. Most of the web sites interlinked and so a threshold of 100 links was used as the minimum to draw an arrow. Hence this diagram shows the most frequently interlinking web sites. As an additional step for clarity, universities not connected to any others in the diagram were removed. The diagram is successful in the sense that patterns are relatively easy to distinguish. For example, it is not difficult to notice from the top level domains that universities from the same country tend to interlink and that UK universities form the core of the network.

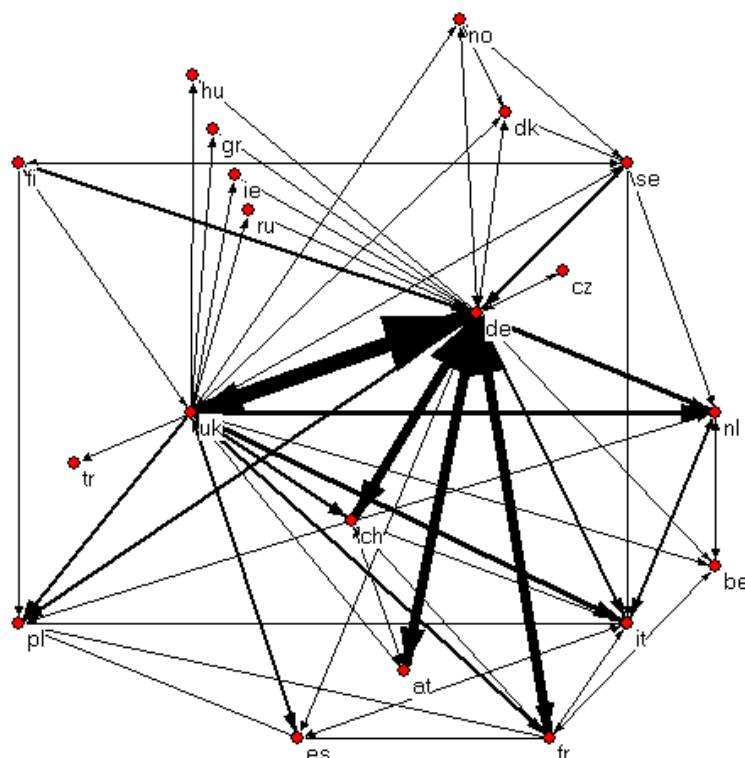


Fig 3.2. Variable arrow-width diagram of interlinking between European universities (Thelwall & Zuccala, 2008).

Figures 3.3 and 3.4 are of similar data to Figure 3.2 – European university interlinking - but are drawn in a different way. The universities are positioned in a circle at random and arrow widths are proportional to the number of pages in a given language (Swedish or Spanish) linking from universities in the source country to universities in the target country, divided both by the number of pages in the source and in the target country university system (because of the greatly differing sizes of web sites involved). A node positioning algorithm like Kamada-Kawai could have been used for this data but it was judged not necessary since the pattern was clear from the circular graphs. The top diagram shows that Swedish interlinking mostly connects Sweden (se), Norway (no) and Denmark (dk), whereas the lower diagram shows that in European universities, Spanish interlinking almost exclusively originates within Spain.

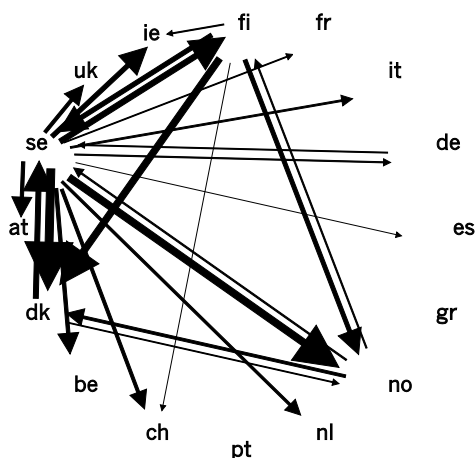


Fig 3.3. Variable arrow-width diagram of interlinking between European universities in Swedish pages (Thelwall, Tang, & Price, 2003).

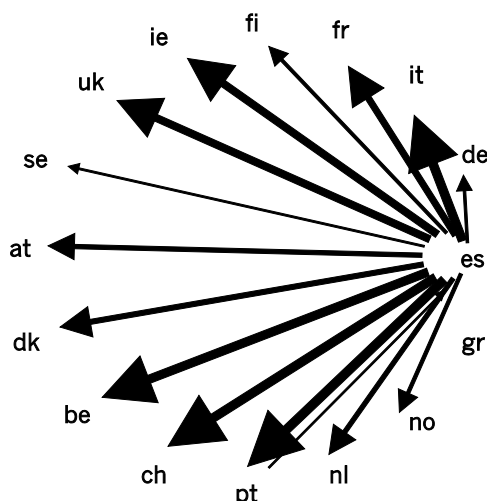


Fig 3.4. Variable arrow-width diagram of interlinking between European universities in Spanish pages (Thelwall et al., 2003).

Figure 3.5 is a Multi-Dimensional Scaling map that does not include any arrows representing links but which uses link data to position the universities (the acronyms are web site names of UK universities). This approach, which clusters together universities that tend to interlink a lot, is appropriate due to the large number of web sites involved – too many to give details of each individual university.

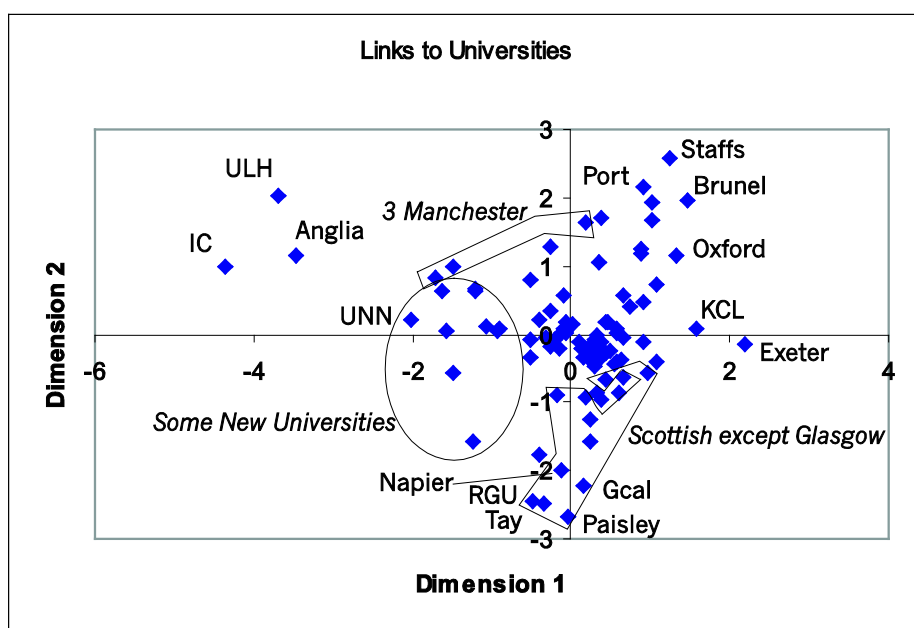


Fig 3.5. A multi-dimensional scaling positioning of UK universities according to the links between them (Thelwall, 2002).

Co-link relationship mapping

For some types of web site interlinking is rare and so they produce poor link networks. For instance, competing companies in the same market sector understandably seem almost never to link to each other. It may still be possible to generate a relationship map for such collections of sites using types of indirect links known as co-links (Björneborn & Ingwersen, 2004). Two web sites A and B are said to be *co-linked* (or co-inlinked) if there is a page in a third site C that links to both A and B. On average, there are more co-links than direct links and co-inlinks seem more likely to exist for similar companies than direct links. As a result, co-inlinks or co-inlink counts are useful for mapping sets of commercial web sites and also

for other cases where there is not enough direct link data to generate a reasonable map (Vaughan & You, 2005). The majority of hyperlink-based mapping research has used co-links rather than links (e.g., Vaughan, 2006; e.g., Zuccala, 2006), although some have used both (Ortega, Aguillo, Cothey, & Scharnhorst, 2008).

The format of the URL citation col-ink search is "A" "B" -site:A -site:B, where A and B are the domain names of the web sites for which co-inlink counts are being obtained. In fact this query matches pages that link to any page in site A and any page in site B but are not already in site A or site B. Fewer co-link searches are needed than link searches ($(n^2-n)/2$ for n sites) because this is a symmetric measure – although links can go either from A to B or from B to A, there is only one type of co-inlink between A and B. As a result, co-inlink network diagram should have lines rather than arrows between pairs of co-inlinked web sites. In most investigations using co-links multidimensional scaling maps have been used instead of network diagrams but there is no reason why a network diagram could not be used (Vaughan, 2006).

Figure 3.6 is an example of a co-inlink diagram, produced to illustrate the web environment of the ZigZagMag web site. For those who know the Iranian web(!) it shows that ZigZagMag associates strongly with Persian news sites and Persian blogs, and more weakly with international news sites, and blogs and with other Persian sites.

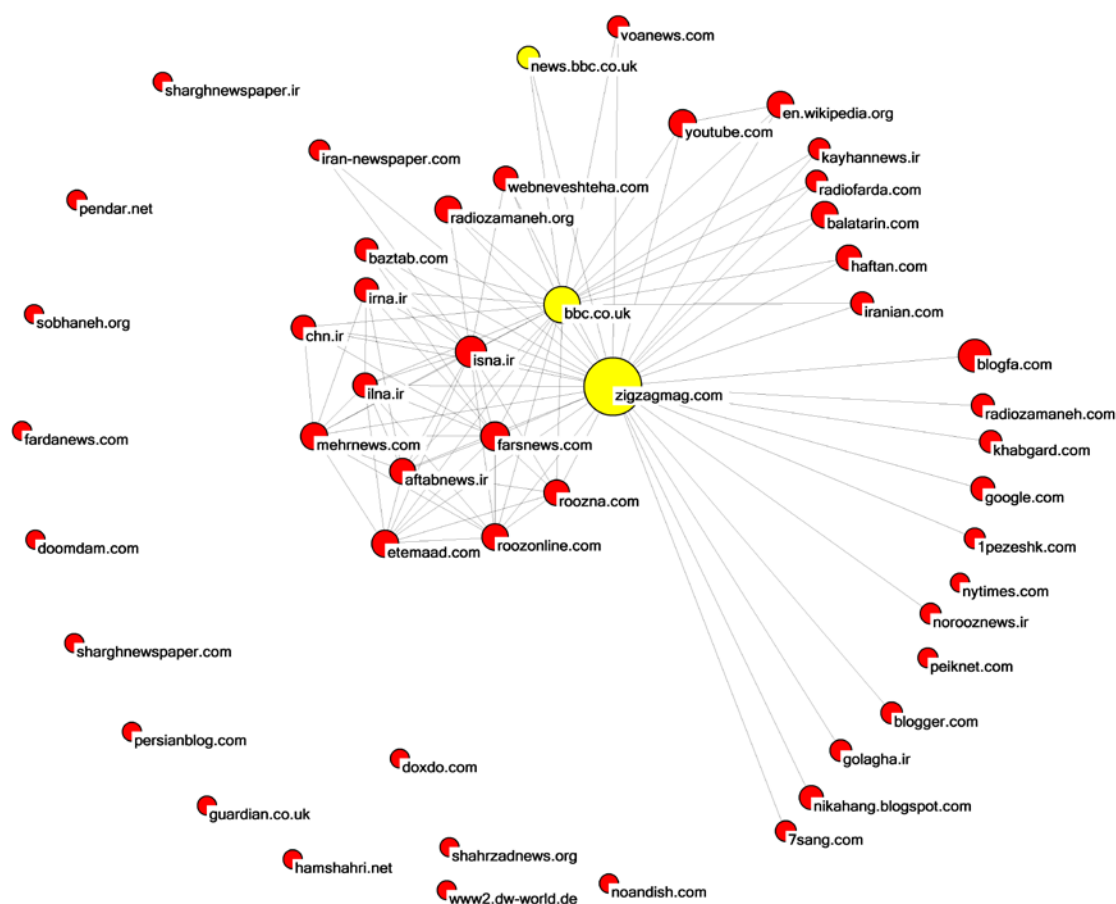


Fig 3.6. A co-inlink diagram of web sites linking to ZigZagMag.com, with lines connecting the most strongly co-inlinked pairs of web sites, using Fruchterman-Reingold positioning, circle areas representing total inlink counts and light shaded circles being associated with the BBC (Godfrey, Thelwall, Enayat, & Power, 2008).

Link differences between sectors – An information science application

This section briefly describes some link analysis research that is useful background for anyone designing a link analysis study. The underlying question is: are hyperlinks used in

similar ways by everyone? A few studies have shown that there are significant differences between organisational sectors in the way that web links are created. This result can help manage expectations about what is possible from a link analysis. The differences are clearest in an investigation into hyperlinking between university, industry and government web sites in the UK West Midlands. The results indicate that academic web sites contain far more links than the others, with commercial web sites containing the fewest.

It is part of academic culture to share information (e.g., in journal publications) and to draw upon the work of others (e.g., in reference lists). This is true in both research and teaching. Hence it seems natural for academic web sites to link to other web sites for information and to contain much information that other web sites could link to. In addition, academic hyperlinks are sometimes created to acknowledge a relationship, such as research collaboration or a government or commercial source of funding. Academic web sites also tend to be quite large and to have parts that are created and managed by various university members, such as staff personal home pages and research group web sites (Stuart & Thelwall, 2006; Stuart, Thelwall, & Harries, 2007).

Commercial web sites, in contrast, seem to be primarily marketing devices, designed to showcase a company's products or services and often controlled by the marketing department. In this context it does not make sense to link to competitors' web sites since this may lose customers. There is hence an incentive to avoid having all links to other web sites. Nevertheless, some links may be created to a parent company, the web site design company, an online store for purchases and useful supporting services (e.g., for related insurance).

In between these two extremes, government web sites (including local government) seem to be tightly managed but allow linking to acknowledge organisational structures, such as to parent or sibling organisations. In addition, there may be many public service links to useful and trusted sources of information (e.g., online health sites, bus timetables) or services (e.g., pest control, local attractions) (Holmberg & Thelwall, 2009).

In consequence of the above, link analyses based upon hyperlinks between, to or from academic web sites are likely to reveal clear patterns - perhaps related to information flows. The same is true for government web sites except that the patterns may reflect organisational structure and services instead (and hence show strong geographic trends) and the patterns may be weaker due to fewer links. In contrast, studies of links between commercial web sites are likely to be unfruitful. Links from commercial web sites may well also be unfruitful but could reveal organisational or symbiotic relationships between businesses. Links to businesses may be more interesting, but these could easily be dominated by links from online directories, which would undermine any patterns found. As describe above, one way to identify relationships between companies by link analysis is to use co-inlinks instead (Vaughan & You, 2005). This method estimates the similarity between two web sites by the number of web pages that simultaneously link to both (i.e., co-inlink counts). For a collection of business web sites a similarity graph can be drawn, using the co-inlink counts between all possible pairs of sites.

Summary

Link analysis is an approach for generating and analysing data about the *impact* of information or organisations on the web, or the online *relationships* between documents, web sites or organisations. The purpose can either be to directly study the web itself or to use the web as an indirect source of evidence about offline phenomena, such as inter-organisational relationships. The data for link analysis can originate from commercial search engine searches or from research crawlers and can be reported in a variety of summary formats, such as inlink counts and network diagrams. Whilst link analysis seems relatively quick compared to most social science research methods, the results need careful interpretation with the aid of content analyses of links. Links can be an ideal source of up-to-date information and are particularly useful for pilot or large-scale studies and when used in conjunction with other methods (e.g., interviews) or data sources.

4. (L3) Web Impact Assessment

Web impact assessment (WIA) is the evaluation of the “web impact” of documents or ideas by counting how often they are mentioned online. The underpinning idea is that, other factors being equal, documents or ideas having more impact are likely to be mentioned online more. This concept essentially originates from an early study that counted how often prominent academics were mentioned online and the contexts in which they were mentioned (Cronin, Snyder, Rosenbaum, Martinson, & Callahan, 1998). The results were proposed as a new way of getting insights into “academic interaction”.

Researchers that are interested in examining the impact of specific ideas or documents may therefore benefit from a web impact assessment. For instance, the goal may be to compare the influence, spread or support of competing academic theories, political candidates or a number of similar books. In such cases, comparing the web impact of the theories/candidates/books (e.g., the number of web pages mentioning each one) can be a proxy for their offline impact or a direct measure of web impact.

- For academic theories, it would be interesting to find out in which countries and universities they are mentioned. Whilst it would be impractically difficult to find this out offline (e.g., via phone calls or interviews with random academics across the globe) it would be much easier to do this online by counting web mentions using the webometric techniques described in this chapter. This would give useful results but of course the absence of a theory from a university web site would not always mean that the theory was unknown in the university. It might be known and even taught but not mentioned online. Hence the web results would be indicative of a minimum level of awareness rather than definitive. Nevertheless, in a comparative study of two or more theories, each one would suffer from the same limitations and so comparing the results between them should be informative about which was the best known overall.
- For the example of two (or more) political candidates, the ultimate indicator of success would be the votes polled. A web impact assessment would be useful to evaluate the online component of the candidates’ campaigns, however. This could judge how successful each campaign was as well as where it was successful, with whom and why. For such an issue, the restriction of the data collected to just online information would be an advantage rather than a limitation. A web impact assessment could also be used beforehand to help judge the relative success of their campaigns or even to predict a winner, but opinion polls would be better for both of these tasks.
- To compare the sphere of influence of two books, web searches could be conducted to identify how often each one was mentioned and in which countries. In this case, publishers’ sales figures would give more accurate information about the geographic spread of purchasers but would not be able to give data on the context in which the books were read or the opinions of the purchasers. This gap could be filled by web searches finding information such as blog postings reviewing or discussing the books.

A related commercial application of web impact assessment is for organisations producing or promoting ideas or information and needing to be able to provide evidence of their spread or impact. For example, a public body charged with promoting innovation may need to provide evidence of the impact of its work. It may not be possible to directly measure the desired outcomes (e.g., business innovation) and so such an organisation may be forced to turn to indirect indicators, such as the number of attendees at its events, its press coverage (via the LexisNexis database) or the number of requests for its literature. The Web provides a new way to generate indirect indicators of this kind. Web searches can be used to count the number of times that its publications or concepts have been mentioned online. A publication that has been mentioned frequently online, for example in blogs, news web sites and online reports, is likely to have had a big impact. Moreover, an analysis of web impact could be conducted comparatively, with mentions of the reports or ideas of similar organisations also measured in order to benchmark performance.

The practice of evaluating the impact of documents through counting how often they are mentioned predates the web in the form of citation analysis (Borgman & Furner, 2002; Garfield, 1970; Moed, 2005; Nicolaisen, 2007) but the principles are very similar. In citation analysis, the number of times an academic article is cited by another article (e.g., included in a reference list) is taken as an indicator of the impact of the research described in it. The basis for this, drawing upon Merton's (1973) sociology of science, is that science is a cumulative process and that articles which get cited by other articles are likely to be demonstrating their contribution to the overall scientific endeavour. Numerous scientists around the world are now assessed partly on the basis of how many citations their published articles attract. Web impact assessment cannot claim the same deep theoretical foundations as citation analysis and the evidence provided by online mentions is consequently less strong but it can nevertheless give useful web impact indicators. The following sections give advice about conducting a web impact assessment based upon mentions in general web documents, and other sections discuss the sub-topic of web citation analysis.

Web impact assessment via web mentions

A simple type of web impact assessment is to take a collection of ideas or documents and then submit them as phrase searches into a commercial search engine, using the reported hit count estimates as the impact evidence. The hit count estimates are the numbers reported by search engines in their results pages as the estimated maximum number of matching pages (e.g., the figure 373,000,000 in "Results 1 - 10 of about 373,000,000 for Obama" at the top of a Google search results page). If a comparative assessment is needed for benchmarking purposes then the same procedure should be repeated for an additional collection of ideas or documents chosen from a suitable similar source. Table 2.1 gives a simple example of a comparison of the web impact of four medieval history books.

Table 2.1. A comparative web impact assessment of four books.

Book title	Google query used (title and author)	Hit count estimate
Handbook of Medieval Sexuality	"Handbook of Medieval Sexuality" Bullough	3,780
The End of the House of Lancaster	"The End of the House of Lancaster" Storey	2,610
The Reign of King Henry VI	"The Reign of King Henry VI" Griffiths	1,340
Medieval Women: A Social History of Women in England 450-1500	"Medieval Women: A Social History of Women in England 450-1500" Leyser	973

The simple procedure above has a number of drawbacks but some refinements can be used to make the evidence more robust. The first refinement is that the keyword searches should be checked to ensure that the overwhelming majority (e.g., 90%) of the pages returned correctly mention the desired document or idea. It is likely that some of the searches will return many spurious matches, especially if they are short or use common words. In these cases the search should be modified to filter out false matches. For example, in the case of a book the author could be added to the search (as in Table 2.1). In the case of an idea, words very commonly associated with it could be added. Unfortunately, modifying the keyword searches makes subsequent comparisons unfair but there may be no alternatives and so any unfairness of this nature should be reported alongside the results. For some searches it may be impossible to modify them in any way to get mainly correct matches and these can either be dropped from the analysis or their results manually checked to identify correct matches. An example of the latter situation is the physics journal *Small*: it is impossible to construct a web search with reasonable coverage such that the majority of matching pages mention this journal because its title is such a common word.

A second refinement is that it is best to identify and count all matching pages rather than rely upon search engine hit count estimates. This is because these estimates can sometimes be unreliable – typically they overestimate the number of results that the search engines would return. It is not always possible to count all matches because search engines report a maximum of 1,000. This technique is therefore only practical for searches that generate fewer total matches. There is an alternative, however, which is discussed later in the book: to automate the URL counting and to use the “query splitting” technique to get additional matches when the maximum of 1,000 is reached. Table 2.2 contains web pages counted via their URLs in the results pages for the specified Google search. Note that all the results are significantly smaller than the estimates reported in Table 2.1.

Table 2.2. A comparative web impact assessment of four books – improved figures.

Book title	Google query used (title and author)	URLs
Handbook of Medieval Sexuality	"Handbook of Medieval Sexuality" Bullough	782
The End of the House of Lancaster	"The End of the House of Lancaster" Storey	653
The Reign of King Henry VI	"The Reign of King Henry VI" Griffiths	541
Medieval Women: A Social History of Women in England 450-1500	"Medieval Women: A Social History of Women in England 450-1500" Leyser	498

A third refinement is to count matching web *sites* rather than matching web *pages* (i.e., URLs). This is important because some web sites may mention an idea repeatedly whereas others may mention it only once. Whilst repeated mentions probably suggest more interest in the idea than a single mention, it could also just mean that similar content has been copied across multiple pages in a site. Hence it seems safest to count mentioning web sites rather than web pages. Although there is no single accepted definition of “web site” it is convenient and reasonable to equate web sites with domain names so that two pages sharing the same domain name are always deemed to come from the same site. As with URL counting, this process can be automated (e.g., using Webometric Analyst), as discussed in other chapters in this book. Table 2.3 contains web pages counted by first extracting their URLs from the results pages for the specified Google search, then identifying the domain name for each URL, then counting the number of unique domain names in the resulting list.

Table 2.3. A comparative web impact assessment of four books – counted by web site, on the basis of the full domain names of URLs matching the Google query.

Book title	Google query used (title and author)	Web sites
Handbook of Medieval Sexuality	"Handbook of Medieval Sexuality" Bullough	401
The End of the House of Lancaster	"The End of the House of Lancaster" Storey	378
The Reign of King Henry VI	"The Reign of King Henry VI" Griffiths	327
Medieval Women: A Social History of Women in England 450-1500	"Medieval Women: A Social History of Women in England 450-1500" Leyser	288

Even with the above refinements there are disadvantages with web impact measurements. The principal disadvantage is that the unregulated and anarchic nature of the web means that the scope of the web sample is likely to be unclear. For example, suppose that an idea of “Farm Widgets” is important to the organisation PromoteBusinessWidgets. Farm Widgets are likely to be mentioned by PromoteBusinessWidgets in its web site (and hence picked up in a web

impact report) if PromoteBusinessWidgets has a web site and if that web site is sufficiently informative to mention the importance of the idea. This seems more likely to be the case if PromoteBusinessWidgets is a large organisation or if it is an organisation that regards its web site as important. As a result of cases like this, it is possible for an idea or publication to have a (small) web impact even when it is not well-known outside the originating organisation. This impact could be multiplied if the organisation maintained several web sites or placed adverts online. Conversely, a document could be widely read by a community that does not extensively blog or publish online. For example, a leaflet with guidelines for the safe use of Zimmer frames would presumably be targeted at the elderly, who seem to be a sector of society that does not extensively publish online. As a result of factors like these, web impact evidence should be interpreted as indicative rather than definitive. Web impact evidence would be strongest in cases where it would be reasonable to assume that web impact would fairly reflect offline impact, or for cases where a web-based phenomenon is being assessed (e.g., blogging or online electioneering).

The following are guidelines for an initial web impact assessment. All the necessary raw data for this is provided by Webometric Analyst reports.

- Compare the overall hit counts to discover which documents or ideas have the most impact, notice any particularly weak or strong performers and comment on why they might be weak or strong.
- Identify the most popular *types* of site for each document or idea (e.g., news, blogs, universities) and comment on why they might be popular.
- Identify differences in the most popular *types* of site between the different documents or ideas and comment on why the differences might be there. Are any of the differences surprising or glaring omissions? For instance, an academic document not attracting mentions from any universities would be surprising.
- Identify the most popular countries for each document or idea (e.g., from the Top-Level Domains of the results, ignoring .com, .net, .org) and comment on why they might be mentioned so often in these countries.
- Identify differences in the most popular countries between the different documents or ideas and comment on why the differences might be there. Are any of the differences surprising or glaring omissions? For example, perhaps one document has attracted a lot of attention in China or another document is the only one not mentioned in Spain.

Bespoke Web citation indexes

An impact assessment exercise may less concerned with general web mentions of an idea or report than with mentions in online academic documents or other formal reports. In such cases general web searches would likely produce too many spurious matches to be useful. It is relatively easy to narrow down the general web search results, however, using the knowledge that academic articles and reports are normally in PDF format when posted online, although they may sometimes be posted as Word documents or other word processing formats. As a result, the majority of mentioning documents can probably be found by running an appropriate keyword search but turning it into an advanced search by restricting the results to only PDFs, for example by adding the advanced search command `filetype:pdf` to the search. To include Word documents in the results, additional searches can be run with `filetype:doc` or `filetype:docx` to capture Word documents. The number of matching documents could then be counted and used as the impact evidence, as in Table 2.4. Totalling the hit count estimates in Table 2.4 gives an overall report impact estimate as follows.

- Ready steady innovate!: $3+0 = 3$
- Agricultural widgets in Norfolk: $1+1 = 2$
- Innovative agricultural widget designs: $0+0 = 0$

Table 2.4. A comparative web impact assessment of three reports based upon the number of citing PDF documents found.

Report title (made up)	Google query used	Hit count estimate
Ready steady innovate!	"Ready steady innovate" filetype:pdf	3
Agricultural widgets in Norfolk	"Agricultural widgets in Norfolk" filetype:pdf	1
Innovative agricultural widget designs	"Innovative agricultural widget designs" filetype:pdf	0
Ready steady innovate!	"Ready steady innovate" filetype:doc	0
Agricultural widgets in Norfolk	"Agricultural widgets in Norfolk" filetype:doc	1
Innovative agricultural widget designs	"Innovative agricultural widget designs" filetype:doc	0

Searches restricted to a single document type, as above, typically generate many fewer results than generic searches. This often gives scope for additional human processing to filter, check and add extra value in the form of context checking. If accurate results are important then the results should be filtered to remove spurious matches and duplicate documents. Spurious matches – results matching the search but not mentioning the desired document – should always be checked for, of course. Duplicates are possible because reports may be circulated and reposted in different places online by readers, if permission is granted to do so. Also, there may be PDF and Word versions of the same document. In a large set of results, duplicates can be identified by listing the titles of the identified documents and using the titles to suggest duplicates (e.g., by sorting alphabetically by title to place common titles together for easier identification).

In addition to general search engines, specialist search services like Google Scholar may also be able to reveal additional online (and even offline) citing documents (Jascó, 2005; Mayr & Walter, 2007; Vaughan & Shaw, 2008). In some cases, it is also possible to search directly for citations in these specialist search services if they contain their own citation index.

Significant extra contextual information can be generated from the search results by building a *bespoke web citation index*. A web citation index is simply a list of the titles and types of each correctly identified citing document, as shown in Table 2.5. This can be built by visiting all the results obtained as in Table 2.4 and extracting the document titles and identifying the document types. This index succinctly summarises the *contexts* in which the reports have been found useful. Note that duplicate elimination can be conducted at the same time as building the web citation index. If necessary, additional columns could be added reporting more information about why each report was cited, but this is more of a content analysis, as described in the section below. Although there are many citation indexes on the web already as part of services like Google Scholar, the purpose of the bespoke index is to summarise relevant information for one particular purpose rather than to generate a widely used searchable resource.

Table 2.5. A web citation index (made-up cited report names) generated by human evaluations of the search results, as produced in Table 2.4.

Citing report title	Citing report type	Cited report
Regional Improvement and Efficiency Strategy for Yorkshire (PDF)	Research report	Ready steady innovate!
Report of the UK parliamentary agricultural select committee September, 2008 (PDF)	Research report	Ready steady innovate!
We know what to do but we don't always do it – aligning policy and practice (PDF)	Conference paper	Ready steady innovate!
Assessment of evidence about the effectiveness of rural development schemes (MS Word)	Research report	Agricultural widgets in Norfolk
England's rural areas: steps to release their economic potential (PDF)	Research report	Agricultural widgets in Norfolk

Content analysis

It can be difficult to interpret or explain the significance of the statistics in a web impact assessment. This is because the variety of reasons why a web page could be created, including negative reasons like spam marketing make it difficult to give a simple explanation of what a count of online mentions really means. This gap can be filled by finding out what kinds of web pages are commonly represented in the results in order to give a general description of what the statistics represent. In consequence, qualitative investigations should always be conducted as part of any web impact assessment, except perhaps when a web citation index is the primary output and the clients wish to read the documents in the citation index. Qualitative information about the web citations can be obtained by visiting a random sample of web pages from the study (i.e., pages matching the keyword searches) and reading them to find out what they are about and the context in which the keywords were mentioned. This is best formalised as a content analysis.

A web impact content analysis is a systematic categorisation of a set of search results based upon human inspection of the contents of the matching URLs. The end result is a set of categories and an estimate for the number of search results fitting each category. The categories themselves can be predetermined, perhaps from categories previously used in a similar exercise, but it is best if the categorisation scheme is implemented flexibly so that it can be expanded if pages appear that do not fit the existing categories well. The object of such an expansion should always be to give additional relevant information about the context of the citations. It is also possible to use an inductive content analysis: starting with no categories at all but grouping similar documents together to start with and then later formalising this into defined categories. This approach is recommended for new types of web impact exercises.

Category choices

The choice of categories for the content analysis should be related to the objective of the web impact assessment. Although it is possible to classify web pages in many different ways (e.g., visual attractiveness, colour scheme, national origins, owning organisation size, main topic, industrial sector, embedded technologies) the categories should primarily be chosen to inform the web impact exercise. In particular, it is likely that the categories will address who created the citing pages or the purpose of the citing pages. When the content analysis is complete, it should be used to complete sentences like: “Document/idea X was mainly mentioned online by A and B” or “Document/idea X was mainly mentioned online *because of* A and B”. If the primary topic of interest is the organisational origins of the online citations, then it is likely that many classifications would include categories for the main organisations represented (e.g., universities, the press, companies, government) as well as a category for individuals (e.g., bloggers, personal home pages, social network profiles). If the primary topic of interest is the citation types then the categories could encompass common document types (e.g., press

stories, academic papers, blog postings) or common topics (e.g., business information, recreational information, geographic information).

The categorisation process can either be informal or formal. A formal categorisation should use a recognised content analysis method (e.g., Neuendorf, 2002) and this is recommended for research to be published in academic outlets. Formal content analysis can be time consuming, however, because it involves extra states to ensure of the validity of the results, such as cross-checking between multiple coders and the use of a formal coding scheme. An informal content analysis, which could be done by one person using intuitive judgements of categories, is appropriate for pilot studies or for situations where a high degree of validity of the results is not essential.

Sampling methods

In theory, a content analysis could include a classification of all pages matching the keyword searches. Nevertheless this comprehensive approach is often likely to be impractical because there are too many pages to classify in the time available. In such cases a sample of an equal number from each keyword search should be used. If possible, this sample should be selected at random from the URLs or sites (depending upon which are being counted – see above) matching each search using a random number generator. A rough-and-ready simple alternative would be to use a systematic sample instead, however. For example if there are 200 matches for a keyword search and the sample for the content analysis has been selected as 10 then starting at the 10th match and taking every 20th result from then on would give a good spread of results and seems unlikely to introduce a systematic bias.

How many sites should be classified if a sampling approach is being used? This is a difficult issue because classification is quite time-consuming and for robust results a large number of pages must be classified. This is particularly true if it is important to distinguish between the proportions in categories for the different keyword searches. In most cases, however, it is sufficient to give an overall approximate proportion of web pages in each category because the purpose of the classification is to inform the analysis of the results rather than to identify statistically significant differences between categories. For a quick investigation, such as for a pilot study, as few as 30 classified pages could be sufficient although 100 is recommended as the normal minimum number. If it is important to distinguish the proportions in categories between different keyword searches then a larger sample size is likely to be needed and a formal approach should also be taken for the classification process itself. In this case, a standard text on content analysis is recommended to give guidelines on the classification process and the minimum number of classifications needed (e.g., Neuendorf, 2002).

Example

The case below illustrates the results of a classification exercise. It is taken from a web impact analysis for publications produced by the UK National Endowment for Science, Technology and the Arts (NESTA) in 2006-7. After extracting a list of web sites mentioning each publication, a random sample of 324 web pages was classified (a maximum of one per web site) and the results are shown below. The classification was created inductively, primarily focussing on the type of *organisation* mentioning the documents. The scheme below was adopted and the results are in Figure 2.1.

- Academic – University or other similar academic institution, including research-only government and non-profit research institutes.
- Press or blogs – Online newspapers and online versions of offline newspapers, unless the newspaper is specific to a company or affiliated to an academic organisation (e.g., a regional research forum). Includes all blogs, whether written by journalists, professionals or the general public.
- Industry – Commercial organisations.
- Government – Government departments and government-funded organisations.

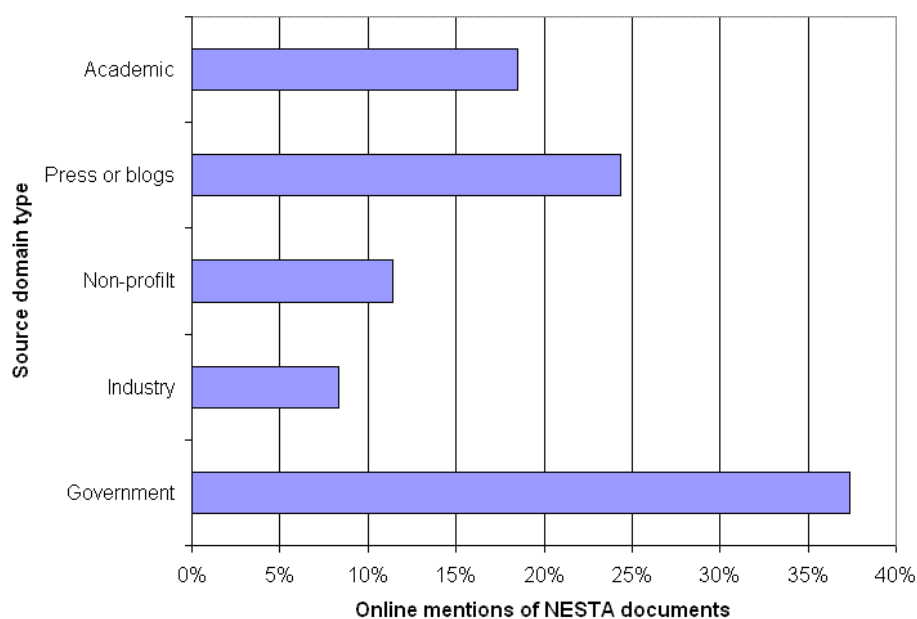


Figure 2.1. Sources of mentions of NESTA publications (324 classified).

The categories in Figure 2.1 were chosen to reveal the organisational sectors that mentioned NESTA publications so that the organisation could understand and develop its audience reach. As a result of Figure 2.1, the web impact statistics for the NESTA reports can be interpreted primarily as evidence of government and press/blogger interest and should not be interpreted as evidence of commercial interest. This is something that would not be obvious without a content analysis.

Note that if the main web impact report statistics for a study are reported on the basis of counting web sites rather than web pages then the classification sample should be randomly selected on the basis of web sites with a maximum of one page classified per site. This was the case for Figure 2.1: a maximum of one web page per site was classified. The random selection of pages on a per-site basis can be awkward, but if Webometric Analyst is used for data collection then appropriate random lists are automatically produced as part of its output (see the chapter on automatic search engine searches).

Validity

An important issue is the validity of the human judgements used to make the classifications. Most web impact assessments have used a single human classifier and this seems appropriate in cases where the purpose of the classification is to give context to the results rather than to accurately distinguish between categories. Nevertheless, it is more reliable to use multiple classifiers using the same scheme and then to assess the extent to which their judgements agree. Information about levels of agreement can then be reported as evidence of the robustness of the classification results. If this level of classification accuracy is needed then a standard content analysis textbook is recommended as a guide for the procedures involved (e.g., Neuendorf, 2002).

As the above discussion illustrates, whilst a simple “quick and dirty” content analysis can be straightforward, a robust and accurate classification is more complex and far more time-consuming, involving many more classifications and the use of multiple classifiers. If a classification is for a formal academic publication and distinguishing between categories is important then the latter type will be necessary. If the classification is for an organisation that does not need, or want to pay for, very accurate results then a simpler exercise is likely to be sufficient. Similarly students practising the technique should start with a simple classification as the time involved with a larger one might not be justifiable.

URL analysis of the spread of results

URL analysis is the extraction of information from the URLs of the results returned by a search engine search for a web impact assessment. This takes advantage of the structure implicit in most URLs, such as the domain name of a URL tending to reflect the organisation hosting the page.

When assessing the impact of documents or ideas online, it can be useful to identify where the impact comes from in geographic terms or which web sites hosted the most mentions. Geographic spread can be estimated to some extent by extracting the top level domains (TLDs) of the URLs or web sites matching the keyword searches. Most TLDs correspond directly to nations so that, for example, if 50% of the matches to a keyword search have .fr domains then they probably originate in France. It is relatively easy to identify and count TLDs in a list of matching URLs or sites and so TLD statistics can be calculated relatively quickly, especially if the process is automated (e.g., by Webometric Analyst). An important drawback, however, is the prevalence of generic TLDs, such as .com sites, which could originate from anywhere in the world. If accurate geographic statistics are needed then a random sample of the generic TLD URLs or sites should be visited to identify their origins.

Table 2.6 illustrates a table of TLDs extracted from a set of search engine results. The table confirms the Anglophone bias that might be expected for a book about a medieval English King (UK, US (edu), Canada, Australia (au), Ireland (ie)) but also contains a potential surprise in the form of six German web sites (de). No significance can be drawn from the com, org, net or info results.

Table 2.6. TLDs of web sites matching the query: "The Reign of King Henry VI" Griffiths.

TLD	Sites	%
com	39	32.0%
uk	25	20.5%
edu	10	8.2%
de, org	8	6.6%
net	6	4.9%
ca	4	3.3%
info, nl, jp, es, au, ie, fr	2	1.6%
ee, cn, pl, mx, nu, it, eu, at	1	0.8%

It can sometimes be useful to identify the *sites* that mention the documents or ideas most frequently. This can be achieved by counting the number of pages represented for each web site in the results. Here web sites can again be equated with full domain names. Whilst this information can be calculated manually in principle, it is easily automated with software like Webometric Analyst. Table 2.7 gives an example extracted from searches for web pages mentioning a particular Persian document. It shows that the site www.pendar.net mentioned the document most often and that many web sites mentioned the document frequently. If such a table is used, then the web sites should be visited to check that the document is not mentioned for a spurious reason, such as advertising or an incorrect match. Note that some search engine result automatically restrict the number of matches per site to two per results page, and if this is the case then the search engine option to disable site collapsing should be selected, if possible, to get the best results.

Table 2.7. Web sites containing most pages mentioning a Persian document.

Web site (domain name)	Pages (URLs)	% of all URLs
www.pendar.net	114	10%
www.dreamlandblog.com	100	9%
www.ahmadnia.net	72	6%
sibestaan.malakut.org	63	5%
www.fasleno.com	55	5%
noqte.com	52	5%
www.mowlanayear.ir	39	3%
www.khabgard.com	25	2%

A similar summary is of the *hosting* web sites that contain the most subdomains mentioning the document or ideas from the queries used. This can be achieved automatically by equating the domain name endings with hosting web sites and counting the number of derivative domain names (subdomains). For example, blogs hosted by www.blogspot.com all have domain names ending with *blogspot.com* (e.g., *webometrics.blogspot.com*) and so counting the number of domain names ending in *blogspot.com* in the results reveals the number of *blogspot* blogs mentioning the documents or ideas from the original query. Table 2.8 illustrates the results from this kind of process for a search for a Persian document name. The table shows that the Persian blogging service *blogfa.com* hosts the most blogs mentioning the document but that several other blog services, including international sites like *wordpress.com* and *blogspot.com* also host blogs.

Table 2.8. Domains containing most blogs mentioning a Persian document.

Hosting web site	Contained web sites (domain names)	% of all domains
blogfa.com	84	18%
blogspot.com	42	9%
wordpress.com	19	4%
blogsky.com	11	2%
blogfa.ir	10	2%
malakut.org	6	1%
ketablog.com	5	1%

Web citation analysis – An information science application

This section describes web impact assessment research addressing a specific type of problem within the academic discipline of information science. This example illustrates how web impact analyses can drive extensive academic research.

Web citation analysis is a type of web impact assessment that has been developed within the library and information science discipline. It typically involves web impact assessments using simple counts of web mentions for large collections of academic documents, such as journal articles, in conjunction with content analysis of the citations. The assessments tend not to use URL citation indexes because their purpose was theoretical and tend not to use URL analyses of the international spread of results because this is relatively unimportant for the research hypotheses.

Web citation analysis was pioneered by Vaughan and Shaw (2003) as an online compliment to traditional offline citation analysis. The latter focused on counting the number of times an article had been cited in other academic articles (e.g., mentioned in reference lists). This figure is widely used as the indicator of the importance of an article (Moed, 2005). Historically, (offline) citation analysis has mainly used databases of citations collated and maintained by the Institute for Scientific Information (ISI, now Thomson-Reuters). Previous authors had realised that search engines could be used to discover how often any article had

been mentioned on the web but had applied this to whole journals rather than individual articles (Smith, 1999).

Vaughan and Shaw (2003) used a new approach, extracting long lists of article titles from library and information science academic journals and comparing how often they were cited in ISI citation databases with how often they were cited on the web. They found that web citations were more numerous but correlated significantly with ISI citations and that just under half seemed to directly reflect scholarly or educational impact. As a result, web citations could be used as a replacement for ISI citations, for example for journals not covered by the ISI. Nevertheless, there were some disadvantages: some web citations were for relatively trivial reasons, such as online tables of contents for journals. In addition, constructing the searches was time-consuming. A follow-up study of biology, genetics, medicine, and multidisciplinary sciences gave similar results, but a lower proportion of citations reflecting scholarly or intellectual impact (Vaughan & Shaw, 2005). A later study, of citations to the publications of library and information science faculty, found a smaller proportion of scholarly or educational impact citations, and almost 40% of the web citations came from lists of articles, but confirmed that there were many more web citations per publication than citations recorded in the Web of Science academic citation databases. In contrast, Google Scholar citations almost exclusively reflected intellectual impact (92%) and were also more numerous than Web of Science citations, although less numerous than web citations (Vaughan & Shaw, 2008).

A number of other studies have also taken large samples of academic journal articles and evaluated the number of online citations to each, using a range of different techniques to identify citations. For example, one investigation coined the phrase "Google Web/URL citations" to encompass citations to either the title or URL of online articles. In a comparison with Web of Science citations for open access journals in eight science and social science disciplines, the web citations were found to be relatively numerous and correlated with Web of Science citations in all disciplines except psychology. This study also confirmed that Google Scholar was a good source of citations, correlating significantly with Web of Science citations (Kousha & Thelwall, 2007).

In summary, a significant body of research has developed that is concerned with using web impact analyses for the important but narrow task of developing methods to identify the online impact of academic journal articles. This shows the potential for web impact analyses to underpin research into specific online issues.

Advanced web impact studies

Advanced web impact studies are those that employ additional techniques beyond those described above. One way that this can be achieved is by modifying the basic keyword searches used to reduce their scope. For instance, one study took a list of searches for 70,700 academic journal articles and modified them by adding the text `syllabus OR "reading list"` to change them into searches designed to match online syllabuses. The purpose of the study was to assess the extent to which recent academic journal articles were mentioned in course syllabuses in a range of disciplines (Kousha & Thelwall, 2008). For this kind of web impact study no content analysis is needed because the web pages matching the search are already tightly defined as academic syllabuses. The research found evidence that there were large disciplinary differences in the extent to which academic research was used in academic teaching. Such a conclusion would have taken much more time to reach using other methods such as questionnaires sent to a random sample of educators in various disciplines.

Summary

Web impact analyses are based on counting and analysing the URLs returned by search engines in response to queries designed to match documents or ideas. The purpose of the exercise is to produce indicators of the extent to which the documents or ideas are mentioned online, either to directly assess web impact or to indirectly assess predominantly offline impact through measurement of the online component of that impact. Web impact assessment

is relatively easy to conduct in a simple “quick and dirty” manner and in this form is useful to give a broad overview of the impact of a set of documents or ideas. Web impact analyses are more time-consuming to conduct in a robust manner, however, because of the increased human effort involved in conducting high quality content analyses. Content analysis is normally an important component of web impact analyses because of the need to interpret the figures produced. Content analysis is most needed for documents and ideas that have a broad appeal and least necessary for highly specialised documents or searches that are mentioned online in predictable specialised searches, as with the syllabus searches discussed above and probably with most academic publication web impact studies.

5. (L4) Web Network Thick Description [new]

Introduction

The rise of the web has led to the emergence of new phenomena like social network sites and the migration of old phenomena to the web, such as online political campaigning. This has spawned many research projects as academics attempt to understand the implications of each new phenomenon. As part of this, scholars have articulated new online research methods or have assessed the implications of moving existing research methods (e.g., surveys) online. Studies of web phenomena may seek to develop new theory using a method that supports this, such as grounded theory, or may seek to test existing theories in a new web context. It seems, however, that there is space for methods that are explicitly descriptive and exploratory when new web networks emerge. This is because new web phenomena are often socially complex and technically awkward to investigate and so there is space for descriptive research that can serve as a starting point for research.

Although researchers investigate native web phenomena and the online reflections of offline phenomena there are no general strategies for simple descriptive investigations of new online networks. In response, this article describes two general and flexible methods to investigate topics on the web through a combination of quantitative and qualitative approaches. The hybrid Web Network Thick Description (WNTD) is for web networks with known offline characteristics or counterparts. The pure WNTD is for web networks without known offline components. WNTDs have an emphasis on being descriptive and exploratory rather than assessing existing theory. The methods are illustrated, but not evaluated, with an example of a Twitter network and an example of a hyperlink network.

Web Network Thick Descriptions: A rationale

The basic idea for a WNTD is to use multiple methods, including both qualitative and quantitative, to give a more rounded picture of the topic from its web reflection than would be gained from a single method or by focusing on testing a theory-driven hypothesis. Whilst pre-web thick descriptions were qualitative, a combined qualitative-quantitative approach is appropriate for the web because quantitative methods are relatively easy to apply to the web in order to extract topic-related patterns, as the list below shows. Qualitative methods are also needed to concretely connect quantitative patterns to the topic so that they are not interpreted in a misleading way by making simplifying assumptions that are not valid (e.g., that hyperlinks between university departments reflect some kind of collaboration). WNTDs are thus a type of mixed methods research.

In anthropology and sociology the term thick description has been used for the strategy of describing not only the phenomenon investigated but also its context so that the reader has a better idea of the meaning of the activities or behaviours described. The term has been previously introduced into information science in the context of bibliometrics. WNTDs are similar in that they are essentially webometric methods enriched by qualitative contextual information and applied to online issues. The issues can either be issues in the traditional sense, such as social or political issues, or can be anything that is describable through a set of websites matching a theme.

It is important to recognise that a WNTD, like most web-based studies, is likely to have serious limitations. The reason is that web content is not systematically regulated or subject to strong behavioural imperatives and so information derived from the web is likely to be partial, with gaps and with anomalies of various kinds. For example, if the purpose is to investigate collaboration between psychology departments through hyperlinks between their web sites then problems may be caused by departments that do not link to those that they collaborate with and by departments that link to all other departments, irrespective of whether they collaborate or not. Similarly, if this collaboration is investigated by identifying whether the members of different departments follow each other on Twitter then there can be gaps in the data due to members that are not active Twitter users or who do not follow others or anomalies due to members who follow others indiscriminately. As a result of this, any

investigation focusing on a single data source, such as a WNTD, is likely to produce biased and partial findings and hence the results should be interpreted cautiously. This issue could be avoided to some extent by using multiple WNTDs with different data sources (e.g., tweets and links) in the cases where this is possible.

Methods overview

In order to standardise the terminology, minimum requirements are described for an investigation to count as a WNTD, although any WNTD can include additional components. The rationale for partially prescribing the methods to be used and giving a specific name to the approach is to provide guidance for future studies in the form of a minimum requirement for an effective analysis and, in parallel, to promote testing of this minimum requirement so that it can be replaced with an alternative, if necessary, with the same ultimate goal: to provide guidance for an effective web analysis of a topic. The descriptions below mention the free software *Webometric Analyst* when it might not be clear how to apply a method.

WNTDs explicitly do not use theory but focus instead on describing the phenomenon investigated. Hence they are suitable to investigate topics that have not previously been examined in detail. As part of the descriptive remit, they do not need theory driven hypotheses or research questions driven by prior theory and concerning what to expect from the analysis (unlike most social sciences articles). Nevertheless, they can have such theory-driven hypotheses or research questions if they are expressed in a general way and do not obstruct the primary aim of finding out about the topic. In addition, prior theory can be harnessed in the analysis stage to help explain any patterns found in the networks or to generalise the findings.

Pure WNTD

A pure WNTD investigates a topic network online but does not attempt to systematically connect it to any similar offline network or information, either because no similar offline network exists or because it is difficult to find information about the offline network. A hybrid WNTD includes all of the components of a pure WNTD but some are modified and there are additional components.

Data type choice: Data for a WNTD can be of any web-based type that can be used to imply online connections of any kind between web objects. Data must be a (node, connection) pair. Examples include (tweeters, tweets targeted with @ or friendship connections), (web sites or web pages, hyperlinks or URL citations or web mentions), and (YouTube users, Subscriber connections or reply-to implicit connections). The data can also be indirect connections, such as co-mentions. For example a co-inlink connection between two web sites A and B is a third web site C that links to both A and B. Similarly, in a network in which the nodes are words, the strength of co-word connection between two words A and B may be counted from the number of texts that mention both words A and B.

Data set selection: Any coherent method can be used to identify the set of nodes to be investigated and to get the data for the connections. The network should have a specific scope to enable the connections within the network to be as comprehensive as possible. Any important omissions from the network will undermine the results. Methods to create node lists include online searches with a specific scope (e.g., tweeters tweeting a link to a specific journal article, following a specific conference or mentioning a topic-relevant keyword), authoritative online directories, and ad-hoc combinations of these methods. In the latter case, the methods should be as systematic as possible to avoid biases.

Node importance ranking: If the network is directed, then nodes should be ranked in order of importance using indegree centrality. The indegree centrality of a node is simply a count of the number of connections pointing to it. If the network is undirected, then importance ranking should be achieved using degree centrality instead. The degree centrality of a node is the number of connections that it has. These are the most commonly used importance measures for nodes in a network. Alternative centrality measures can also be used, if a justification can be provided. In particular, outlink centrality may be important for a directed

network to identify the nodes sending the most connections and betweenness centrality may be important for directed and undirected networks to identify nodes that occupy strategic positions between clusters. The most central nodes should then be identified and discussed.

Network clustering: Network clustering splits the nodes of a network into a set of groups such that the nodes within each group tend to connect more to each other than to nodes outside of the group. There are many different methods for splitting a network into clusters (or communities) and there is no single agreed best method. A simple method is visual identification of clusters through the network diagram but it is better to use a recognised automatic clustering method, such as community detection, social network analysis methods or multi-dimensional scaling. For example, the Webometric Analyst Partition menu contains two similar clustering algorithms that put the nodes into clusters and colours the nodes according to their cluster. Once the network has been clustered, the clusters should be investigated in order to assess whether they are coherent and named, if possible to summarise their meanings. A name should match the nodes inside a cluster and not match nodes outside the cluster, as far as possible. Once the clusters have been named, the overall pattern created by the clusters can be discussed as well as any significant connections between clusters. Illustrating the patterns between clusters can be helped by merging all the nodes in each cluster together in a new version of the network showing inter-cluster connections (in Webometric Analyst, this is achieved with the option to merge nodes with the same colour).

Content analysis of connections: A content analysis of a random sample of the online connections should be conducted in order to aid the interpretation all the results. A content analysis is time-consuming but essential. This content analysis is a human-based categorisation of a random sample of 100-400 nodes using categories that are relevant to the topic. It also is possible to use multiple unrelated sets of categories to make a faceted classification scheme. A content analysis takes time to do well because it involves devising and describing effectively a relevant set of categories, training 2-3 human coders and evaluating the level of inter-coder agreement. A sufficiently high inter-coder agreement is needed to justify the categories as being robust. The results of the content analysis, typically represented as a table or graph are useful to give insights into how and why the issue is mentioned or linked to online, as well as how the results of the other methods should be interpreted (e.g., the node importance ranking and the clusters).

Qualitative investigations of outliers, clusters and important nodes: The results of all the analyses need to be tied together by investigating qualitatively and in more detail the key aspects of the network identified. This stage should look for reasons why outliers exist in the network (e.g., disconnected nodes), why the important nodes are important, and why the clusters found exist. This step involves reading the connection data carefully, formulating hypotheses and looking for additional evidence to back up these hypotheses. This part should typically form the discussion section of a write-up.

Summary: The summary should bring together the findings and give an overview description of the online network analysed.

Hybrid WNTD

A hybrid WNTD investigates a topic online but also attempts to systematically relate it to a similar offline network. A hybrid WNTD is essentially an extended version of the pure WNTD but some of the pure WNTD parts are slightly modified to incorporate offline information.

Data type choice: Same as for Pure WNTD.

Data set selection (modification of pure WNTD): For a hybrid WNTD, the nodes in the data set must be something with an offline reflection for the online-offline component to be used. For instance, if the online nodes are tweeters then offline nodes might be the people behind the Twitter accounts used. The offline connections investigated do not have to be the same as the online connections but there must be some intuitive or theoretically driven reason why the online and offline connections are worthy to be compared. For example, if the online

connections are tweets between academics then the offline connections could be any kind of connection between academics, such as citations or co-authorships.

Comparison of online with offline node importance ranking (extension to node importance ranking in pure WNTD): After calculating node importance ranks, as described in the pure WNTD section, the results should be compared with a relevant offline ranking. There should be a theoretical justification as to why any chosen offline ranking is relevant. The Spearman rank correlation coefficient should be used for the comparison because web data is typically skewed. A positive result can be taken as evidence of a relationship between the online and offline phenomena but not necessarily cause-and-effect.

Examination of links between and within offline node categories (addition to pure WNTD or replacement for network clustering): If the links in a network reflect the categories to some extent then the links will tend to be within categories rather than between categories. A simple test for this is modularity (e.g., available in the Webometric Analyst Stats menu), with a positive value indicating that links tend to occur within categories at a higher rate than predicted by chance. This can be used as a test for a connection between the online and offline patterns. Links within categories (i.e., links between pairs of nodes within the same category) could also be qualitatively compared to links between categories (i.e., links between pairs of nodes in two different categories) to see whether they are different types but this is unlikely to give useful results and so it only recommended if it seems that there are likely to be substantial differences in link types between the two. This can be achieved formally with a content analysis of the two groups or informally by examining a selection and looking for patterns in the differences.

Content analysis of connections: Same as for Pure WNTD.

Qualitative investigations of outliers, clusters and important nodes in terms of offline properties (modification of outlier investigation in pure WNTD): This section is as for the pure online qualitative investigations except that offline properties may also be investigated to look for evidence of the reasons behind the identified key aspects of the network.

Summary (modification of summary in pure WNTD): The summary should bring together the findings and give an overview description of the online network analysed in comparison to what is known about the offline network.

Technical requirements for a web network thick description project.

The following core requirements for a WNTD Project are minimum technical needs for a network to be creatable, at least in theory. These are formulated for a researcher-specified list of the entities of interest.

- The entities must all have a web presence of some kind and the web presences must be of the same type. For example, all the entities might have a personal web page, a web site, or a profile in the same social web site, such as Twitter, YouTube or Mendeley.
- The web presences must be reasonably likely to connect together in a recognised implicit or explicit way. For example, if the web presences are home pages or web sites then the connections could be hyperlinks, title mentions, URL citations, linked title mentions or co-variants of these, such as URL citations. If the web presences are social web profiles then the connections could be friend, subscriber or follower connections. They could also be implicit connections, such as reply-to: A is connected to B if A replied to any of B's comments. Again co-variants of these are also possible connection types, such as co-subscription (the number of profiles that A and B both subscribe to). Whatever the connections are they must be frequent enough to generate a network that is rich enough to analyse. The connections can be weighted (i.e. with a numerical strength associated), or binary and can be directed or undirected.
- It must be possible and practical to identify the web presences and connections for the whole network, either automatically or manually. Automatic mechanisms might involve a web crawler finding hyperlinks between a set of web sites or software like

Webometric Analyst getting follower connections from a social network site like Twitter or YouTube.

These technical requirements ensure that it is possible to conduct a WNTD Study but do not guarantee that the results will be useful or meaningful. An assessment of the value of the results should be part of the final discussion or conclusions of any write-up. A possible and valid outcome of a WNTD is therefore a conclusion that the network analysed is not meaningful. In this case the results may serve as a warning to others than the network is not useful to analyse.

Examples of possible WNTD Studies

This section gives some ideas about types of WNTD study that may be possible.

Business: The scope could be a collection of businesses within a specific sector and perhaps also within a single country or other defined region. The entities could be the web presences of any single type, such as company websites, official Twitter accounts and official YouTube channels. An indirect connection type is likely to be needed since the organisations are presumably competing and ignoring their competitors on line. Due to the indirect connections, the results are likely to reflect others' perceptions of, or interactions with, the organisations rather than their interactions with each other. For example, some studies of business web sites from different sectors by Liwen Vaughan have aimed to cluster them by sector as the outcome of the research.

Politics: The scope could be the politicians that are members of a specific national or regional government. The entities could be their personal websites, blogs, or Twitter accounts. These would need to be checked to ensure that they were genuine rather than echoing unmodified party content. If not all politicians had a presence of the type investigated then it may still be useful to analyse them but the results should be more cautiously translated to an offline context. Presumably, the networks generated would primarily reflect party affiliations but the links between parties may be interesting as well as clusters within parties.

Distributed Organisations: The scope could be all individual groups that are members of a larger organisation, such as Greenpeace, the Slow Food movement, the International Trade Union Confederation, Oxfam or the Hari Krishna Organisation. The web presences and connections could be any of the types previously mentioned and the results may shed light on how the organisation is structured, how it grew, or on the existence of likeminded or co-operating clusters within it.

Issues: The scope could be all groups that contribute to the debate on a specific issue, such as nuclear power, nuclear disarmament, alternative energy, action against poverty, water resource management, vegetarianism, cruelty to animals, decriminalising cannabis, anti-racism, women's rights, LGBT equality or on a specific piece of legislation. If the organisations involved are not focused on a single issue then it is important that the connections between them reflect this issue rather than anything else. This may need a modification to the data collection to remove irrelevant connections. For instance, a simple approach would be to remove all connections not containing a mention of a keyword related to the issue (e.g. if using hyperlinks or web sites concerning the nuclear power issue then hyperlinks might be removed if the originating page did not contain the phrase "nuclear power"). If this extra step is not taken then the content analysis results are likely to include many irrelevant items. The work of Richard Rogers is particularly relevant to analysing networks of issues on the web.

Academic Fields: The scope could be all individual academics, research groups or departments that research a specific topic, field or discipline. The results may reveal collaboration structures or intellectual structures in the sense of similar topics researched. Since academics seem not to use social media uniformly, the analysis seems to be likely to be most complete if the web presences used are personal or organisational web sites. Nevertheless, it could also be useful to analyse the subset of active users of a particular social media, such as Twitter, to focus on social media use rather than the topic itself.

Hobby Organisations: The scope could be all organisations participating in a specific hobby or craft within a country or region such as Knitting, Model Aircraft, Construction, Amateur Football, World of Warcraft, Amateur Dramatics or Allotment Farming. It seems likely that most hobby organisations would not have an online web presence and so the analysis could only reflect the online active subset, which must be acknowledged as a limitation.

Example 1: Twitter networks

Twitter is a natural environment to study networks in because it is widely used for communication and is public, except for users that protect their accounts. Moreover, at the time of writing it was possible to automatically harvest tweets by searching for them, including tweets to and/or from specified users using free software like Webometric Analyst. Hence, given a topic that is tweeted about, it seems reasonable to either investigate the network formed by the known key actors for that topic or to identify and investigate the network of the most active users.

Network of pre-existing key topic actors in Twitter

To create a Twitter network for a pre-existing collection of key people or organisations for a topic, the first and most important stage is to identify relevant Twitter accounts. Depending upon the topic, these might be the accounts of individuals or the official accounts of organisations. These may be found by searching the web sites of organisations or searching Twitter itself. This step is important because all the data collected from Twitter relies upon it and if any one of the accounts is wrong or missing then the network could be wrong. Not all people have Twitter accounts and so there may be unavoidable gaps in the network as a result. Once the list is complete, it can be submitted to Webometric Analyst or other software as a list of terms to monitor¹. Monitoring the names will return both tweets from the users and to the users. Since Twitter searches may return only tweets from the previous two weeks, if the users are not very active then it may be necessary to continue the monitoring for several weeks or months – or throughout a specific event that is likely to cause a lot of activity - in order to get enough data.

Once the Twitter monitoring is complete, software like Webometric Analyst can turn the Tweets gathered into two networks: a direct mention network and a co-mention network. The nodes in both networks are the users monitored. The direct mention network contains an arrow from user A to user B if user A sent tweets to user B during the time monitored. The arrow thickness will be proportional to the number of tweets sent and clicking on the source or target user node in Webometric Analyst will reveal a list of the tweets sent or received. In contrast, the co-mention network contains a line between two nodes if a Tweet was simultaneously sent to both of them.

Topic-based Twitter users network

To create a network to represent discussion on a topic on the web in a situation where the key actors are unknown and too difficult to discover without a large analysis of Twitter, a project must start by building up a list of keywords to search Twitter for in order to identify topic-relevant tweets. This is the most important step of the analysis and takes time to do well. The keywords should be found by brainstorming for relevant terms and then testing them using search.twitter.com. There may also be widely used hashtags relevant to the topic and these should be added to the term list. The selected terms should be narrowly focused so that almost all tweets mentioning them are topic relevant otherwise the network created may not relate properly to the topic. This may mean discarding terms that are relevant but that are also used in other contexts. For example, a study of soccer could not use the term football since it is also used to describe American football. In addition, the terms should be used at least a little in Twitter otherwise there would not be any point monitoring Twitter for them. If it is difficult to find a few highly relevant and popular terms that get many matches then it would be OK to

¹ <http://lexiurl.wlv.ac.uk/searcher/twitterNetworks.html>

create a long list of terms that are rarely used but are highly relevant. Once the list has been compiled and checked, it can be monitored for a period of time using Webometric Analyst.

Once the monitoring period is complete and enough data has been collected, Webometric Analyst can be used to create a list of all the users tweeting and tweeted to in the collected data and this list can be used to identify the top 50 users to create a network for. For example, these might be the most active tweeters, the people receiving the most tweets, or some combination of the two. The list should be manually filtered to remove spam or otherwise irrelevant accounts. This list should then be fed into Webometric Analyst or other software to create direct tweet and co-mention networks from the tweets already collected.

A pure WNTD is relevant for the network generated because the nodes are unknown and nothing may be known about the users behind them. The analysis is the same as for the pre-existing key topic actors in Twitter network, except for the comparison with pre-existing categories, which will not exist. Instead a method should be used to cluster the network, such as the automatic clustering routine in Webometric Analyst, and the analysis described above conducted on these clusters. In addition, an attempt should be made to name the clusters by identifying what the users have in common in each cluster or why they form a cluster.

Example: Influential Twitter climate accounts

This example is a brief examination of the network formed by a UK-centred collection of Twitter users that are influential for climate change, using a list of 55 published in the UK Guardian newspaper². A focus on investigating this source allows the analysis to bypass the key first stage of finding relevant accounts to follow. It also allows the user names to be published and not anonymised because they have all previously been published in a national newspaper, which would give them much more publicity than an academic article and so there is no risk of an invasion of privacy by publishing their names.

Since many of these accounts were active, they could be monitored over a short period of time and still yield meaningful networks. They were monitored for one hour, giving tweets to and from all the accounts for the previous two weeks due to the two week cut-off for Twitter searches.

A content analysis of a small random sample of 100 tweets showed that 38% were retweets (RT or MT) and the complete set fits into the categories below, suggesting that tweets associated with users in the Guardian list are used to send facts and to encourage or support climate change activism.

- 30%: Fact or news related to the environment, the climate or climate change.
- 24%: Encouragement or support of climate change activism.
- 3%: Opposing the climate change argument.
- 34% irrelevant, 6% unclear and 3% Spam.

Table 5.1 suggests that the most important user in the network is James_BG (James Murray, BusinessGreen.com editor and in the News category). He connects to six other users in the network. The second most important is probably ClimateRealists (in the Opponents category), because was targeted by 4 tweets even though it only sent tweets to one other user. The owner of this site, a climate change sceptic who believes that recent climate changes are not man-made, seems to have become less active recently according to his blog but seems to be targeted or referenced by other tweeters with arguments (e.g., "@PlanetSpeaking: @ClimateRealists What a confusionist you are! Who claims 'low CO2 melts ice'? Cited study doesn't. @michaelemann" and "@CO2Insanity: @ClimateRealists Niagra Falls frozen in 1911 must have been that global warming again.").

² <http://www.guardian.co.uk/environment/blog/2010/may/11/top-50-twitter-climate-accounts>

Table 5.1. Centrality scores for the 19 climate users with non-zero degrees.

User	About	In-degree	Out-degree	Degree
James_BG	BusinessGreen editor.	3	3	6
ClimateRealists	solar variation causes climate change blog	1	4	5
kate_sheppard	US energy and climate change blogger	3	2	4
ClimateGroup	Business climate news	3	2	4
DECCgovuk	UK Department of Energy and Climate Change	3	1	4
350	Campaign to reduce CO2 to 350 parts per million	3	1	4
worldresources	Climate change policy, economics and science	2	3	4
billmckibben	350.org climate campaign founder	2	2	4
energyaction	Youth clean energy coalition	2	2	4
grist	US green news and comment	2	2	3
metoffice	UK Meteorological Office	2	1	3
globalactplan	Global Action Plan UK environmental charity	1	2	3
wwwfoecouk	UK Friends of the Earth	1	2	3
the_ecologist	UK magazine	1	2	3
Climatecamp	Direct action on climate change	1	2	3
peopleandplanet	UK student environment group	1	2	3
tcktcktck	Campaign for a global legally binding climate agreement.	1	1	2
algore	US climate change politician Al Gore	1	0	1
EPAgov	US Environmental Protection Agency	1	0	1

The directed network in Figure 5.1 shows that the newspaper categories do not match the network well with the exception of the campaign groups. The 6 campaign groups form one interconnected group, with the exception of the Climate Group. The only other cluster that is connected is that of the two linked bloggers, *grist* and *kate_sheppard*. *The lack of correspondence between the categories and the connections suggests that the climate change issue tends not to be organised by sector, at least in Twitter, but that cross-sector links are common and important.*

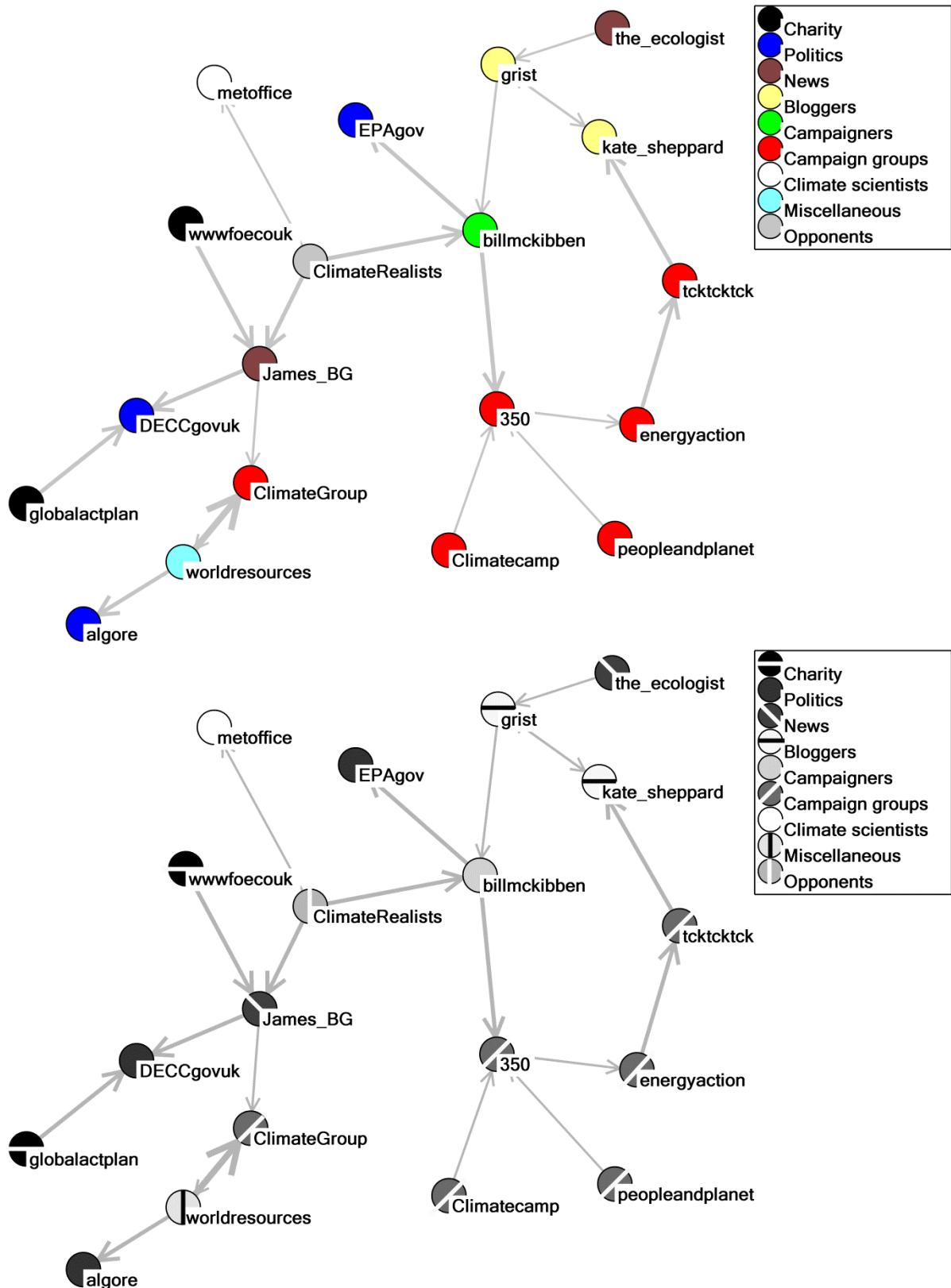


Figure 5.1. A network of tweeting between the 55 climate-related tweeters, with disconnected tweeters removed. Nodes are coloured (top diagram) or shaded (bottom) by the categories assigned by the Guardian newspaper.

In summary, using data collected during the period examined 19 of the 55 Guardian-recommended tweeters formed a single connected but sparse network, based upon the @addresses of the tweets sent. The BusinessGreen editor James_BG was the most connected

within the network, tweeting three others and being tweeted by three others. Second was the Climate Realists anti-global warming campaign, mainly due to its tweeting of others. Four other tweeters were sent messages by at least three others within the network. The different types of node within the network interconnected with each other, with the exception of the campaign groups, which tended not to connect to other types of organisation and formed a coherent cluster. In this case the links within the campaign groups cluster tended to be to share news information and ideas, and the links between the cluster and other nodes were similarly sharing information so there was no apparent difference between the two types of link.

Example 2: Web link or mention networks

Networks of websites based upon the links between them have been extensively studied before and are a natural source for WNTD studies. Despite the rise of the social web, traditional websites are still important for organisations and individuals to create a presence on the web and to impart information about themselves, as well as for publishing various kinds of information and for marketing purposes. In some cases related individuals and organisations might naturally hyperlink to each other as a result of collaborations, connections or information use. This seems to be particularly common in academia due to the scholarly culture of collaboration and free information sharing, and also in politics due to the need to show affiliations and to debate with opponents. Hence hyperlink networks seem likely to show interesting patterns in academia and politics. In contrast, business web sites are unlikely to connect to other sites and especially not to competitors but to focus instead on marketing the business itself. Hence a hyperlink network of businesses within a specific sector would be likely to be sparse and uninteresting. For collections of websites that do not naturally interlink, the logical alternative is to use co-inlink connections instead of direct links. Recall that a co-inlink between two websites A and B is a third website C that links to both A and B. Pairs of websites with many co-inlinks are likely to be similar or related in some way and so the patterns in a co-inlink network may be interesting.

Direct link networks can be created in two ways. The first way is time consuming and only possible for small or medium sized websites (e.g., blogs, university departments or research groups, individual politicians' sites) rather than large sites (e.g., universities, news web sites, Wikipedia). This method is to crawl the sites using a web crawler that is capable of creating networks from the links between the sites, such as SocSciBot (socscibot.wlv.ac.uk). The second, and quicker, way to create direct link networks is using webometric software that harnesses the power of commercial search engines. It is not possible to create hyperlink networks from the current major search engines because they do not allow searches for hyperlinks anymore. Nevertheless, there are three inter-document connections that can be identified from searches in commercial search engines: URL citations are mentions of the URL of one website in another website; title mentions are mentions of the name of one website in another website; and linked title mentions are mentions of the name of one website in another website associated with a hyperlink. These can all be searched for automatically and converted into a network using the Webometric Analyst software; only a list of websites and their names is needed³.

Indirect link networks based upon co-inlinks cannot be created using SocSciBot but can be created using from commercial search engine searches using Webometric Analyst. As with direct link networks, URL citations, title mentions or direct title mentions should be used.

The first stage in either a pure or hybrid WNTD is to create a list of website to investigate the links between. The ideal size for this list is 50 – many more sites than 50 will create a cluttered network diagram and many fewer than 50 may result in too little information to usefully analyse. An authoritative source of the sites, such as on a website of an umbrella organisation or published in the press, would be perfect. If this is not available, then it is important to spend time attempting to create a comprehensive and accurate list from

³ <http://lexiurl.wlv.ac.uk/searcher/usingWebometricAnalyst.html>

searching the web. If the list contains either category information or ranking information for the websites then a hybrid WNTD can be conducted, otherwise a pure WNTD should be used.

There are many different options for creating a link network, as described above (direct links or co-inlinks; using hyperlinks, URL citations, title mentions or linked title mentions; with SocSciBot or Webometric Analyst) and so it may be necessary to try different methods to see which gives the best network to analyse. Direct links are normally easier to analyse than co-inlinks since they are more direct, so these should be used if possible. The best network of any type is likely to be the one with the most data (i.e., the most links) and so the direct link network with the most links should be selected unless it has too many Spam or irrelevant connections to give useful information.

Example: US mathematics schools

This example is a hybrid WNTD of the links between US mathematics schools, using the US News & World Report rankings and website addresses⁴. These schools extensively interlinked so it was possible to use direct link networks. In fact the main problem was to reduce the connections in the network to a level that would make it possible to extract patterns. Hence the network with the fewest connections (linked title mentions) was chosen as the one to base the analysis on and the weakest 75% of the connections were removed. The resulting network (Figure 5.2) was based upon the strongest 25% of the linked title mention connections and is therefore a network of strong web connections between US mathematics departments.

Linked title mentions correlated significantly with rankings from the US News & World Report website (Spearman's rho 0.516, $p=0.000$), suggesting that links were created for reasons related in some way to scholarly or educational activities or excellence. The correlation might also be partly due to the sizes of the schools

Table 5.2. Indegrees for schools with at least 50.

School name	Linked title mention indegree	US&WN Score
Massachusetts Institute of Technology	125	5.0
New York University	105	4.4
University of California Los Angeles	84	4.5
University of California Berkeley	68	4.9
University of California San Diego	61	3.9
Columbia University	53	4.4
Princeton University	52	4.9
University of Washington	50	3.7

A content analysis of a random sample of links in the network gave the results below, which clearly illustrate that the links are overwhelmingly associated with the temporary movement of an academic from one department to another. Hence, the link network is actually a human exchange of ideas or personnel network. In consequence, the patterns may reflect similar disciplines, assuming that speakers and personnel are chosen on the basis of interest in similar fields to those practiced in the host departments but interviews with seminar organisers would be needed to check this.

- 58% Mathematician from one department giving a talk at another.
- 24% Current or future affiliation of a mathematician.
- 10% List of conferences, schools or other departments.
- 8% Other.

⁴ <http://grad-schools.usnews.rankingsandreviews.com/best-graduate-schools/top-science-schools/mathematics-rankings>

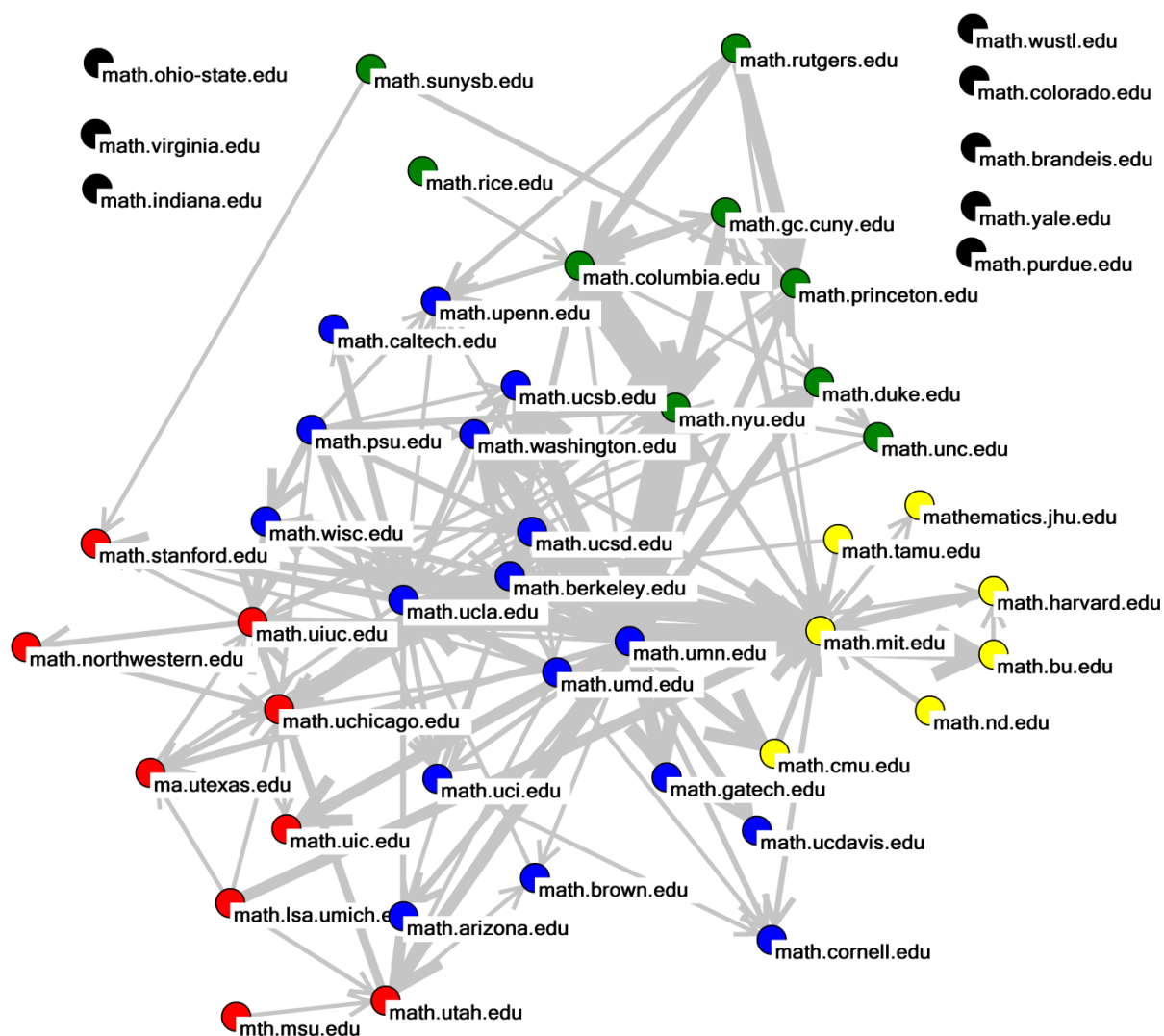


Figure 5.2. A network of title mention links between US mathematics schools, with arrow thicknesses proportional to the number of links and the weakest 75% of the connections removed. Nodes are coloured by the categories assigned by Webometric Analyst clustering.

In summary, the websites of US mathematics schools are highly interconnected. Even after artificially reducing the number of connections between schools, the network was quite dense and the automatically identified clusters highly interconnected with nodes outside of the cluster. Interlinking tended to reflect researcher mobility, either in terms of moving for jobs or giving talks and so it is clear that US mathematics schools form an active and dense virtual college of personal contacts. Nevertheless, prestige is still an important predictor of connections: higher ranked schools tended to receive more links from other maths schools. This suggests that these higher ranked schools also tend to have denser human connections with the rest of the US academic maths community.

Summary

This chapter introduced two new methods, hybrid WNTD and pure WNTD, as frameworks for descriptive analyses of online networks of various types. Both methods contain a set of core components that aim to give a thick description of the network rather than to test specific hypotheses or theories, although theories can be drawn upon as part of the analysis of the results. A WNTD is particularly suitable for a first investigation into a topic or issue on the web: subsequent investigations may benefit from being more focused on a specific aspect of the topic or issue, or on testing a theory relevant to the issue.

The two proposed methods have not been evaluated; instead the article has given ideas about how to use them and two brief examples to suggest how they may be used. Hence, the methods have the status of unverified proposals and will only be validated if they are used in the future and shown capable of producing interesting studies.



6. (T1) Analysing comments and sentiment [new]

- **Key example:** *Creationism vs. Intelligent Design (ID)*.

This chapter is about examining comments and discussions in the Social Web. A text analysis project of this kind may focus on a particular topic of interest, to find out what is being said, where and why. One recent debate is that between creationism and intelligent design. Creationism is the scientific belief that the universe originates in a big bang and progresses without the influence of any external being. ID, in contrast, hypothesises that the world, or aspects of it, are too well designed to have been created without a guiding hand (this does not apply to Heathrow Airport). A researcher studying this debate might be interested in how widespread it is online, in which web sites and countries it appears, what kinds of people argue on both sides, which points are raised during the discussions and where and when it gets most heated. The researcher may also have specific theory-driven hypotheses that social web texts may help them to answer, such as, “Children from religious backgrounds are more likely to display extreme emotions when debating ID than children from non-religious backgrounds”.

Introduction

Comments are a key aspect of the social web. Many social web sites allow users to post short text comments in various places. A person might post a derogatory comment whilst watching a YouTube video, attach some text to praise a Flickr photograph, reply to a controversial blog entry or exchange a few rapid public text messages on a friend’s Facebook wall. This chapter is concerned with analysing social web comments. Whilst most of the methods described could apply to many kinds of internet texts, such as blog posts, the focus is on short informal texts. Without defining this scope precisely, it includes online textual comments posted to specific videos, photographs, blog posts or news stories. It also includes microblogging and text messages distributed or sent publicly to Friends in SNSs. Comments make it possible to study interactions between people and audience reactions to online content. In both of these cases comments seem much easier to study than the equivalent offline phenomenon. For instance, to investigate offline audience reactions to a movie at the cinema, a researcher would need to use time-consuming methods, such as surveys, questionnaires or focus groups. In contrast, popular YouTube video viewers’ reactions may be partly investigated via comments posted to the video.

It is important to recognise that comments and other short texts perform different roles across different platforms. For example, comments may be posted to a media artefact, a person or to a previous comment. In some platforms, comments may be allowed to connect to other comments or link to web pages, but in others only plain text may be allowed. Similarly, some systems allow comments to be rated as good or bad by their readers.

This chapter discusses methods to sample, download and organise collections of comments to study. It also gives brief descriptions of four appropriate methods to research comments: simple statistics, content analysis, grounded theory and sentiment analysis. For any specific project, the source of comments to investigate, the technique for sampling them and the analysis method depends upon the research questions. As always, multiple methods may be appropriate in some cases to make a more complete study of the phenomenon investigated, as will analyses of related content, such as associated videos or images associated with the comments. The final methods section describes sentiment analysis, which is the newest and least written about in social sciences methods texts.

- Content analysis involves manually assorting a sample of comments into researcher-defined categories. It is most suitable for exploratory investigations into new phenomena or contexts.
- Sentiment analysis is a quantitative method to automatically identify sentiment in text. It can be used to identify sentiment patterns in communication. It is best used in conjunction with other methods to provide sentiment-related insights or when directly researching the role of sentiment in the social web.

- Summary statistics are simple quantitative summaries of a comment data set, such as the average length of comments or the average age of commenters. Summary statistics are best used for exploratory research into new phenomena or to give background information to add context to a report that primarily uses another method.

Gathering comments to analyse

A text analysis project must decide which texts to study before the analysis can start. This decision is often not simple to make. It may depend not only on the research questions but also on what is possible with the social web sites investigated. In addition, a decision must be made about how to store or access the texts.

For some projects the choice of texts to analyse is clear from the research question. For instance, an investigation of comments on a particular news story in a specific online newspaper would naturally study all of those comments, unless there were too many. In contrast, a study of religion-related messages in social network sites would not be able to identify all such texts and would need to design a strategy to obtain an appropriate collection to study.

Units of analysis

The first issue is to decide on the basic unit of analysis. In particular, should it be the individual comment or a collection of comments, such as all comments sent by one person or all contributions to a thread within a forum? Typically the smallest unit of analysis consistent with the research questions should be chosen. The reason for this is that larger units of analysis take longer to study and therefore limit the size of the sample that can be analysed. This in turn reduces the power to generalise the results. An example where taking individual comments would be inappropriate is an investigation of the overall structure of discussions in threads. For this, individual comments would be useless and whole threads must be sampled instead. Similarly, if a project focuses on differences between authors then all texts associated with an author could be the basic unit of analysis. Such a project might study a particular set of “units”, i.e., collections of author’s texts.

Basic sampling strategies

Assuming that it is impossible to analyse all texts relevant to the research question, a method must be used to select an appropriately sized sample. Two commonly used methods for this are random and stratified sampling. A random sample is a subset of the full collection created using a random number generator in such a way that all of the data has an equal chance of being selected. If the data is in a spreadsheet, one text per row, then an easy way to achieve this is to add an extra column of random numbers (e.g., in Excel using the random number generator function =RAND()), then sorting the spreadsheet using the column of random numbers and using the top rows as the sample for analysis. For example, if a random sample of 100 is needed then the top 100 rows after sorting could be used.

A second strategy is stratified sampling. This method selects specified numbers of texts with given properties to allow comparisons between them. For instance a gender-related study of YouTube comments may select 50% written by females and 50% written by males even though the majority of YouTube comments have male authors. This stratified sampling can be achieved by splitting the texts into strata and then randomly selecting from each stratum. This may be impossible if the key property for the strata (e.g., gender) is unknown when creating the sample. In such a case the standard random sampling method could be used instead but with an additional step. This step is to identify the key property of each selected text (e.g., author gender) and then place it in the appropriate stratum. When the stratum is full then subsequent texts for it can be rejected. This filtering can continue until all strata have the required sample size.

A problem occurs when the potential set of texts to analyse is too large to identify and number so that it can be sampled using a random number generator. An example of this would be an investigation of comments on news stories in the BBC News web site. There are

too many of these to easily identify and number and so an alternative strategy is needed to identify and number them. This strategy will necessarily be inferior to the methods discussed above, but may still be adequate for the research goals. For instance one strategy is to list all stories in each major category every day for a month and then use a random number generator to select a sample from this set. This has the disadvantage that the results may have a significant number of stories from a single long-running major news event. This problem could be reduced by sampling over a longer period of time, if this is practical. This also illustrates the general point that if it is impossible to get a complete set of texts then it is a good idea to obtain as large a set as possible to sample from. In addition, it is worth giving some thought about how to ensure that the initial set is as representative as possible. In any case it is vital to have a thorough and frank discussion of sampling limitations when writing up the results.

Identifying samples via search engines A quick and dirty way to identify a sample of web pages is to use a little-known feature of modern search engines: their `site: search`. This supports site-specific searches by adding the web site domain name and path after the colon. For instance, to search to pages with comments in the BBC News website (<http://www.bbc.co.uk/news>) the query `site:www.bbc.co.uk/news/ "Your comments"` could be used. This matches pages with URLs containing `www.bbc.co.uk/news/`, i.e., BBC News pages, which contain a comments section. This is quick because Google gives results almost instantly. The results are poor quality for research, however, because the process by which a search engine chooses which results to show is not transparent. This means that the sample has an unknown bias. Nevertheless, this method can be useful for quick pilot studies and also if no alternative is available.

A sampling method that is sometimes used for internet research is the snowball sample. This method starts with a few web pages or sites and then finds other web pages or sites by following links. This method is only suitable in contexts where other methods are not possible because the sample generated is likely to be biased by the starting point or points. Nevertheless, it may allow research in otherwise difficult types of social web environment, and particularly those that are not searchable. If a form of snowball sampling is used then, if possible, a sample that is much larger than needed should first be created and then a smaller sample randomly selected from it. This should help to reduce the bias in the method.

Sampling topics via keywords

Research that investigates a specific topic rather than a small collection of texts has an additional issue to resolve when sampling: how to identify topic-relevant texts. For instance, if investigating religious comments in YouTube then some method must be devised to identify such comments. There is no way to identify all religion-related comments and so a way must be devised to identify a reasonably representative sample.

One logical and general method is to brainstorm for a list of topic-relevant words and then use them to search for texts. This would only find texts including the terms and so is dependent on a good starting set of terms. It will be imperfect because texts can discuss a topic without using any topic-related terms if they are replies, e.g., "I disagree with every statement that you made" could be a reply within any topic. Resuming the example of religion, a keyword list could be constructed by combining the names of the world's major religions and the names given to their followers. For instance the list might start: Christianity, Christian, Islam, Muslim, Hinduism, Hindu... Another approach might be to list words related to god and religion generally such as: god, deity, believer, apostate, atheist, agnostic, deist, religion, religious... The two lists would be likely to generate religion-related texts but with different biases. The former seems likely to produce texts describing or discussing specific religions whereas the latter may generate more discussions about religion itself. Which collection is preferred would depend upon the research questions. This illustrates the subtle biases that keyword lists can introduce.

Example Suppose that a project is concerned with discussions that take place about intelligent design in the comment sections of YouTube videos. The basic unit of analysis for this project would be a single YouTube video and all associated comments. To identify a sample, ideally a complete list of all ID-related YouTube videos should be found and then a sample of the appropriate size drawn randomly from this. Since no such list exists, keyword searches could be used to create an approximation to such a list. A search for “Intelligent Design” in YouTube yields 12,800 videos and it seems that most should be relevant since this is a specific phrase. Since Intelligent Design is sometimes abbreviated to ID, a search for this could be tried, but this gives too many false matches to be useful. Another possible term is *creationism*, but this term yields too many false matches in the sense of pre-intelligent design debates about creationism. Hence, the best strategy seems to be to use the single phrase search “Intelligent Design” alone.

Suppose that a sample of 100 is needed. In fact YouTube only returns 1,000 results per query so the sample of 100 would have to be drawn from these first 1,000 results. This could be done by copying all the results into a spreadsheet, as described above, randomising their order, and picking the top 100. If software is available to automate the downloading, then this would save a lot of time. Human filtering of the results would be needed, however, to weed out false matches, such as music videos (e.g., *Kraftwerk Aerodynamik (Intelligent Design Mix By Hot Chip)*) or videos about engineering designs.

Storage

When investigating a collection of texts a decision must be made about whether to analyse them online or offline. If using offline analysis then a storage medium and format must be chosen. It is strongly recommended to save online texts and analyse them offline, if possible. This is because a web site may close, be changed or clear out some of its pages and this could be a disaster if it occurs midway through a study. Moreover, if web pages are removed after the study then this can prevent future checking and re-analysis. The simplest way to save data is to save the web page containing it from a web browser (often by selecting Save As from its File menu). This has the advantage of preserving the text exactly as a user would see it. It has the disadvantages that the saved files can be awkward to organise due to long file names and large file sizes. A tip is to add a number to the start of each file name and record the key information for the page saved in the file against the file number in a spreadsheet or single document. A second disadvantage of working offline is that saved web pages are slow to load and information about them, such as the results of a content analysis, cannot easily be stored within the saved web pages and hence normally would have to be stored separately. This makes the analysis more complex.

An alternative generic method is to copy the texts investigated from the live web pages into a single spreadsheet or a document. This would be very convenient for future analyses and would also preserve the texts. A spreadsheet is the preferable choice if a content analysis is to be used because it is convenient for recording and totalling the categories used. If using a word processor instead, then a table inside the document is quite a good alternative. The copying method described here is recommended for studies in which the context in which the text occurs is not very relevant.

In addition to the two generic methods described above there are specialist data gathering techniques that can be applied in some cases and specialist software that can save text in a way that supports a particular type of analysis. As an example of the latter, the Nvivo software can store collections of web pages and has functions to aid content analysis and other methodologies. An example of software to aid downloading data is Webometric Analyst (<http://lexiurl.wlv.ac.uk>). This is able to automatically query some social web sites, such as YouTube, and download sets of comments together with information about the commenters. This data is saved in a format that can easily be loaded into any spreadsheet for analysis.

Finally, note that copyright law often allows researchers to make limited copies of copyright material for private study purposes under conditions known as “fair use” or “fair dealing”, as long as it is not republished online.

Graphs and simple statistics

Some text analysis methods produce lists of numbers of categories that need to be reported somehow. This section gives some advice on constructing graphs to display the figures and on calculating basic summary statistics. It does not describe the theory of probability and statistics, which would be needed for a full analysis.

The numbers associated with a text analysis may be measurements derived from the texts or categories for the text authors. Text measurements include: the number of words, sentences or paragraphs in each document, its age (e.g., in hours or days) or more complex measures such as the reading index scores reported by some word processors. Author categories or figures include: age, gender, geographic origin and profession.

Category data for authors or texts can be reported in a table or graph. A graph is normally a better choice if there are enough categories (at least three) because visual information is easier for humans to understand. In particular, it is easier to identify trends in an appropriate graph than in a column of figures. A table may be preferable in cases where the categories need extensive descriptions because these can be awkward to display in a graph.

For category data, the main choice of graphs is a bar chart or a column chart. Bar charts are recommended for data with text descriptions unless they are short. This is because longer text descriptions in a column chart must be vertical to fit in and this makes them more difficult to read. As a general rule pie charts should only be used if one category includes the majority of the data. This is because it is difficult to visually compare the sizes of pie slices if they are similar whereas it is easy to compare similarly sized columns or bars in the main two types of chart. Here is some general basic advice for charts that is forgotten surprisingly often.

- Put meaningful labels on the x-axis and on the y-axis.
- Ensure that all text on the graph is legible.
- Use a clear caption for the graph. This should be self-contained so that it is not necessary to read the text to understand the graph.

For numerical data, such as author ages and text lengths, in order to display it on a graph it is often useful to split it into equal-sized categories. It can then be displayed in a bar or column chart as described above. If the data is in a logical order and the categories have equal sizes then a line graph can be used instead. The choice of intervals to group the data into could be obvious from the data, otherwise it is recommended to use round numbers (e.g., 0-10, 11-20, etc.) or even-sized intervals (e.g., 14-18, 19-23, 24-28, etc.) assuming that 14 is the smallest.

It is often useful to calculate the average of a set of numbers. The normal average, the mean, is not usually useful for numbers derived from text because it can be significantly affected by individual large values. It is reasonable to use it when there are no extreme values, such as for most age data, but not normally for text length data because in many sets of texts a few are much longer than the rest. The safest measure to use is the *median* because it is not affected by extreme values. The median is the middle value if the numbers are listed in increasing order. The means or medians of two data sets or categories can easily be compared to see which has the larger average. To check if any difference is genuinely significant, a statistician or statistical text book should be consulted for an appropriate test (e.g., the Mann-Whitney U test).

Example Figure 6.1 illustrates the range of ages for commenters on an intelligent design YouTube video: “Ken Miller on Intelligent Design” and Table 6.1 shows the reported commenters’ genders. This information might be presented and discussed as part of an analysis of the debates taking place within the comments.

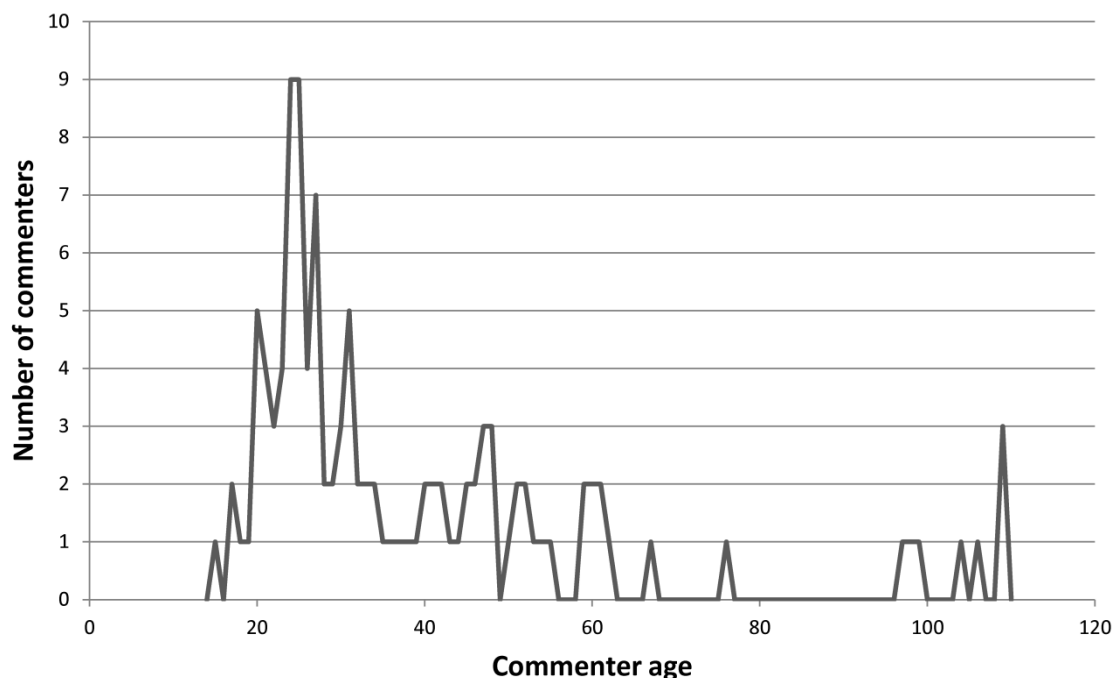


Figure 6.1. Reported ages of the 129 different commenters on the most recent 1000 comments on the YouTube video “Ken Miller on Intelligent Design”, as gathered on 12 October, 2011.

Table 6.1. Reported genders of the 129 different commenters on the most recent 1000 comments on the YouTube video “Ken Miller on Intelligent Design”, as gathered on 12 October, 2011.

Gender	Frequency
Male	109 (84.5%)
Female	11 (8.5%)
Unknown	9 (7.0%)
Total	129

Another common measure is the standard deviation. This measures the dispersion of a set of numbers but it is probably rarely useful for text-related numbers because it is greatly affected by the extreme values that are common in text-related figures. One of the exceptions is again author ages. For instance in some cases it might be interesting to use the standard deviation to compare how spread out the author ages are for two different data sets or categories.

Are statistics alone enough for a social web project? A set of tables and graphs might convey a lot of information but this is not enough for a social sciences project and so should be combined with other more qualitative methods to give additional insights into the topic investigated. Statistics are often more informative if they are comparative – in other words if similar tables and graphs are presented and compared for different aspects of the data (e.g., boys vs. girls in terms of commenting frequency). Purely statistical projects normally use more powerful statistics than tables and graphs, however, such as statistical hypothesis tests. These can be read about in statistics books by anyone interested in this approach.

Creating graphs Although there are many specialist graph drawing packages, for most people a spreadsheet is the easiest graph creator. Normally, a graph can be created by entering the category names and data in adjacent columns of the spreadsheet, then selecting the data and labels before clicking on the graph-creating button. Prompts should then allow a choice of graph types and there will be ways to add labels to the axes. It is a good idea to experiment with spreadsheet graph creation in order to be able to create attractive diagrams. Only plain graph styles should be used so that the report looks professional and the graphs are not confusing or distracting. For example, avoid three-dimensional graphs and strange-shaped columns or bars. The overriding goal should be making the information easy to understand by the reader rather than making the document attractive (sorry).

Summary

- Graphs are useful to summarise category data.
- Tables are less easy to quickly understand than graphs but are recommended in some cases.
- Graphs should be plain, clearly labelled and of an appropriate type.
- The median is the normal measure of central tendency (average) for text-related numbers.

Content analysis

The method of content analysis can be used to investigate a collection of texts to categorise one or more aspects of those texts. The aspects investigated are usually things that are not simple enough to be automatically extracted by a computer program but need human interpretation to identify. These depend upon the research questions and objectives. The result of a content analysis is information about the percentage of texts that fall into two or more categories as decided by one or more human coders. For instance, suppose that a research project had the objective of investigating political discussions in a collection of blogs based in the West Bank: one research question might be: what proportion of blogs support Hamas? This could be answered using content analysis by a human coder reading all of the blogs and categorising them as pro-Hamas or not. Such a content analysis would probably attempt to gain more information by expanding the categories to be more fine-grained (e.g., pro-Hamas, pro-Fatah, pro-Israel, neutral, unclear). It could also be expanded to add other categories relevant to the research objectives, such as blogger type (e.g. political party member, activist, journalist, other) and blog type (e.g., political, personal diary, professional, other).

There seems to be a paradox at the heart of content analysis. Social science issues, theories and research questions are typically complex, so how can a method like content analysis help by giving a simple set of categories and percentages for these categories? The reason is that although content analysis rarely fully resolves a complex social science issue, it can help its resolution in at least the following different ways:

- *Providing background information to help explore a new phenomenon to aid future, more in-depth research.* For example a content analysis of contributions to a particular forum might seek to find the typical age, gender, geographic location and occupation of contributors. This would be a task for a content analysis if the information was not systematically available in a way that a computer program could extract.
- *Identifying key aspects of a new phenomenon by devising appropriate categories.* In new contexts, little may be known about what occurs and what the most important phenomena to study are. In such cases, devising appropriate categories may be the key outcome of a study. A classic example of this is research into blogging genres that categorised blogs into several common classes: personal journal; filter; knowledge log; and mixed format (Herring, Scheidt, Bonus, & Wright, 2004). The background information about the general types of blog in existence helped subsequent researchers studying blogs. In particular, it dispelled the myth that most blogs are about politics, technology or news by showing that most are diary-like personal journals.

- *Testing a theory by categorising its predictions or something closely related to its predictions.* For instance early computer mediated communication researchers believed that text-based messages in informal contexts would be widely misunderstood because they did not encode the affective non-verbal communication that is typical of face-to-face exchanges (e.g., smiling, shrugging). This issue could be partly addressed by coding a sample of informal texts to find out what percentage contained affective expressions of any kind (Walther & Parks, 2002). This would not directly answer the belief but would shed light on how widespread the problem was. As another example, some studies have hypothesised that counting the number of hyperlinks pointing to a university web site would give an indication of its research quality. This hypothesis derived from the belief that hyperlinks typically point more often to high-quality content than to poor-quality content. Hence, universities conducting higher-quality research would have better-quality web content and would attract more hyperlinks. A content analysis addressed this issue by taking a random sample of links targeting university web sites and categorising the content linked to. It found that about 1% of links targeted research outputs but that 90% of links targeted wider academic-related content (Wilkinson, Harries, Thelwall, & Price, 2003). The study concluded, amongst other things, that the hypothesis was false because too few hyperlinks targeted research outputs.

Planning a content analysis

A content analysis needs a set of texts to categorise, a set of categories to use and people to code the text into the categories. The set of texts should include at least 100 to ensure reasonable accuracy in the results. The categorisation process can be time-consuming to do well but more than 100 texts should be classified if particularly accurate results are needed. This might be the case if the category sizes are expected to be quite similar and it is important to discover which categories are genuinely different in size. The classified texts should be the entire set, if possible, otherwise as close as possible to a random sample.

The choice of categories should be driven by the research questions and objectives. For instance, in a study about gender and linguistic style then the categories should include author gender (unless this is obvious and already known) and one or more sets of categories related to linguistic style. Hence there are two decisions to be made: how many independent sets of categories should be used and what should the classes be in each set. It is difficult to give advice on these, because the answers are study-specific, but there is a trade-off between having too many sets of categories to be able to code in a reasonable time, and having too few categories to give enough information.

For each aspect to be categorised, a set of categories is needed. The categories need to be clear, not overlapping, and covering all possibilities. To make sure that all possibilities are included, it is common to include an “other” or an “unknown” category. Sometimes the choice of categories is determined by the research questions or is predetermined because they are recycled from a previously published paper. Often, however, some effort is needed to decide which categories to use.

When designing categories, a combination of logic, pragmatism and imagination is needed. The categories must be chosen so that they are relevant to answer the research question by separating the texts into meaningful groups. This is normally done before starting the content analysis. Each category should have a name and a short description to help the human coders to decide how to apply it. The descriptions are particularly important because categories that are self-evident to the person creating them can be obscure or interpreted very differently by others unless the descriptions are clear.

It is a good idea to conduct a small-scale pilot study to classify a small set of texts, say 20-40, to ensure that the categories are clear, exhaustive and appropriate for the texts. This saves time if the categories or descriptions need to be changed because it will be easier to make changes at this stage. In some cases the best categories to use may not be known in advance and can only be decided by examining the texts. To cope with this issue, a grounded theory-like approach can be used in the sense of constructing categories to fit the differences

found in the texts. This can be done by starting with a set of categories and then adding new categories for texts that do not fit the existing set. Alternatively, the classification process could start with no categories and produce the first categories by examining some texts and deciding upon how to group them. In either case, when adding a new category, all previously coded texts should be reclassified to check whether they fit the new category. If using multiple coders then only the first and main coder, probably the researcher, should complete the classification before the others start, otherwise different classification schemes will be produced.

When adding new categories, two key factors should be remembered. First, the key distinguishing features of each category should be relevant to the research questions or objectives. If this is forgotten then the results may be unhelpful. Second, it is not good practice to have categories with too few or too many texts. Consider merging categories that are too small, if it makes sense to do so in terms of answering the research questions. Consider also splitting categories that are too large – for instance containing most of the texts – if there is a natural way of doing so that fits with the research questions or objectives.

Example. The table below shows the results of a content analysis on a random sample of dialogs between pairs of MySpace members. MySpace was a very popular early youth-oriented social network site. In this example the factor contains a brief description and a lot of examples to illustrate the category. This level of description is rather light, however, and content analysis category descriptions should normally be longer than this. The purpose of the content analysis was to discover what types of things were discussed between Friends in MySpace, with a theory driven choice of gossip, information and offline coordination as the three categories. In this case, each dialog could be classified as belonging to any, all or none of the categories. Usually a content analysis splits texts into separate, disjoint categories, however.

Table 6.2. Frequency of three uses of MySpace social network site dialogs from a random sample of 100 (Thelwall & Wilkinson, 2010).

Factor	Percentage of dialogs	Examples (modified for anonymity)
Dialog contains any gossip - discussion of own or others' behaviour, attributes or activities	53%	I moved to Houston, Tx. I come home at the beginning of July well i just diyed my hair nearly black!! i regret not going to UMSX bc MZU is so much harder i sooo messed up :((i went out with car on friday night for a white guy tim knows a lot of rap song Tina talks about you all the time. she looked so much happier with you Nigel said you were feeling bad How far away does Tony go to school? J keeps askin if your comin to the wedding I heard u might be gettin a bike thats cool. so i heard that you started at the mall and just wondering how it is working out for you
Dialog contains any other information or facts	1%	Oprah lives in a mansion because she's rich.
Dialog contains any arrangement or coordination of offline activities or other communication	18%	CALL ME WHEN YOU GET A CHANCE hey text me sometime.. [number] i hope to see you toniite <3 I'm gonna be in ABD in Jan. for like a week, we gotta hang out Hey I can call you 2day?!!

The above table could be criticised for not having detailed enough descriptions of the categories. Although several examples are given for most categories, this is not as good as a more detailed category description.

The above categories may be surprising to some people if they have different understandings of gossip. There are similar but distinct dictionary definitions of meanings of this term and this can be confusing. For instance, gossip can be used to describe something negative – as in gossiping behind someone’s back. It can also be a more neutral term (as in the table above) to describe any kind of exchange of information about other people or the gossipers, whether positive or negative. This simple example shows the importance of descriptions and examples for the categories because without them a reader could completely misinterpret the findings.

Conducting a content analysis

Ideally, a content analysis should be conducted by three independent human coders classifying the same texts using the same classification scheme and descriptions. This is desirable because the use of multiple coders guards against individual biases. It also allows inter-coder consistency to be assessed by comparing the results of the different coders. If the use of multiple coders is impossible then the write-up of the results should include a discussion of the possible influence of coder bias on the results.

Each coder should be given a brief set of written instructions telling them what to do and how to do it. These instructions should ensure that they fully understand the task and the need to carry it out accurately. In addition, the coders need a method to record their classifications. Any appropriate method is acceptable, including the following suggestions:

- A printed sheet with the texts and a grid in which to enter the category selections.
- A spreadsheet containing the texts and columns to record the categories in.
- A word processor document containing the texts in a table with columns for the categories.
- Specialist software with support for content analysis coding, such as Nvivo.

A spreadsheet seems to be the best option for those without specialist software since it is possible to tally the categories using inbuilt spreadsheet functions.

Reporting the results

To report the results of a content analysis, the main task is to count up the number of texts in each category and then use a table, bar or column chart to display these totals. The numbers should also be used to calculate the percentage of texts in each category, which is the key outcome. If a sample of texts was classified rather than the complete collection then the percentages are estimates rather than exact values for the “population” of texts that the sample was taken from. This is because the sample might be a little biased even if it was properly selected at random. This is unavoidable and so the most important thing to do is to state that the percentages are estimates in the report discussing the data.

If multiple coders were used for the same texts then there is a special statistic that should be calculated to assess the extent of agreement between the coders, Cohen’s Kappa (Cohen, 1960). This statistic is a bit complicated to calculate but there are online calculators that can help. Essentially the statistic gives a number that reflects the degree to which the coders agree with each other, with higher numbers being better. There is no single agreed way to interpret the Kappa statistic but it is normally compared against a standard range of values to translate it into an intuitive interpretation. One such standard range is: at least 0.75 indicates excellent inter-coder agreement; 0.40 - 0.75 is fair to good agreement; and below 0.40 is poor agreement (Fleiss, 1981). As long as the figure is at least 0.40 then it seems reasonable to use the results. If not, a new coder could be enlisted as a replacement for an unreliable coder. If all the coders disagree with each other then this suggests that the

categories or their descriptions are too ambiguous and this aspect needs to be redesigned and the coding should then be redone.

If more than one coder is used then there are different ways in which their results can be combined. The easiest is to merge them all together. For instance, if three coders classify the same 100 texts then combining them would give 300 classifications, three per text. The duplication does not matter because this will disappear when the percentages are calculated. Alternatively, an attempt could be made to identify the best category for each text on a case-by-case basis. If all or a majority of the coders agreed on a category then the majority decision should be accepted. In cases where there is a tie or complete disagreement then the text could either be removed altogether or an additional coder could be enlisted to make a final decision based upon the classification scheme descriptions and code book.

Summary and recommended reading

- Content analysis is a systematic method for categorising texts into separate classes.
- It can be used to answer specific theory-driven hypotheses (e.g. are most contributions to a political forum argumentative?) or to give a descriptive characterisation of a new phenomenon.

For more information, see:

- Neuendorf, K. (2012/2013). *The content analysis guidebook (2nd Edition)*. London: Sage.
- Krippendorff, K. (2004). *Content analysis: An introduction to its methodology*. Thousand Oaks, CA: Sage.

Grounded theory

Grounded theory is a widely used qualitative research methodology that attempts to develop theory from data. The results are therefore “grounded” in the data rather than driven by existing theories. It is a good approach to use with new types of data or in contexts where there is not a strong pre-existing theory to channel a research project. There are different grounded theory variants and it can be applied to images, films or other objects as well as documents. This section, however, focuses on the most popular Anselm variant and its application to sets of texts.

At the heart of grounded theory is coding. It is similar to content analysis in this respect, but the way in which the coding is done, and its purpose, are both different. The coding does not begin with predefined categories but by examination of a sample of the texts to look for aspects or concepts that are relevant to the research questions or objectives. These concepts are then labelled in the texts. Once some texts have been labelled in this way, the labels themselves are examined to see whether they naturally group together in any relevant way. If they do, these clusters form categories in the data. The creation of these categories may then suggest the need to examine and code more data to expand the categories or to clarify their boundaries. More data examination may also be needed to find appropriate categories for the concepts that have been left out, although it is not necessary to fit all concepts into categories. Hence, the process described in this paragraph can be repeated until the categories are saturated with data – in the sense that additional text analysis does not produce new information.

The next stage is to examine the set of categories to look for relationships between them. This should lead to the formulation of hypotheses about the texts based upon the relationships between the categories. This leads to the need for more text analysis, a coding and possibly category formulation or reformulation in order to test the hypotheses i.e., revisiting the above paragraph).

Once hypotheses have been tested and found to be consistent with the data they assume the status of (substantive) *theories*. These theories are the primary outputs of a grounded theory study.

Although the above very brief summary describes the core of a grounded theory study, there are several other factors that need to be considered.

- Constant comparison is a critical part of the process. This primarily means comparing the concepts against each other to ensure that they are genuinely different and that the differences make sense. The same applies to the categories.
- Iteration is central. All the stages need to be repeated until a satisfactory theory is developed.
- The texts analysed do not have to be a random sample but can be purposively selected to help clarify categories or to test a hypothesis.

As an example of purposive non-random sampling, suppose that after several iterations of the grounded theory steps, a hypothesis had been developed that different age groups discussed a particular issue with differing strategies. Purposive sampling could be used to select texts from age groups that had not been sufficiently analysed so far, perhaps because they produced a low proportion of the texts. In this way all the age ranges could be saturated with texts analysed without having to analyse many more texts from age ranges that were already saturated.

Software for grounded theory Grounded theory can be conducted with the texts in any format that suits the researcher. Perhaps the simplest way is to print out the texts to be analysed and then code concepts using sticky notes or coloured highlighter pens. It is also possible to annotate the texts in a word processor or spreadsheet. The most sophisticated option is to use a specialist content analysis program, such as Nvivo or ATLAS.ti. This requires a time investment to learn but helps the research process by making it easier to do most of the tasks. For instance, such software will offer ways to help cluster concepts into categories and to view all texts that have been coded with specified concepts or categories to support the constant comparison process.

When loading texts into a software package a decision must be made about what is the basic unit of text: the document. In traditional offline analyses each individual document may be quite long, such as a complete transcript of an interview or focus group session. In contrast, internet texts are often quite short, such as individual blog posts or tweets. Unlike a content analysis, with grounded theory multiple codes (concepts) can be assigned to a single document and so it is not necessary to break up the texts analysed into small segments. On the contrary, it can be helpful to have all associated texts within a single document so that when the document is coded the context of any particular section of text is clear. For instance, suppose that a data set is a collection of discussions in a particular online forum. It would be helpful to have all contributions to a discussion stored in the same document rather than creating a new document for each contribution. This is because the meaning of any contribution would be clearer if the contributions before it could also be read.

In comparison to content analysis, grounded theory has the advantage that it produces theories from the texts and the outcomes can be more interesting as a result. A key disadvantage is that the method is non-scientific in the sense that the results are not normally based upon random samples of texts and there are no real safeguards against researcher bias. In contrast, content analysis can use multiple coders to guard against bias. With grounded theory the onus is on the researcher to make the case that the findings are a valid interpretation of the texts.

Summary and recommended reading

- Grounded theory is based upon identifying concepts in texts, grouping the concepts into categories, and developing and testing hypotheses from the categories.
- The outcome is one or more substantive theories, which are tested hypotheses.
- Grounded theory is an iterative process with frequent revisiting of the data. It uses purposive sampling rather than random samples.

For more information, see:

- Charmaz, K. C. (2006). *Constructing grounded theory: A practical guide through qualitative analysis*. London: Sage.

- Lewins, A., & Silver, C. (2007). Using software in qualitative research: A step-by-step guide. London: Sage. [This gives information about Computer Assisted/Aided Qualitative Data Analysis (CAQDAS) software for grounded theory and other purposes.]

Sentiment analysis

Sentiment analysis is a method for automatically identifying sentiment in text using a specially-designed computer program. Sentiment analysis programs identify indicators of sentiment in text and use them to predict what the overall sentiment of the text is likely to be. For example, a text containing many positive words and no negative words is likely to be classified as being positive. Sentiment analysis developed for commercial applications – such as identifying unhappy customers from their online comments. Now it is also used in social science research, primarily to address research questions concerning the role of sentiment in various types of, or contexts for, online communication.

There are several different kinds of sentiment analysis but only two are discussed here. *Trinary sentiment analysis* classifies texts into three different types: those that are positive overall, those that are negative overall and those that are neutral overall. Applying a trinary sentiment analysis is a bit like conducting a human content analysis for the three categories of positive, neutral and negative sentiment, except that a computer does the work of classification. Although the time saved by the automation is offset by the time needed to learn how to use a sentiment analysis program, the program can process an almost unlimited number of texts quickly. Hence sentiment analysis can be carried out on a much larger scale than content analysis. As a result it can identify patterns that would not be evident in smaller scale analyses. Sentiment analysis categories can be cross-referenced with other categories to detect sentiment-related differences in these other categories. For instance do female users of a given system use positive sentiment more often than do male users? Are political blog posts more often negative than news-related blog posts?

A second type of sentiment analysis is *sentiment strength detection*. Instead of assessing whether a text is positive, neutral or negative, it estimates the strength of positive and negative sentiment in the text. Hence, each text gets a positive sentiment strength score of 1 (no positive sentiment) to 5 (very strong positive sentiment) and a negative sentiment strength score from 1 (no negative sentiment) to 5 (very strong negative sentiment). A neutral text would have a score of 1 for both (i.e., 1,1) or equal positive and negative sentiment strengths. It is possible for even a short text to contain both positive and negative sentiment, in which case both sentiment scores would be greater than 1. Sentiment strength scores for texts can be calculated automatically with software such as SentiStrength, as described on the book web site. As with trinary sentiment analysis, the ability to automatically predict sentiment strengths is an advantage because large numbers of texts can be quickly processed. In contrast, however, the sentiment strength results are scores (1-5) rather than categories. Table 6.2 gives examples of sentiment strength scores for some sample texts.

Table 6.2. Some examples of texts and computer-annotated sentiment strength scores. Words triggering the sentiment scores are in italic.

Text	Positive (1-5)	Negative (1-5)
Your blog is <i>really wonderful!</i> Thank you for sharing.	4	1
The prime minister is a <i>disgrace</i> and should <i>resign</i> .	1	4
Can we meet up next week? <i>x</i>	2	1
I am <i>tired</i> but <i>happy</i> after finishing the half marathon.	3	2

The natural way to investigate sentiment strength is to compare it with another category associated with the text, such as author gender or text topic. The *average* positive and negative sentiment strength for each category can then be compared. This can answer questions such as: do male or female authors use stronger average positive sentiment?; Which topic is associated with the strongest average negative sentiment? Sentiment analyses works best when combined with categories that can be automatically or easily identified for large

numbers of texts. The results can be summarised in a table or graph to reveal differences. A proper comparison needs a statistical method to test for genuine differences, however, such as a Wilcoxon signed rank test for differences in sentiment strength between two categories.

Example Sentiment may relate to age and gender in intelligent design debates. To investigate this, the comments on a set of 50 randomly selected ID videos in YouTube were downloaded, categorised by author gender, and classified for sentiment automatically (with SentiStrength, free online). Only one comment was classified per commenter, even if they contributed many. Figures 6.3 and 6.4 show the percentage of comments in each sentiment category, broken down by gender. The graphs show that males tend to use stronger negative sentiment and females tend to use stronger positive sentiment.

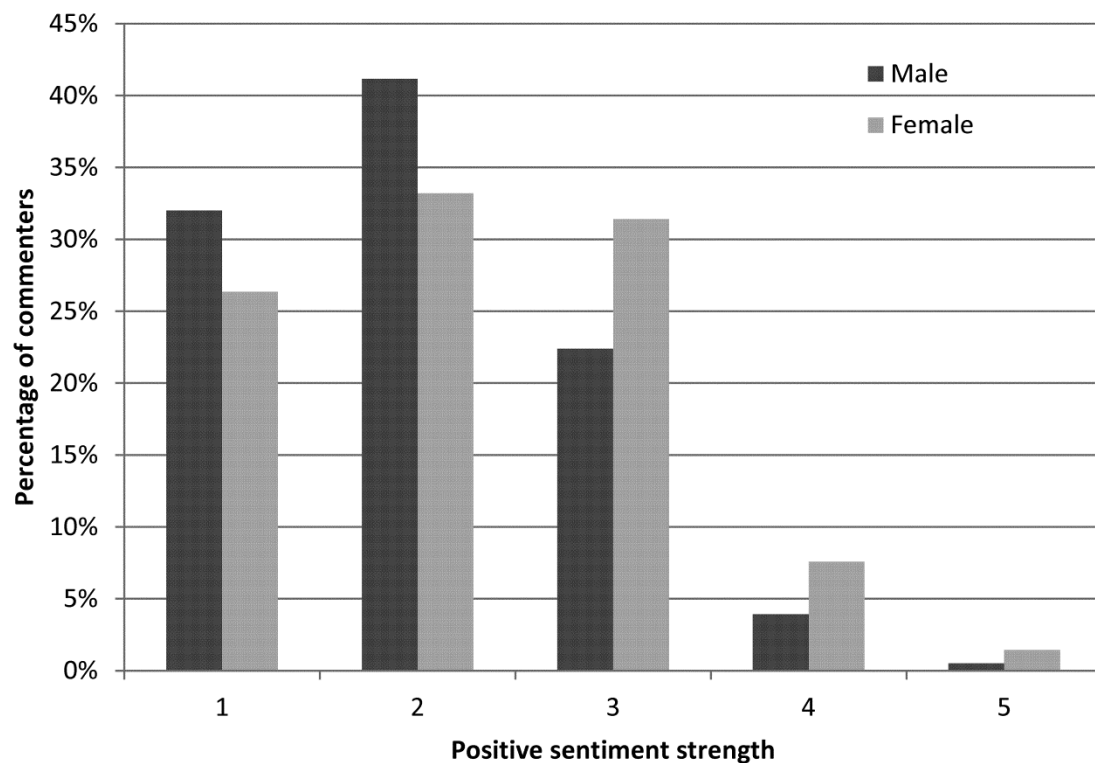


Figure 6.3. Sentiment strength of commenters on 50 random Intelligent design YouTube videos. Positive sentiment strength is classified from 1= no positive sentiment to 5 = very strong positive sentiment. In the graph, male commenters are more likely to express no positive sentiment (i.e., strength 1) or weak positive sentiment (strength 2) whereas females are more likely to express moderate or strong positive sentiment (strengths 3 to 5).

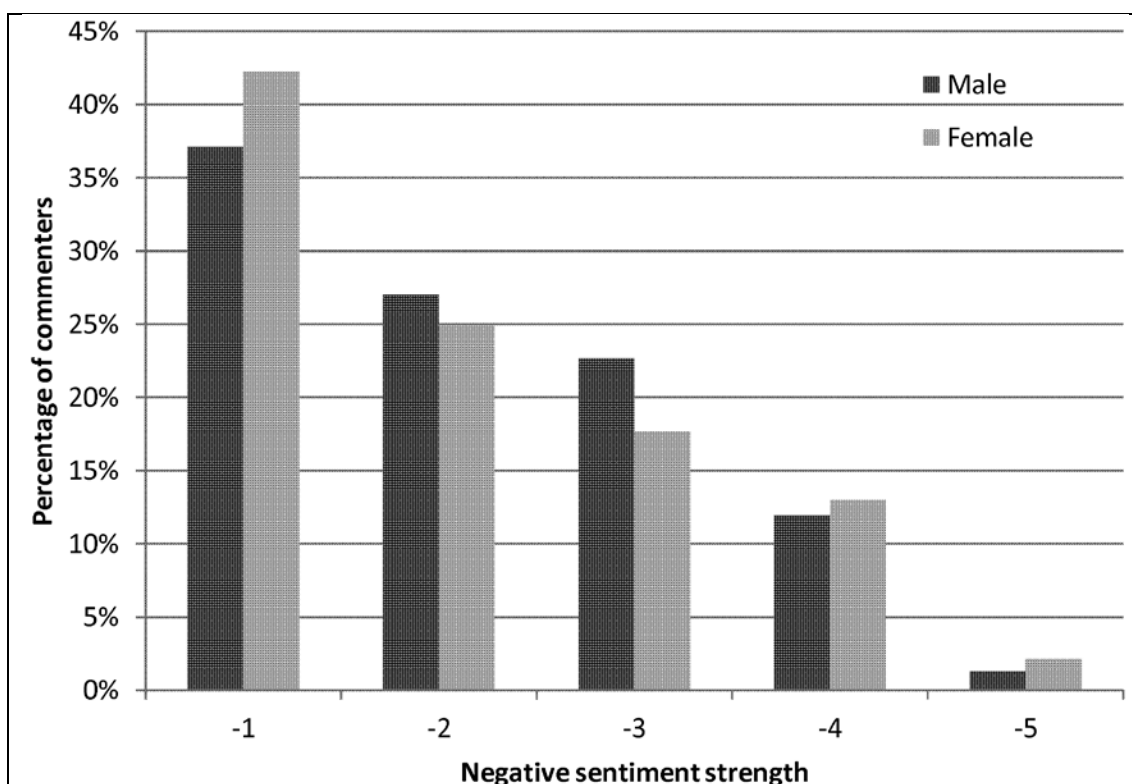


Figure 6.4. Sentiment strength of commenters on 50 random Intelligent design YouTube videos. Negative sentiment strength is classified from -1= no negative sentiment to -5 = very strong negative sentiment. In the graph, female commenters are more likely to express no negative sentiment (i.e., strength 1) or strong negative sentiment (strengths 4 and 5) whereas males are more likely to express weak (strength 2) or moderate (strength 3) negative sentiment.

If there are only a few texts to analyse then sentiment analysis could be conducted manually instead of with software. Here, human coders would classify the texts with the sentiment categories or the sentiment strength scores. If this approach is used then the same method is needed as for a standard content analysis – using multiple coders and clearly defined categories.

Summary

- Sentiment analysis research is normally used to investigate the role of sentiment in various online contexts. This may be investigated by testing for differences in sentiment use across a range of text categories (e.g., topic or author gender).
- Sentiment analysis programs can automatically predict whether a text is overall positive, negative or neutral or estimate the strength of positive and negative sentiment in it.
- Sentiment analysis can be automatically applied to large collections of text.
- Sentiment analysis results are most useful when cross-referenced with categories to detect sentiment differences between categories.

Final thoughts

This chapter has covered sampling strategies for comments in the social web as well as four different methods: graphs and simple statistics, content analysis, grounded theory, and sentiment analysis. If these methods are inadequate for a particular goal then there are many other text analysis methods to choose from, including the following.

- *Discourse analysis* is a set of methods to identify the structure of communication in a given context. It is appropriate when complete sets of comments are available and there are some reasonable expectations about the likely key factors involved. There are many different types of discourse analysis, with variations between disciplines in the style of discourse analysis used.
- *Genre analysis* is a method *to* study the types and structures of document and can be applied to social web texts in heterogeneous environments.

7. (T2) Blog and Twitter Searching [updated throughout]

Blogs are websites containing time-stamped postings written by one or more people and displayed in reverse chronological order. Blogs collectively form a fascinating new data source for social sciences research because there are so many of them – perhaps hundreds of millions – and they tend to be the informal thoughts or diaries of relatively ordinary people. For example, in principle it would be possible to get a snapshot of public opinion related to any major topic by randomly sampling recent blog posts. Of course, bloggers are not typical citizens – they have Internet access and are probably relatively computer literate – but blog sampling can be achieved so quickly and easily that it can be a useful additional research technique for many investigations.

Blogs have a feature that makes them a unique data source for some social issues. The relative permanence of blog posts, due to old posts being archived online rather than deleted make them a source of *retrospective public opinion*. It is possible to discover what bloggers thought about any topic at any point in the recent past simply by selecting posts from that time period. There is no other large scale accessible source of public opinions about major topics in the past, other than about recognised issues via regularly administered standard surveys (e.g., national surveys of social attitudes). Without blogs, therefore, it would be impossible to answer questions such as: what did the public think about Barack Obama before he stood as a presidential candidate?; when did the slang word “minger” become commonly used in the UK?; or when did the Danish cartoons affair become a major issue for the public of Western Europe? This chapter describes techniques for addressing such issues.

Microblog posts, such as from Twitter or Sina Weibo, have some attributes in common with blog posts but some important differences from a search perspective. An important difference is that microblogs are typically much shorter than blog posts and hence describe their topic less clearly and in much less detail. Moreover, they are intended to be ephemeral rather than permanent and so they can disappear within a short period of time, either from the creator's profile, or from search engines. As a side effect of their short length, microblog posts do not have their own page and URL but are only listed in user profiles.

Micro/blog search engines [modified]

The simplest blog investigation technique is blog-specific search engine keyword searching. In early 2013 there was a choice of blog search engines, including Technorati, IceRocket and Google Blog Search (for a review see: Thelwall & Hasler, 2007). Each one can return results in reverse chronological order so that the most recent post is returned first. Figure 7.1 shows results for the search “Barack Obama” in Google Blog Search, with the first result being only 1 minute old. A less popular topic might find the first result being days or months old instead, however. A simple qualitative way to get an insight into public reactions to a topic, then, is to construct a suitable keyword search for the topic, and spend some time reading through the matching blog posts. This could be formalised by conducting a content analysis of some or all of the matching posts (see elsewhere for more on content analysis).

Google Blog Search BETA

"Barack Obama" Search Blogs Search the Web [Advanced Preferences](#)

Blog results Results 1 - 10 of about 5,517,695 for "Barack Obama". (0.22 seconds)

[Browse Top Stories](#) New! [Sort by relevance](#) [Sorted by date](#)

Published: [Last hour](#), [Last 12 hours](#), [Last day](#), [Past week](#), [Past month](#), [Anytime](#), [Choose Dates](#)

Subscribe: [Blogs Alerts](#), [Atom](#) | [RSS](#)

Related Blogs: [Welcome to Obama for America](http://www.barackobama.com/) - <http://www.barackobama.com/>
[Barack Obama - U.S. Senator for Illinois](http://obama.senate.gov/) - <http://obama.senate.gov/>
[Against Barack Obama](http://www.againstobama.com/) - <http://www.againstobama.com/>
[Barack Obama : Pictures, Videos, Breaking News](http://www.huffingtonpost.com/news/barack-obama) - <http://www.huffingtonpost.com/news/barack-obama>
[Flickr: Barack Obama's Photostream](http://www.flickr.com/photos/barackobamadotcom/) - <http://www.flickr.com/photos/barackobamadotcom/>

[Catholic Church Conservation: Catholic bishop thanks God for Obama](#)
 1 minute ago by Gillibrand
 The Catholic Bishop of Portsmouth, Crispian Hollis, has used his diocesan website to thank God for the election of the pro-abortion **Barack Obama** – a clear abuse of his position that has outraged one of Britain's leading pro-life ...
[Catholic Church Conservation](http://cathcon.blogspot.com/) - <http://cathcon.blogspot.com/>

Figure 7.1 Google blog search results for “Barack Obama” in November, 2008.

The simple blog search has an important commercial application in market research, especially in support of marketing initiatives. A company that has just launched a major advertising campaign may want quick public feedback on its impact. One way it can do this is to search for relevant product or brand names in blogs and to read entries made after the start of the campaign. Computing organisations like IBM and Microsoft monitor large numbers of blogs so that they can automatically deliver relevant results to customers or allow customers to process the raw data themselves. In addition, specialist web intelligence companies like Nielsen offer a similar service. Companies without a large market research budget could use blog search engines as a substitute, however.

There are also some microblog search engines, including the main search.twitter.com Twitter search facility and independent search services, including Topsy (<http://topsy.com/tweets>). Unlike blog search engines, these seem to search only recent tweets from the previous 2-4 weeks.

Date-specific searches

The uniquely valuable information contained in the world's blogs is the time-stamped comments of a large number of ordinary bloggers. Several blog search engines, including those mentioned above, allow date-specific searches: normal searches but with a single day or range of dates selected for the results. Hence it is possible, for example, to search for opinion about an issue one month or one year before it became major news. Of course, if nobody blogged about the issue before it became news then the blog search results would contain no matches.

As with standard blog searching, insights could be gained by spending time reading the blog posts in the data-specific search results or a more formal content analysis could be used. Unlike standard blog searching, however, there is normally no way of directly corroborating the results because there is no other large repository of public thoughts about any issue and humans are poor at recalling their past thoughts on an important topic. Hence, it may sometimes be that date-specific searches must be used unsupported as evidence for a hypothesis in research. In such cases the researcher should discuss and evaluate the likelihood of bias in the blog data source. For example, they should assess whether an important constituency would be missing because it would have been unlikely to blog or would have blogged in a different language or in inaccessible blogs. The researcher should mention in their report the weaknesses in blog searches as a data source, as discussed in this chapter.

In addition to date-specific searches some blog search engines provide a range of advanced search facilities, typically on a separate page reached by clicking an “advanced search” link on the home page. These may include factors like blog importance, language and

topic. Searches for blogs that link to a given blog are also possible in some search engines. All of these are probably less useful than the simple date-specific search for the typical researcher, however.

There do not seem to be any search engines that currently allow date-specific searches of microblogs but these searches can be run indirectly for Twitter by requesting a graph of tweets matching a query (see below) and then clicking on the graph to identify tweets from any specific date within range (currently up to a month in Topsy Analytics).

Trend detection [modified]

A second unique feature of blogs and tweets, also derived from their time-stamped postings, is their ability to reveal trends in public opinion over time. This is supported in some of the blog and microblog search engines through the production of graphs of the daily volume of blog posts matching a keyword search. Currently only IceRocket (trend.icerocket.com) offers this free for Blogs (for the previous 3 months) and Topsy Analytics (analytics.topsy.com) for tweets (1 month). Note that this is the first genuine webometric technique in this chapter in the sense of being quantitative.

Figure 7.2 illustrates the results of a search for *Cartoons AND (Denmark or Danish)* graphed by BlogPulse in May 2006 (Thelwall, 2007). The graph shows that interest in the Danish Cartoons affair in English language blogspace began on about the 26th of January, 2006. From then it grew rapidly and then gradually subsided. The January start date is surprising because the cartoons were published on September 30 of the previous year and generated media attention in Europe and condemnation by various Muslim groups and politicians but attracted almost no English language blog posts before January (at least as reported by BlogPulse and corroborated by Google blog search). This seems to be strong evidence that the importance of the issue in the English-speaking world was not recognised at the time of publication and that other factors must have triggered the major debate. The evidence for this case appears particularly compelling because blogspace is a natural arena for airing and discussing important issues relating to politics and the news and so it seems inconceivable that there would be a public perception of the importance of the Danish Cartoons without a corresponding significant number of blog postings.

The BlogPulse Danish cartoons graph can also give evidence for the cause of the explosion in debate. Clicking on the start of the explosion on the 26th of January and just afterwards gives a date-specific search for blog posts on the date. Reading these posts reveals two main events being discussed: the withdrawal of the Saudi Arabian ambassador from Denmark and the boycott of Danish food in Saudi Arabia. It seems that the coincidence of these political and economic events triggered the public debate.

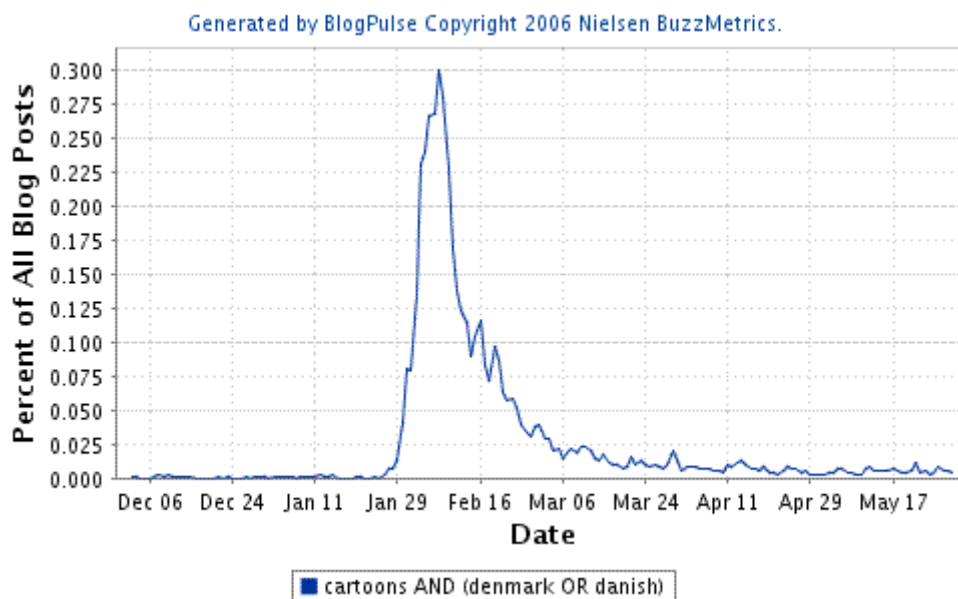


Figure 7.2. A blog trend graph of the cartoons debate: volume of blog postings related to the Danish cartoons debate (reproduced with permission) (Thelwall, 2007).

Blog and microblogs trend graphs, produced as above, are useful for several different related purposes.

- *Identifying the starting point for discussion* of an issue, verifying that an issue had not been significantly discussed before its assumed start date, or identifying micro/bloggers who predicted or discussed an event before it became popular. The Danish Cartoons graph illustrates this application, with blog posts at the trigger point revealing the cause of the explosion in discussion.
- *Identifying key events within a broad issue.* This can be achieved by constructing a keyword search for the broad issue (e.g., “stem cell research” or “abortion”), producing a micro/blog search graph and looking for spikes in the graph. Each spike is likely to represent a significant related news event that triggered blog discussions. For example, the Figure 4.3 spike on September 10, 2008 was triggered by discussion of stem cell research ethics as part of US presidential election campaigning. These news events can be identified by running date-specific micro/blog searches for the dates of the spikes (e.g., by clicking on the graphs). This technique works best for issues that occasionally generate significant blog discussions because the graphs produced by less popular issues are likely to be too spiky, dominated by random fluctuations in blogging rather than external events.
- *Identifying long term trends.* A micro/blog graph can reveal trends in the long term interest in a topic – such as whether there has been an overall increase or decrease in interest over time. For instance, Figure 7.3 suggests that public interest in stem cell research increased in the second half of 2008. This could also be used, for example, to help gauge whether a retired politician had been quickly forgotten or had retained a lasting influence.
- *Comparative time series analysis.* It is possible to construct multiple micro/blog time series with the explicit purpose of comparing how the trends relate. For example Figure 7.4 compares the proportion of blog postings that mention *Obama* with the percentage that also use the word *black*. Comparing the two time series in the future, it may be possible to detect presidential incidents where race is invoked unusually frequently by bloggers. No such incidents are clear from Figure 7.4, however, although precise measurements taken from the graph (e.g., the height of the top line divided by the height of the bottom line) might reveal some unusual ratios.

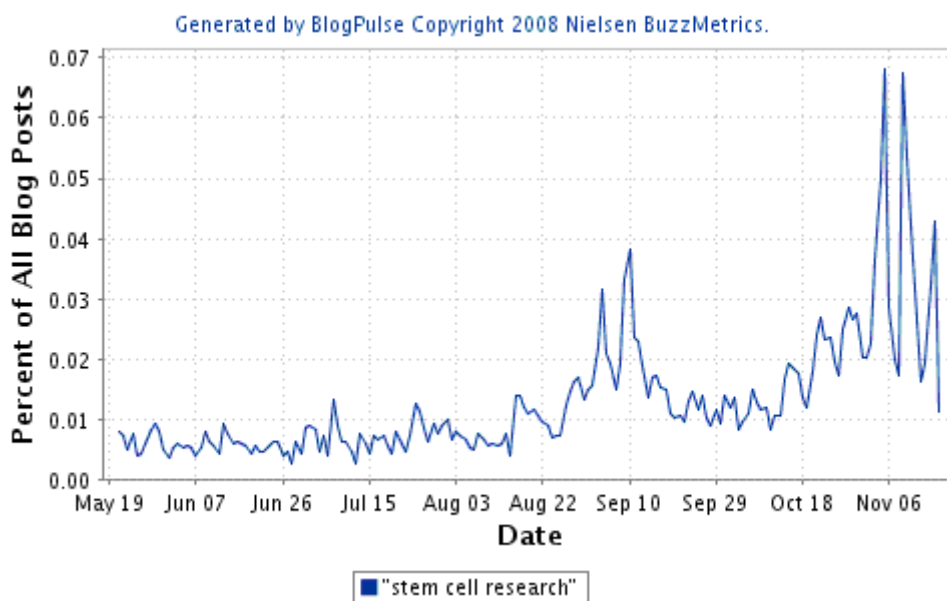


Figure 7.3. A blog trend graph of blog postings for stem cell research, indicating an apparent long term increase in interest and at least two significant events (reproduced with permission).

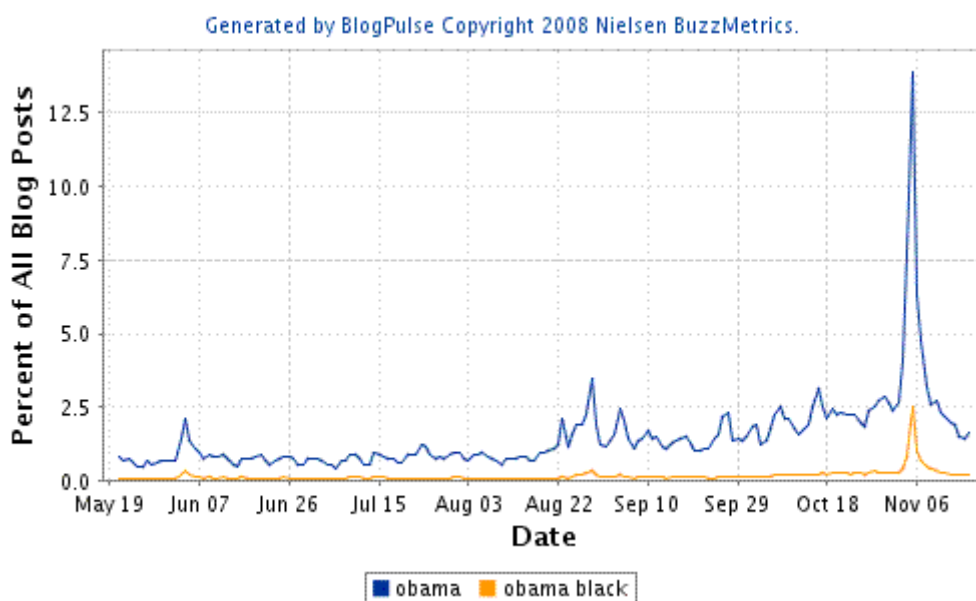


Figure 7.4. A comparison of the volume of blogs mentioning *Obama* with the volume mentioning *Obama* and *black* (reproduced with permission).

Checking trend detection results

When using micro/blog trend graphs, two further techniques are useful to check the results: keyword search testing and search trend comparison. Keyword searches should always be tested by reading a random sample of matching blog posts to ensure that most posts discuss the desired issue. Without testing it is possible that the trends graphed may be unrelated to the desired issue. For example, if the Danish cartoons graph had been produced with the single keyword "cartoons" then it would have contained a second bump caused by an international cartoon convention.

Search trend comparison means using general search volume graphs, like those produced by Google Trends (www.google.com/trends/) to produce parallel graphs to the blog trend graphs so that the shapes and key dates can be compared. This corroborates the blog information with general search information to assess whether the phenomenon is specific to

micro/blogs. The Google trends tool shows the daily volume of Google searches matching the keywords. In the Danish Cartoons case the shape of the blog graph Figure 7.2 is replicated by the Google Trends search volume graphs for the keyword search *danish cartoons* (Figure 7.5). Of course, the two graph shapes will not always match as there may be topics that people search for but rarely blog about. Google Trends graphs are ostensibly a more authoritative source of evidence about trends in public opinion than blog graphs because more people search with Google than keep a blog. Nevertheless, Google Trends graphs cannot be validated by finding the cause of the searches graphed in the same way that blog trend graphs can be validated by reading the underlying blog posts from the graph. Hence, Google Trends graphs are best used to corroborate the blog trends graphs rather than as primary evidence. Google Trends can also give some insights about why Google users have searched for particular terms by showing related terms.

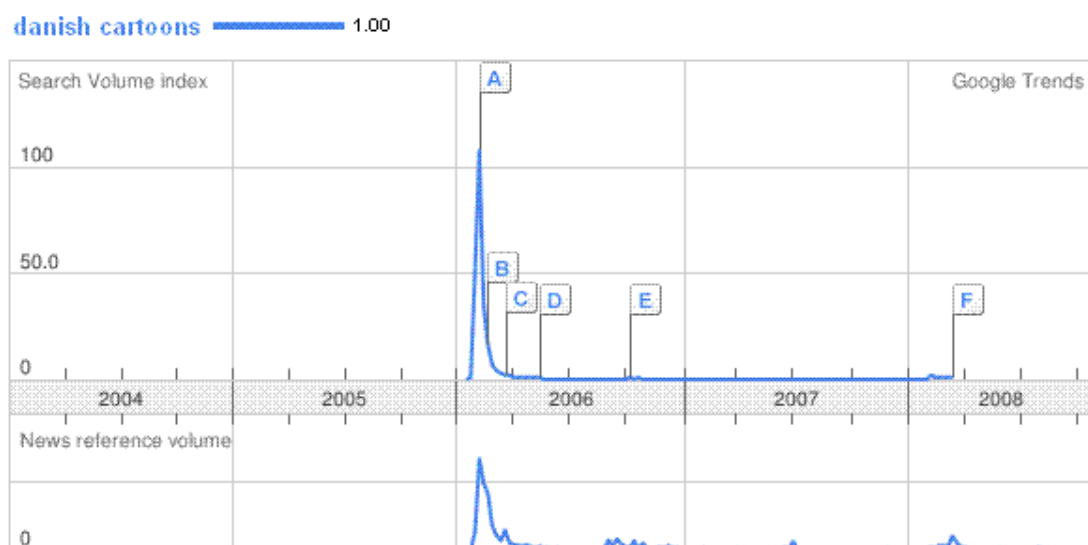


Figure 7.5. A Google Trends graph of the volume of *danish cartoons* searches (reproduced with permission)

Limitations of micro/blog data

A disadvantage of micro/blogs as a data source for public opinion was introduced above: that micro/bloggers are atypical citizens to some extent. This should not be exaggerated for developed nations since a majority have Internet access in one form or another and a small but significant proportion of the population have kept a blog (Lenhart, Arafeh, Smith, & Macgill, 2008). This proportion would be larger if the blogs of large social network sites like Facebook were included, although they are not (currently) open to blog search engines.

A second disadvantage of micro/blogs in practice is that the coverage of micro/blog search engines tends not to be reported and so there is an unknown bias in micro/blog search engine results. Because blog search engines need to be engineered to process each different blog format efficiently, it is likely that only the most common formats are supported. Fortunately this probably includes most or all large free blogging sites like blogger.com, but it may exclude all individually-designed blogs. Moreover, blog search engines may include blogs with a Really Simple Syndication (RSS) feed or similar because these are easy to process. In addition to these considerations, a blog can only be included if a blog search engine has found it. It may find blogs by a range of methods, including user submission of URLs and following links. These methods are likely to favour more popular blogs because their owners are more likely to promote them and others are more likely to link to them. Non-English blog formats are also less likely to be included because of the difficulty for US-based blog search engine programmers in understanding them.

An important limitation of micro/blog searching, for international issues, is that it is normally language-specific, with the exception of proper nouns and accepted international expressions. Moreover, there is variation in the extent to which micro/blogging is popular in particular countries – for example it seems particularly common in the US and Iran but is probably rare in most of Africa due to poor Internet access. Micro/blog search engines are almost always international in coverage – partly because it is difficult to automatically identify the nationality of a micro/blogger since it is often not reported – and so the results of any search may include results from an unknown proportion of an unknown range of countries, which complicates their interpretation.

As a result of the disadvantages of micro/blogs in terms of their ability to provide a representative sample of public opinion they are not recommended as a *primary* data source for public opinion research if alternatives are available. Instead, they are useful as part of an initial exploratory research phase, to gain background information about an issue or to help the initial formation of research hypotheses to be subsequently tested against more robust data. Ignoring micro/blogs because of the poor quality data could be a mistake if a research project subsequently suffers from inadequate hypotheses or poor background knowledge. Another use of blog searching is as part of a longitudinal study, for example taking weekly samples of micro/blogs, in which the task of identifying changes would cancel out some of the limitations of the data source.

Advanced blog analysis techniques

All the blog analysis techniques discussed so far use freely available blog search engines and are easy to set up and use. For some purposes, however, more sophisticated tools operated by the researcher may be needed to produce the necessary results. This may involve a considerable increase in effort and difficulty. Blog search engines collect data daily on millions of blogs and this requires considerable computing resources in terms of software development, computing power and bandwidth. As a result, this undertaking seems justified for a researcher only if it would give the main evidence for a research investigation. This section gives some ideas about employing blog monitoring and analysis programs.

- *Automatic spike detection in broad issue scanning.* The visual spike detection technique discussed above for identifying key events within a broad issue can be automated. If a large collection of blogs is monitored daily then a blog search program can be written to match all posts relating to a broad issue, for example via a general keyword search for the issue. It is then possible to detect overall spikes via testing for increases in the volume of matching blog posts. This is an automated version of manual spike identification. More significantly, it is possible to monitor each word mentioned within each blog posting to identify if and when each individual *word* spikes (Thelwall & Prabowo, 2007). This technique can detect hidden spikes for a sub-issue of a large broad issue. Part of the power of this technique is that through its exhaustive search of every word it is able to detect previously unknown sub-issues within the known broad topic. An example of this technique is a project to automatically detect public fears about science from blog postings. This project monitored tens of thousands of blogs, filtering for the broad issue through a keyword search comprising a list of synonyms of “fear” together with a list of synonyms of “science” and “scientist”. This extracted blog posts expressing concern and mentioning science. Automatic scanning of these filtered posts revealed a range of new issues, including concerns about CISCO routers in the Internet, stem cell research events and the Terri Schiavo medical case in the US. The strength of this method is its ability to detect previously unknown events hidden in a mass of data but its disadvantage is that it is most effective for events associated with distinctive and unusual words.
- *Vocabulary analysis.* Some research requires exhaustive lists of words in blog postings, for example in linguistic analyses of word variations either for linguistic research or as part of a wider investigation into the spread of ideas. These vocabularies can be constructed by collecting blog postings, extracting lists of words from all postings and constructing a list (vocabulary) of all words found. This is a particularly useful technique

when the words or word variations sought are not known in advance but can only be found by scanning word lists (Thelwall & Price, 2006). The method has been used in a project to find and explore the evolution of words that are derivatives of Franken- as a framing device in debates over genetically modified food (e.g., Frankensoya, Frankenwheat). The results may also suggest when each new word first appeared in blogs, which could be corroborated through commercial blog search engines.

- *Long term trend graphs.* At the time of writing, none of the commercial blog search engines would produce a blog search graph for longer than the previous year. Hence projects needing long term graphs could either purchase the data from the blog search engine companies or collect their own data. The latter option is a major undertaking because at least 100,000 blogs should be monitored to give reasonable quality graphs for anything but the biggest issue. If too few blogs in the collection discuss an issue then the graph produced is likely to be dominated by random spikes.

Advanced microblog analysis techniques

Microblogs are currently easier to gather than blogs because they originate from fewer sources, and are smaller and less formatted, making the software needed much easier to create. For instance, the time series analysis methods discussed above can be carried out on a collection of tweets gathered by Webometric Analyst and then processed using Mozdeh. Instructions for this are available online (<http://mozdeh.wlv.ac.uk/installation.html>).

It is possible to gain access to additional microblog searching features and older data by purchasing a subscription to a service such as Topsy Analytics Professional. Figure 7.6 is a graph produced by the free version of Topsy Analytics. The professional version gives over a year of Twitter data and creates a number of different time series graphs for it, including sentiment graphs that illustrate changes in sentiment over time for tweets matching the query. Companies can use services like this to monitor their brands in real time but they are also available for use by researchers.

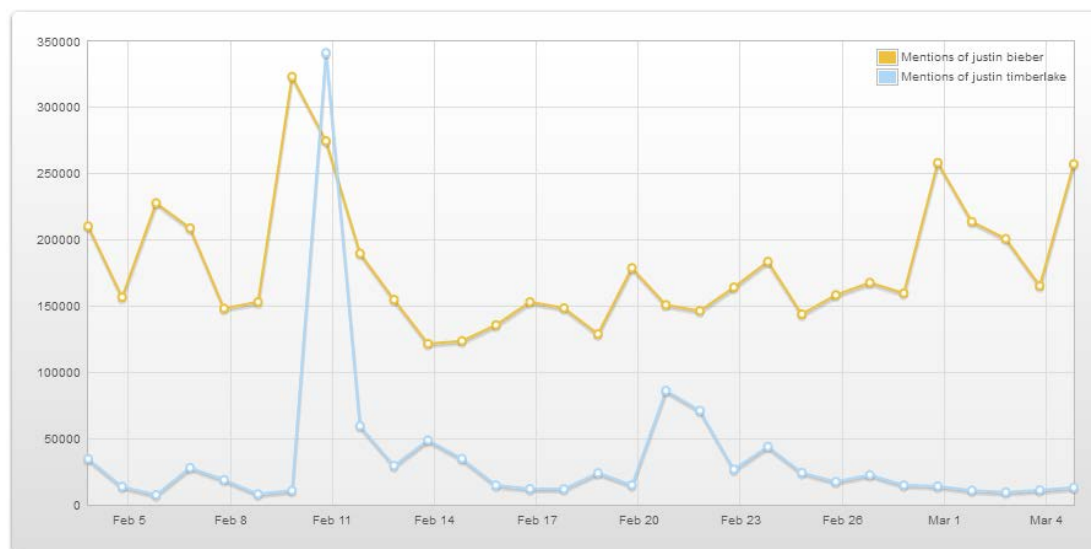



Figure 7.6. A Topsy Analytics graph of the volume of tweets matching the queries *justin bieber* and *justin timberlake* over a month.

Summary

This chapter discusses a range research techniques related to micro/blog searching. The first and simplest techniques are not examples of webometrics because they do not involve any counting but the remainder are. The techniques make micro/blogs a useful data source for public opinion and in some cases give information that is not available elsewhere. Moreover, most of the methods discussed are relatively easy to carry out via commercial micro/blog search engines. Nevertheless, micro/blogs provide weak evidence in many cases because

bloggers are not representative of the population and the topics they blog about are not likely to be always representative of the issues that the population considers important. Micro/blog analysis is therefore best used as a “quick and dirty” way of gaining insights into a public opinion topic, to investigate topics that cannot be addressed in any other way (e.g., retrospective public opinion), or in conjunction with other corroborating evidence such as from Google trends.



8. (T3) Web Texts Thick Description [new]

Introduction

This chapter describes a set of methods to analyse a topic or issue discussed within a set of web texts, such as from Twitter, Sina Weibo, blogs or a forum. The methods are primarily descriptive in the sense that their objective is to describe the content of the texts rather than to test a specific hypothesis. Specific hypotheses can be informally tested, however, by using theory to predict in advance what the results might be and then assessing the accuracy of the prediction.

To illustrate how this chapter can be used, suppose that a researcher is interested in online discussions of autism and wonders how people discuss autism online and whether this online discussion could give insights into the general perception of the issue. The researcher might use the information in this chapter to help collect and analyse tweets related to autism over a period of time, such as before, during and after Autism Awareness Week. The results may reveal the common subtopics discussed within the broad area of autism and how the issue changed during the week. This descriptive information could then lead to a conclusion about how useful Twitter was as a source of information about perceptions of autism, and suggestions for follow-up, more targeted research.

This chapter gives a general overview of the thick description of web texts method, describing it in a general way that is applicable to many types of texts. The examples within this chapter illustrate the application of some of the methods for individual topics and offline resources give instructions about how to gather and process sets of texts from specific sources with the methods described here.

Stage 1: Constructing and testing queries

Text data is often gathered by constructing one or more queries and harvesting all texts matching these queries using a search engine of some type. Alternatively, sometimes all texts from a specific source may be harvested (e.g., a particular forum or chatroom) or all texts gathered from a selected group of users in a specific site may be downloaded. This section deals with some key (and quite similar) issues with all of these.

The key underlying factor for all three methods is *precision*. High precision here means that a high percentage of texts gathered are relevant to the project goals. High precision is important because if a significant proportion of the texts are irrelevant then any results from analysing the texts may be spurious in the sense of being irrelevant to the topic investigated. It is more difficult to analyse a data set that is mostly irrelevant than to analyse a dataset that is small and so the data collection priority should be high precision rather than gathering a large amount of texts (i.e., precision is more important than recall in information retrieval terminology).

Keyword queries: To make high precision keyword queries, start with brainstorming for a list of queries, and then test each query by using it to search for matching texts and briefly assessing whether the results contain a significant proportion of irrelevant content. If irrelevant content is identified, then the keyword query should be modified by adding extra terms or exclusion terms (discussed later) to generate higher precision queries. This should be done iteratively to ensure high precision queries, even if this means that the queries contain several terms and match few texts. Keywords that cannot be successfully modified to give high precision must be ignored unless it is practical to engage in enough post-hoc irrelevant content filtering to remove enough of the irrelevant query matches. Adding extra terms works if search results must match all terms in the query. For example, a search for baby might return too many uses of this word as a term of endearment and these irrelevant matches could be reduced if the single query baby was replaced by a set of queries with extra terms chosen to match child-related texts, such as *baby child*, *baby mother*, *baby pregnant*, *baby feeding*, *baby milk*. As this example illustrates, however, adding extra terms is likely to introduce a new source of bias in favour of the keywords added and so the results of the subsequent

analysis should take this into account. Exclusion terms are terms that exclude any matching texts, and in most search systems are indicated by a minus sign before the term. For example, an initial query of *cheddar* for a cheese investigation might return some mentions of Cheddar Gorge in Somerset, UK. The query could be modified by excluding the term Gorge and modifying the query to *cheddar -gorge* so that less matches will be about the Somerset attraction. It is likely that not all irrelevant content can be removed so the goal should be to ensure that the highest possible percentage of relevant content is present in the final version of the filtered text collection.

Specific source: If all texts from a specific source are to be harvested then they are likely to be at least partly irrelevant, even if only because some discussions travel off-topic. There is little that can be done about this at the data gathering stage except by ensuring that the sources gathered from are as narrowly focused on the research topic as possible. Extensive post-hoc filtering (see below) will probably be needed for this type of data gathering.

Specific users: If all texts from a set of selected users are to be harvested then, if possible and fitting the research goals, the users with the narrowest relevant topic-focus should be chosen. This may not be possible to tell in advance, however, and may not be desirable if the research goal is to investigate typical users rather than narrowly-focused users. Extensive post-hoc filtering (see below) will probably be needed to remove irrelevant data from the users. This may take the form of query-based relevant content inclusion (see below) rather than the other methods because it will be easier to match relevant content than irrelevant content with a query.

Stage 2: Post-hoc spam, irrelevant content, and duplicate filtering

Even with carefully constructed queries some spam and irrelevant content is likely to filter through and end up in the data collected. This can be a problem for any kind of automatic analysis because spam is irrelevant to the topic analysed and is likely to be repetitive and hence will fool any automatic algorithm looking for time series or keyword patterns. Hence, once the data collection is complete, the next stage should be data cleaning to remove spam. Optionally, it may also help to remove duplicate content (e.g., retweets) because, like spam, it can dominate the results of any automatic algorithm. It can also dominate the results of manual analysis, such as a content analysis of a random sample, because extensively duplicated content might be frequently selected to occur in any random sample. If it is unclear whether duplicate content might be interesting for an analysis then the texts could be analysed twice, with and without duplicates. Spam and irrelevant content should always be eliminated as far as possible, though.

Stage 2a Irrelevant content

Irrelevant content concerns texts that do not relate to the topic investigated such as discussions of Cheddar Gorge when the topic is cheese. This can be difficult to remove because it needs to be manually identified. If there are a lot of texts then it will be impractical to manually check them all and so methods need to be used for the large-scale elimination of irrelevant content. Essentially, the procedure is to randomly sample the texts to look for systematic types of irrelevant content and then devise a method to mark it all as spam. The following ideas should be practical in most computer programs used to store the texts, including spreadsheets. Most of these methods may not be perfect in the sense that they will remove some relevant content and some irrelevant content will remain. This may be unavoidable but the methods are still worthwhile if they improve the overall quality of the remaining data, by increasing the percentage of relevant texts.

Irrelevant original query. Some of the original queries used to generate the data might produce all or nearly all irrelevant content. This may occur if the terms match unwanted topics. For example a the query *CBI* that was intended to match *Centrum tot Bevordering van de Import uit ontwikkelingslanden* in the Netherlands might instead match content relevant to the *Confederation of British Industry*. In a case like this, all content originating with the badly matching query should be removed.

Irrelevant source. Some data may come from individual irrelevant users. If any users generate many irrelevant or spam texts then all of their texts should be removed.

Partially relevant original query. Some of the original queries used to generate the data might produce a significant proportion of irrelevant content if the terms match unwanted topics in addition to the desired topics. For example a query *UCL* that was intended to match *University College London* might return some matches relevant to the *UEFA Champions League*. This is a tricky case to deal with because it will be difficult to separate out the relevant from the irrelevant content. The methods below are imperfect but designed to improve the overall quality of the set of texts.

- *Manual query-assisted irrelevant content removal* A time-consuming but potentially effective manual method is to make a list of all data matching the partially relevant query and then manually identify and remove the irrelevant data.
- *Query-based irrelevant content removal* If there are a lot of texts then an alternative to manual methods is to look for keyword searches matching irrelevant data and remove all matching searches. In the above case, searches like the following could be used: *UCL football; UCL goal; UCL champions; UCL league; UCL match; UCL game*. If a system is able to list the words that co-occur the most with the broad keyword (e.g., *UCL*) then this may help to suggest words to use in the queries.
- *Query-based relevant content inclusion* If it is easier to design queries to match relevant content than to design queries to match irrelevant content then all texts matching the broad query (e.g., *UCL*) could be removed and then relevant content matching appropriate queries (e.g., *UCL university, UCL London, UCL college, UCL studies*) could be added back in.

It is important to use the query-based methods with care because they may bias the final results. For example, the query *UCL London* might introduce a bias towards texts discussing the geographic location of *UCL*.

Stage 2b Spam

The rules for spam elimination are essentially the same as those for irrelevant content removal above. These rules apply because the spam may match some of the original queries but not others. Spam from a single source may be easy to identify and remove in bulk if it follows a pattern that is easy to capture with a query. For example, if all tweets contain a particular word that is irrelevant to the topic, such as *poker* or *Macys* then all texts matching the word could be removed. Repetitive spam may also be removable by duplicate filtering, as discussed below.

Stage 2c Duplicate filtering

Duplicate content can cause problems with automatic analyses because it can dominate time series and keyword analyses. The underlying issue is that it is easier to duplicate content (e.g., click a button to re-post it, or copying and pasting a text to resend it) than it is to create new content (e.g., compose a new tweet) and so replicated content exaggerates the importance of the texts. For example, if someone posts a funny joke about *Wensleydale* cheese and it gets shared and reposted then an automated analysis of cheese-related texts might conclude that *Wensleydale* was a particularly heavily discussed variety. Filtering out duplicate messages would remove this problem. Nevertheless, the fact that content has been replicated can be an indicator of its importance and so it does not seem correct to completely ignore replicated texts.

The suggested solution to this problem is to analyse replicated content separately from the main analysis and that the main analysis should exclude all replicated data. This can be achieved by removing all texts that are duplicates of a previous text. This is possible in some systems and also in spreadsheets (e.g., by sorting the spreadsheet on the text column and marking for removal all rows in which the text column is identical to the previous, for

example with IF statement like =IF(C3=C2,1,0)). The replicated texts can be analysed separately by compiling a league table of the most replicated texts and analysing them individually to see what insights, if any, they can give into the topic of the investigation.

Stage 3a: Content analysis (all projects)

A content analysis of a random sample of texts is essential to find out what kinds of issues are discussed. Content analysis is likely to be the main research method for projects with few texts for which the other methods do not work. It is also useful to give background context when there are enough texts that more automated methods can also be used.

The texts to be analysed should be selected randomly, starting with a larger random sample than is needed so that irrelevant texts can be discarded without having to start again with the sample. As discussed in the content analysis chapter, a content analysis should be based around multiple sets of categories that are relevant to the research goals and match the type of content found in the texts. If information is also available about the originators of the texts then they could also be included in the content analysis too (e.g., age, gender, nationality, orientation towards the topic investigated). As part of the content analysis, a judgment should be made for each text analysed as to whether it is relevant to the topic investigated and irrelevant texts should be discarded.

If the content analysis is the main method to be used then particular care should be taken with all stages of it: identifying relevant categories; ensuring accurate and comprehensive category descriptions; and testing to ensure high inter-coder consistency.

Stage 3b: Relative word frequency analyses compared to general texts (all except small projects)

A semi-automated way to help identify the most discussed topics within in a collection of texts is to analyse the frequency of the individual words in the collection. For example, if the word *riot* was the most frequent word in a collection of news texts then riots were probably one of the main topics of discussion. Simply examining the most frequent words will not help, however, because the most frequent words in any corpus will be words that are generally common (e.g., *a*, *the*, *it*). Relative word frequencies should therefore be examined instead. Words that are *relatively* more frequent in one collection of texts than another are more likely to be useful pointers to the topics discussed than the generally most frequent words.

Relatively high word frequencies in a collection of texts can be identified by calculating the relative word frequencies for each word in the collection (the number of texts in the collection containing the word divided by the total number of texts in the collection) and compare this to the relative word frequencies of the same words in a general collection of texts. Words with high relative word frequencies in the first compared to the second may point to important topics in the first collection. For example, if the term *baton* occurs in 10% of the texts in a topic-specific collection but in only 2% of the texts in a general collection then this word is likely to be important for discussions of the topic. The significance of the difference between these two values (10% and 2%) can be tested for with a *differences in proportions test*, as built into some software or available online. A differences in proportions test produces a z-value (related to the statistical normal distribution) that is larger when the difference between the two proportions is more significant. Hence, listing the words in descending order of z value ensures that the most significant differences are at the top.

The top few (e.g., 20) words (ordered by z value) in the list of relatively high frequency words must be manually analysed to investigate *why* they are apparently important to the topic investigated. A simple qualitative way to do this is to read some or all texts containing the word to see why it occurred so often. It is possible that random factors caused the word to appear in the list but if there is a systematic cause then this should be reflected in one or more themes recurring in texts containing the words. For instance, if the term *baton* occurred with a high frequency in texts mentioning the police within a riots corpus then reading the matching texts (i.e., those containing both police and baton) might reveal that

there was a big discussion about whether the police used their batons too frequently during the riots.

The end result of the word frequency analysis should be a list of specific sub-topics that received significant discussion within the topic-specific collection of texts. This list should give more specific and detailed insights than given by a content analysis of a random sample of texts and so the two approaches can complement each other. This method can be used on any size collection of texts but is likely to work better on larger collections because random factor might dominate the word frequency lists in smaller collections.

Stage 3c: Relative word frequency analyses for collections of texts within the topic (all except small projects)

Relative word frequency keyword lists can also be obtained from *within* a topic-specific collection of texts by comparing the relative word frequencies in one part of the collection to the relative word frequencies in another part. The easiest way to do this is to compare the relative word frequencies within the topic-specific collection with the relative word frequencies of these terms in the remaining texts. Terms that are relatively frequent in texts matching the query can give insights into the subtopics associated with the query. For example, in a topic-specific collection of texts about riots, words that are relatively frequent in texts mentioning *police* could give insights into the aspects of policing that were most frequently discussed within the original topic.

The statistical significance of the difference between the proportion of texts matching the query and the proportion not matching the query can be tested for with a chi-square test on the 2x2 table illustrated in Table 8.1. Hence, listing in descending order of chi-square value the terms that have a higher relative frequency in the topic-specific collection than in the general collection will tend to give the most topic-relevant terms at the top of the list. These terms can suggest the important sub-topics within the collection, as discussed below.

Table 8.1. Chi-square table for testing the significance of the relative frequency of the term *baton* within texts mentioning the police in the test collection of texts (10%) compared to the topic overall (2%). The chi-square formula (which is in spreadsheets and text analysis programs) can be used to calculate the chi-square statistic from this table (it is 19.020, with Yates' correction).

	Topic-specific texts matching the query <i>police</i>	Topic-specific texts	Total texts
Containing <i>baton</i>	10	20	30
Not containing <i>baton</i>	90	980	1070
Total texts	100	1000	1100

The relatively high frequency words should be ordered in decreasing order of chi square value so that the most significant differences are at the top. The top few words (e.g., 20) must again be manually analysed to investigate *why* they are apparently important to the topic investigated. For instance, if the term *baton* occurred with a high frequency in texts mentioning the police within a riots corpus then reading the matching texts (i.e., those containing both *police* and *baton*) might reveal that there was a big discussion about whether the police used their batons too frequently during the riots. This should give insights into the key factors within the subtopics in the collection.

Stage 3d: Time series for keywords for large collections of texts (large projects collected for at least a week)

Time series analyses can identify specific events within a topic as well as trends over time. A time series analysis is appropriate for large collections of texts that have dates associated with the texts that span a significant period of time – a week would normally be the minimum. The purpose of the analysis is to identify changes in topic or volume over time, including specific

points in time at which significant events happened. For these analyses, software is needed that is able to plot time series trend graphs from the texts gathered.

Detecting general time-specific events. A simple way to detect time-specific events is to plot the *frequency* of texts in the collection over time and then look for spikes or other sudden changes in the shape of the graph that indicate changes in the total amount of texts for the topic. Any sudden change is likely to indicate an event that started at a specific time, unless the change is an anomaly of the data collection process. To identify the nature of the event that caused the change, a random sample of texts at the time of the change should be read to identify a pattern associated with the date.

Detecting temporal trends. Time-specific trends, such as a gradual increase or decrease in interest over time, can be detected by examining a graph of the frequency of the texts over time. It may be difficult to be sure that the trend is due to a genuine change in interest in the topic rather than a general increase or decrease in use of the service producing the texts or fluctuations in the effectiveness of the data collection method, however, so this evidence should be treated very cautiously.

Detecting time-specific events for sub-topics. Events that are significant to a particular sub-topic can be identified by plotting a graph of the frequency of texts within the subtopic. If the sub-topic can be identified by a query then this graph is of the proportion of tweets in the subtopic that match the query. As for the other two methods described above, the graph can show whether there were specific events for the sub-topic that caused changes in the shape of the graph or whether interest in the subtopic increased or decreased over time, relative to the whole topic.

The end result of a time series analysis should be a list of time-specific events that have affected the development of the discussion of the topic, as well as general information about overall trends in interest in the topic or subtopics. These should give insights into how the discussion evolved and should therefore complement the other more static analyses discussed here.

Stage 3e: Time series scanning for large collections of texts (large projects collected for at least a week)

Time series scanning is an advanced technique for automatically identifying significant time-specific events within a collection of texts. It is a more powerful version of the time series for keywords method described above. It works best on large collections of date-stamped texts. Time series scanning can only be conducted by a computer program that will complete the following steps.

- Create a list of all words used in all texts in the collection.
- For each word, create a time series graph of the relative frequency of the word on each day. The relative frequency for a day is the proportion of texts on that day that contain the word.
- For each time series graph, as above, calculate the size of the biggest spike in the graph. The size of a spike is the difference between its peak relative frequency and the average relative frequency for the same word for all days before the spike.
- List the words with the highest spike sizes.

Each word in the list of words with the largest spikes may associate with an event at the time of the spike. The list should give the date of the spike so that the cause of the spike can be identified by reading texts containing the word on the day of the spike. Some spikes may be anomalies or associated with spam or irrelevant events and these should be discarded.

Like the time series analysis stage, the output of this stage should be a list of topic-relevant events, perhaps extending the list generated by the previous analysis.

Stage 3f: Summary and thick description: Integrating the analyses (all projects)

The analyses described above should give an overview of various different aspects of the types of contents discussed in the set of texts (from the content analysis), some of the main

subtopics discussed (from the word frequency analysis) and some particularly important events discussed and any trends in interest in the topic (from the time series analysis and time series scanning). Once they are all complete, these results should be integrated together and summarised. The limitations of the methods should also be acknowledged and explained. The limitations should include information about the presence of spam or irrelevant content that could not be removed during the data cleaning. If hypotheses were proposed before the start of the research then these should also be discussed in the light of the findings. Finally, the implications of the results should be discussed, including: a realistic assessment of the value of the information gained by the method; the implications of the results for any relevant theories that might apply; and the types of future research that would be useful for the set of texts.

Example 1: UK Tweets about the Egyptian revolution

This example monitored the unrest in Egypt from 22 January 2011, just before the initial protests that led to the overthrow of Hosni Mubarak, to 14 February 2011, just after his resignation. The monitoring used the term *Egypt* and the hashtag *#egypt* and, since it started before the protests it did not use stage 1 to test and refine the queries, so the investigation started with stage 2. The monitoring focused on the perspective of the UK and only includes English language tweets with a UK location.

Stage 2: Post-hoc spam, irrelevant content, and duplicate filtering. For stage 2a and 2b, eliminating irrelevant content and spam, this started with browsing the data and then developing queries to eliminate irrelevant texts. The first query was *@stephenfry* because there were many texts related to this user due to him being popular on Twitter and having just returned from holiday in Egypt at the start of the data collection period. Hence, all texts matching the query *@stephenfry* were removed. Additional checking produced a second query *sheikh* because many tweets were about holidays in the popular Sharm el Sheikh resort. Other queries used include *cheap*. The words holiday and airport were considered but not used because many matches were relevant. Instead, selected individual matches from these searches were manually marked as spam. Finally, duplicate items were removed (stage 2c) but only after recording the highest frequency duplicate items for later analysis (e.g., "RT Uninstalling dictator ... 100% complete" with 183 matches).

Stage 3a: Content analysis. A content analysis of 100 tweets gave the results in Table 8.2. The main two types of tweets were simple news fact sharing and analysis or commentary on the events, followed by expressions of support.

Table 8.2. Results of a content analysis of 100 tweets from the Egypt collection, ignoring irrelevant tweets.

Topic	Tweets
Analysis – discussing an aspect of the events in Egypt, including criticisms of leaders	24
News – reporting a recent fact about Egypt	22
Expression of support – message mainly expresses sympathy towards the protesters	18
Internet – information about internet access in Egypt	10
Interest/Worry – expresses interest or concern about Egypt	8
Pictures/Video – information about pictures or videos of events in Egypt	6
Holiday – discusses an aspect of the impact of the events on holidays	4
Question – asks a question related to the events in Egypt	4
Twitter – discusses and aspect of the use of Twitter in Egypt	4

Stage 3b: Relative word frequency analyses. Table 8.3 reports the top 50 words that were relatively frequent in Egypt tweets compared to general tweets from the UK. These contain mainly expected terms but it is clear that there was a focus on the Egyptian leader, Hosni Mubarak, and the protests themselves. Nevertheless, it is clear from the results that

independent activists and news sources were particularly important on Twitter, with official sources like the BBC being relatively marginal for their size. Al Jazeera and its journalists seemed to be the main major news source for the events. The Muslim Brotherhood was relatively little discussed – occurring in 1% of tweets, with many of these stating that they were not heavily involved. It is perhaps interesting that the commentators were not all based in Egypt but also in the UK, South Africa and Australia.

Table 8.3. Words most important in Egypt tweets compared to general UK tweets (z-score). Words that are usernames suggest the person's tweets were extensively retweeted.

Corpus	Word	Gen rel freq	z	Comment and other top related terms not shown
27.3%	#jan25	0.3%	643.6	Start date of the uprising. Also #25jan
9.4%	mubarak	0.2%	319	Overthrown leader, president Hosni Mubarak. Also #mubarak, hosni, president, regime
8.4%	protest	0.2%	287.9	Also protester, protestor, unrest, revolution, uprising
5.9%	egyptian	0.2%	205.6	
4.5%	cairo	0.1%	204	Location of main protests. Also #cairo, tahrir, #tahrir, square
1.1%	@alaa	0.0%	184.9	Alaa Abd El-Fattah, Egyptian blogger and political activist. Prominent in the revolution.
0.6%	@asa_wire	0.0%	181.6	Asa Winstanley, London-based investigative journalist with a focus on the Middle East.
2.3%	jazeera	0.0%	172.3	Al Jazeera news organisation. Also al, @alanfisher, @ajenglish, aje, @dima_khatib
0.5%	@eanewsfeed	0.0%	158	EA Worldview – independent news and analysis source (used LiveBlog). Also liveblog
0.5%	@muschelschloss	0.0%	148.6	German activist blogger.
0.7%	@bbcworld	0.0%	144.1	BBC News (World)
0.3%	@whumba	0.0%	127.1	Activist tweeter from South Africa
0.3%	@riverdryfilm	0.0%	126	Pro-Palestine film.
0.3%	@khadijapatel	0.0%	125.3	South African online newspaper journalist
2.5%	army	0.1%	118.5	
1.7%	democracy	0.0%	118.4	
0.4%	@tweetminster	0.0%	116.5	London-based news identification utility
0.3%	@nohaatef	0.0%	116.3	UK-based Egyptian political blogger.
0.7%	gamal	0.0%	115.8	Gamal Mubarak, son of Hosni Mubarak. Also sw1x, Wilton
1.4%	#sidibouzid	0.0%	114.5	Protests in Sidi Bouzid, Tunisia
0.3%	@jpmlynch	0.0%	112.7	Amnesty International worker in the Gulf
0.4%	@sharifkouddous	0.0%	109.3	Egyptian independent journalist.
0.2%	darkpolitricks	0.0%	108.2	"dedicated to investigating the dark side of politics including corruption, propaganda, the police state, war on terror"
0.5%	#jan28	0.0%	106.9	Friday of Anger protests in Cairo
0.4%	@breakingnews	0.0%	105.4	News aggregator.
0.2%	@guardian_world	0.0%	102.1	Manchester Guardian world news
0.2%	@jinjirrie	0.0%	100.3	Australian writer and political commentator.
1.0%	brotherhood	0.0%	100.1	Muslim Brotherhood, Egyptian political group.

Stage 3c: Relative word frequency analyses for collections of texts within the topic. Most of the subtopics identified in Stage 2b seem to be straightforward so only an analysis of the tweets mentioning brotherhood (Table 8.4) is reported here. The results show that the Muslim Brotherhood was mainly discussed in the context of formal political processes rather than in terms of participation in events.

Table 8.4. Words most important in Egypt UK tweets mentioning brotherhood compared to general Egypt UK tweets (z-score).

Word	% of matches within Egypt tweets	Chi square	Comment
brotherhood	100%	25707	
muslim	90%	13239	
talk	17%	883	Call for talks between the Brotherhood and the government.
opposition	11%	426	Mentioning the Brotherhood as the main elected opposition party, despite being illegal.
member	7%	418	Arrest of a member of the Brotherhood
candidate	2%	160	Brotherhood statement that it will not run a presidential candidate in the next election.
secular	2%	157	Contrasted with secular parties or goals.
group	5%	101	
negotiate	2%	95	Refusal to negotiate with Mubarak.
islamist	3%	80	Discussions about whether the Brotherhood is Islamist and whether Islam will win from the Arab Spring.

Stage 3d: Time series for keywords. Figure 8.1 reports the overall number of tweets per hour relating to Egypt. It is clear that the interest started with the protests on 25 January, but also that there was a large and sustained increase in the level of interest that started on the 28th of January (the Friday of Anger) and continued until about the 4th of February, then died down a little until a big spike at 4pm on February 11, when an announcement was made that Mubarak would step down.

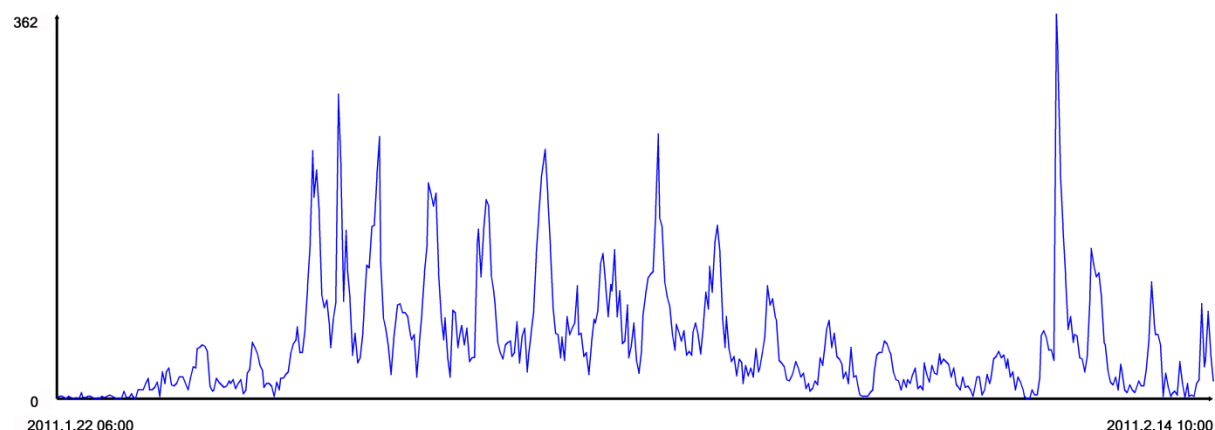


Figure 8.1. Time series for all UK Egypt tweets.

Figure 8.2 shows the frequency of tweets mentioning one of the top terms, @alaa, showing that this blogger was particularly important at the start of the protests, from January 25 to February 1, but was less important after that. Figure 3 shows that the hashtag #jan25 retained

its popularity long after January 25 and only started to decline in popularity on about the 9th of February.

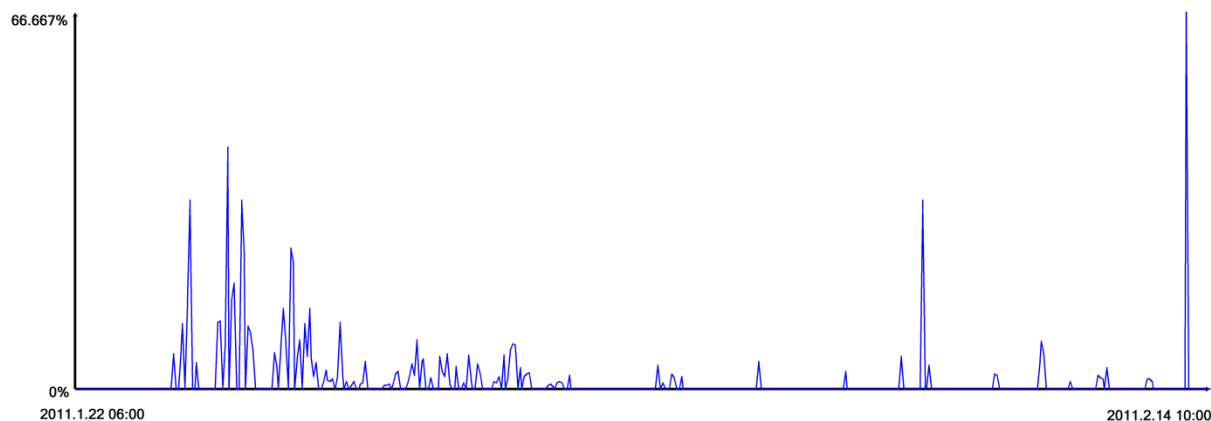


Figure 8.2. Time series for tweets mentioning @alaa.

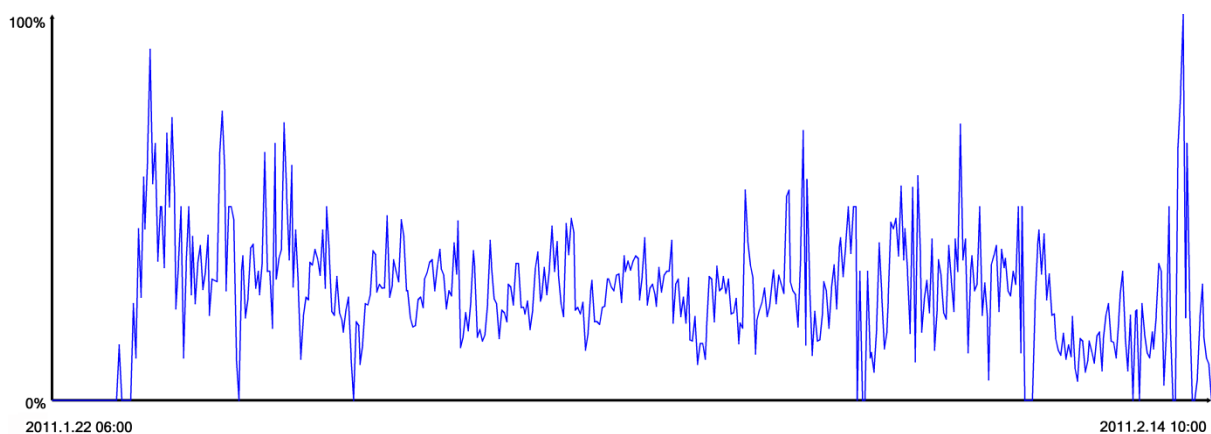


Figure 8.3. Time series for tweets mentioning #jan25 or #25jan.

Stage 3f: Summary and thick description. The results show that interest in the Egypt revolution from UK tweeters started a few days after the initial protests. This interest focused on news analysis and expressions of support, with some particular concerns with internet-related aspects. The specific topics of interest included Hosni Mubarak and events around Tahrir Square in Cairo. The main opposition group in Egypt, the Muslim Brotherhood was hardly discussed at all. The news seemed to come from independent bloggers and journalists as well as Al Jazeera rather than from major UK or US news sources. Interest seemed to flag after a week except for a spike of interest when the resignation of Mubarak was announced.

Example 2: Autism

The purpose of this case study is exploratory and descriptive: to characterise the way in which autism is discussed on Twitter in order to assess whether a future analysis of autism-related tweets might give useful insights into any aspects of autism. For example, if the tweets often gave parents' reactions and worries about a first diagnosis of their child then the outcome of the case study might be a recommendation for an in-depth study of this issue.

The first stage is to create a set of highly focused topic-relevant queries. A brainstorming session led to two queries: autism and "autistic spectrum". These were then piloted by using a program (Mozdeh) to download a sample of matching tweets. Browsing these tweets revealed no systematic causes of irrelevant tweets. An examination of a list of terms frequently co-occurring with autism included many irrelevant terms but also one relevant term, *Asperger's*, which was chosen as a new query term. The list was piloted again and found to be OK. A full-scale data collection was then undertaken by using the same program (Mozdeh) to monitor the queries on Twitter for a month.

At the end of the month the combined set of matching tweets was analysed. First, duplicate tweets were removed so that the results wouldn't be dominated by popular retweets. A co-word analysis of the tweets was then used, with unexpected words co-occurring with any of the keywords checked as potential indicators of spam. This produced the keyword *poker*, which occurred almost exclusively in spam tweets so all tweets containing the term *poker* were marked as spam and removed.

The main investigation started with a content analysis of a random sample of tweets by three coders. Because this was a descriptive study, the categories were chosen after reading an initial sample of 50 tweets and were extended, when necessary, during the main classification by the first coder. The other two coders were given the classification scheme of the first coder and recoded the same tweets. The inter-coder consistency kappa values from this were not high but were high enough to be acceptable, so the results could be used (Table 8.5).

Table 8.5. Content analysis results for 200 autism-related tweets.

Category	Percentage
Autism help ideas or information about autism	28%
People with autism achieving or about their capabilities	20%
Discussing whether someone is autistic or mentioning that a person is autistic	11%
Defending or supporting autistic people	10%
Autism problems	7%
Actions of autistic people	5%
Autism in fiction	5%
Autism as an insult	5%
Learning about autism	5%
Autism carers	2%
Autism meetings	2%
Requests for help for an autism problem	2%

The top 20 words in the autism corpus compared to a general collection of tweets were investigated for specific important themes, with the Table 6 results.

Table 8.6. Top 20 words unusually frequent in autistic tweets.

Word	Corpus	Gen rel freq	z	Comment and other top related terms not shown
autism	39.7%	0.0%	1604	Term describing autism. Also: autistic, aspergers, syndrome, spectrum, disorder
rt	20.5%	0.0%	1211	Lots of retweets – an unusual amount of sharing in Twitter
sap	6.3%	0.0%	512	German company SAP announced an initiative to recruit autistic people. Also: recruit, hire, German, software, programmer
grandin	0.6%	0.0%	199	Actions of autism researcher Temple Grandin
sensory	0.7%	0.0%	174	Discussions of sensory aspects of autism and sensory therapies
dsm-5	0.4%	0.0%	169	American Psychiatric Association's (APA) Diagnostic and Statistical Manual of Mental Disorders.
diagnosis	0.8%	0.0%	167	Aspects of the diagnosis of autism.
keef	0.5%	0.0%	165	Rumour that rapper Chief Keef has Asperger's
children	4.5%	0.2%	151	Ways of diagnosing or helping autistic children
awareness	2.0%	0.0%	150	Discussion of the importance of autism awareness.

Table 8.7 reports the top words associated with the term Children in the corpus to find out what aspects of children are discussed. The results suggest that activities and therapies for children are the most common issues.

Table 8.7. Top 10 words unusually frequent in tweets from the autism collection that contain the term children, compared to tweets not containing the term children.

Word	children	Chi square	Comment
children	100%	23811	
with	52%	558.2	Children with autism.
therapy	7%	498.4	Study finds enrichment therapy effective among children with autism
enrichment	4%	490.6	
treating	3%	455.7	Blog post on how to treat children with autism.
for	35%	191.2	
activities	2%	157.8	Activities for children with autism
ana	1%	152.4	Popular or artificially promoted Blog poster
among	2%	144.5	
parent	7%	141.2	

The time series methods were not used for this issue because it is an on-going topic rather than an evolving topic.

Overall, the findings suggest that Twitter is used for sharing a wide variety of information about autism, but particularly for help or information about autism and for the activities or achievements of autistic people. In a minority of cases the tweets mention people that are autistic or discuss whether someone is autistic. A minority of tweets defend or support autistic people. Retweeting seems to be particularly common for the topic of autism, suggesting that either it is a very supportive environment or information is actively shared.

Children with autism are also a common topic, and treatments and activities for children in particular. Perhaps related to this, the issue of diagnosing autism is widely discussed, including with relatively technical details, such as mentions of an official clinical manual for it.

Summary

This chapter described a descriptive method for analysing web texts that can be used either for an initial pilot investigation into a set of texts or as a full scale detailed descriptive study. The methods described should be enough to give useful findings but can be supplemented with other methods, where appropriate. For example, it may be possible to gain extensive information about the users that wrote the texts and this information could be added, perhaps as tables or graphs. Similarly, the texts might have a particular structure, such as dialogs, threads, or multiple exchanges between named users, and this might be exploited for further analyses.

The main text of this chapter illustrated the application of the methods for specific types of text and topics, but the methods are applicable for any large collection of texts that have associated creation dates. In practice, the methods also depend upon access to appropriate software for gathering and processing texts, and the online component of this chapter describes some examples.

The method depends upon getting enough data that is relatively spam-free and also relatively free of irrelevant content. As a final point of warning, if collecting the texts in real time then it is particularly important to test the data collection method (e.g., queries) before starting the data collection in order to ensure that they get mainly relevant texts and that they get enough texts.

9. (S1) Automatic Search Engine Searches: Webometric Analyst [Updated throughout]

The chapters so far in this book have introduced the main webometric techniques and described how to carry them out without the use of specialist software, although the need for software is mentioned in a few places for tasks that are difficult to do without automation. This chapter introduces the first specialist tool for webometrics, one to automatically submit searches to search engines and to process the results.

An automatic search engine query submitter (not standard terminology) is a program that is capable of automatically submitting queries to search engines and then downloading, saving and processing the results. Although there are several programs that can do this for webometrics, including VOSON (voson.anu.edu.au) and Issue Crawler (www.issuecrawler.net) (Rogers, 2004) this program focuses on Webometric Analyst, a free Windows program available at lexiurl.wlv.ac.uk. Both the automatic downloading and processing capabilities are important to make medium or large-scale webometrics possible, unless a web crawler is used instead.

Although it is possible to write a program to automatically submit queries via a web browser and save and process the results from the browser, automatic search engine query submitters typically use the web services set up by the search engines, typically known as Applications Programming Interfaces (APIs), and the main current source is for Microsoft's Bing. The Bing API currently allows 5,000 searches to be submitted free per month, per user and return the results in a very computer-friendly format (e.g., XML). More searches can be submitted per month but these need to be paid for. The results are not the same as the equivalent results in the online search pages (McCowan & Nelson, 2007) but often seem to be reasonably similar. Automatic search engine query submitters typically start by the user entering either a list of pre-formulated queries or information such as web page URLs or web site domain names that would allow the program to formulate the queries itself. The next stage is that the program submits the queries to the chosen search engine and saves the results. Finally, when the results are all gathered they are processed and displayed in some format, such as a table of results or a network diagram. Sometimes the three stages are automatic and sometimes user actions are required between them. This chapter introduces Webometric Analyst and gives examples of how to use it for common problems.

Introduction to Webometric Analyst

Webometric Analyst is a free program designed to gather data from search engines via their APIs for webometric purposes. At the start of 2013, it could access Bing via its API. In addition to downloading data from the Bing search engine and saving the results in simple text files, Webometric Analyst is able to process the data in order to produce reports or diagrams and can also help with generating link searches for a set of web sites or URLs. The common uses of the program can be accessed via the Wizard that appears when it starts and advanced features are accessible via its classic interface for those who wish to do something non-standard or more complicated.

As introduced above, the typical Webometric Analyst analysis contains three stages. First, the researcher must generate a plain text file containing a list of the searches to be submitted to a search engine. If URL citation or title mention searches are needed, then Webometric Analyst has features to help convert a list of URLs and/or titles into appropriate lists of searches. Second, Webometric Analyst will submit these queries to Bing and save the results in a few simple files. Finally, Webometric Analyst can process the simple results files to give more detailed and formatted results. The second and third stages can be automated by the Webometric Analyst Wizard for a few standard tasks.

Webometric Analyst automatically attempts to get the maximum number of URLs for each query, up to 1,000. Since search engines return results in sets or "pages" of 10, 50 or 100 at a time, to get the full list of results, Webometric Analyst has to submit multiple queries, one for each set of results, and then it automatically merges the results together when saving

or processing them. This process takes time and uses multiple searches from the API limit – typically 20 per search. Hence, if the full URL lists are not needed then Webometric Analyst should be instructed to get only the first results page and not to get any subsequent pages. This is achieved by unchecking the *Get all Matching URLs* option in the *Search Options* menu.

Webometric Analyst Web Impact Reports

Webometric Analyst uses a three stage process to calculate Web Impact Reports – i.e. to summarise the extent to which a set of documents has been mentioned on the web. This process is supported by the Wizard as follows.

A Web Impact Report (WIRE) can be created in a few steps with Webometric Analyst. First, download the program from the web site lexiurl.wlv.ac.uk. Second, create a plain text file with one search term or phrase. The easiest way to create a plain text file is to use Windows Notepad (normally accessed via Start|All Programs|Accessories|Notepad). Each line should contain one search term or phrase, as would be entered into Google for a single search and there should not be any extra spaces in the file or blank lines. If any search is several words then it can be put into quotes (straight quotes rather than smart quotes) to ensure that the words occur consecutively in any matching documents. Unless for example, the following book list could be used.

"Link analysis an information science approach"

"Mining the web discovering knowledge from hypertext data"

"Information politics on the web"

Once the file is created, start Webometric Analyst and select the Web Impact Report option from the first Wizard screen and click OK.

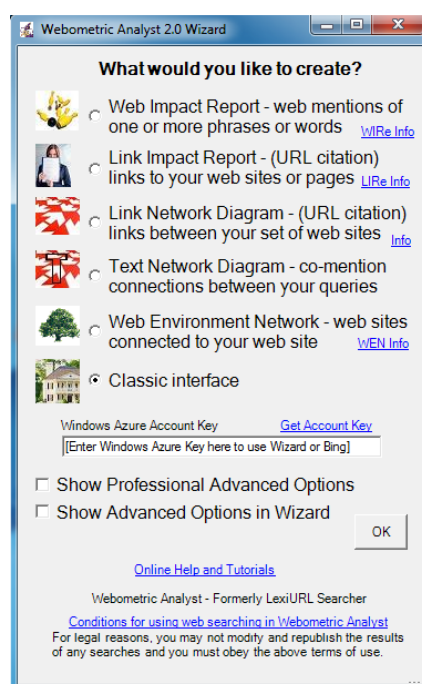


Figure 9.1. The initial Webometric Analyst Wizard (professional version).

The second wizard screen requests the plain text file with the searches: click OK and select the file in the dialog box that appears.

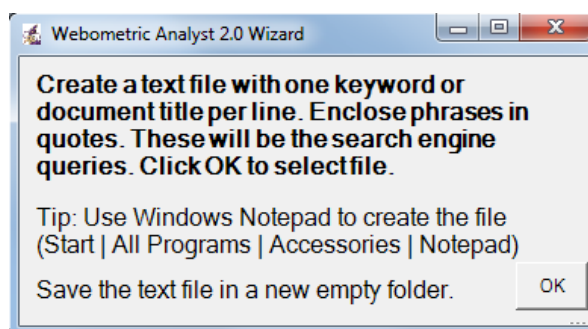


Figure 9.2. The second Webometric Analyst Wizard.

Once the file is selected, Webometric Analyst will start gathering data and after several minutes or half an hour will display a report in the form of a set of interlinked web pages. The main table, accessed by clicking the *Overview of results* link, contains the number of URLs, domain names, web sites, STLs and TLDs matching each search term or phrase. All data in the report is derived from Bing.

Search Engine Results Report

Introduction

This report presents the results of a series of search engine queries, obtained from the URLs returned by the search engine.

- Data from Bing via its **Applications Programming Interface**. The results may be incomplete because search engines do not report all matching URLs for a search and do not index all web pages. Search engines also return a maximum of 1000 results per query and sometimes substantially fewer.
- Data gathered on 04 March 2013.

Any queries with zero results are not shown in this report.

Click to view the [main table of results](#).

Figure 9.3. A section of the main page of a Web Impact Report.

A few less common types of Web Impact Report can also be created but for this the classic interface will be needed rather than the Wizards. The instructions below explain how to create a standard Web Impact Report with the Classic Interface – this method can be customised for alternative WIRes.

Web Impact Reports – classic interface example

This section gives a step-by-step example of using the classic interface to create a Web Impact Report. To give an overview: the user must construct a text file containing a list of searches first. Then the user must select a search engine and any search options and instruct Webometric Analyst to start submitting searches and reporting the results. Finally, when the searches are complete then the user must select a processing option and instruct Webometric Analyst to apply it to the appropriate results file. Below is a simple example to illustrate how this process works – it assumes that the program has been downloaded from lexiurl.wlv.ac.uk.

The example is to compare the web impact of three books: *Link Analysis: An Information Science Approach*, *Mining the web: Discovering Knowledge from Hypertext Data* and *Information politics on the web*.

1. *Input data generation* Create a text file containing the three book titles in quotes, one title per line. This file will be the input for Webometric Analyst and the lines are each single searches. The quotes are included to ensure that the searches are exact phrase matches. The file should be constructed in Windows Notepad (Start/Programs/Accessories/Notepad) or a similar text editing program but not in a

word processor. The resulting file might be called test.txt and contain the following contents.

"Link analysis an information science approach"
 "Mining the web discovering knowledge from hypertext data"
 "Information politics on the web"

2. *Running the searches* Move the file test.txt to a new empty folder in Windows to protect your computer from Webometric Analyst accidentally overwriting or deleting files. Now start Webometric Analyst and click the "Search" button and select the file test.txt. The searches will be submitted over a period of a few minutes and Webometric Analyst will report when they are complete. When the search is complete, three new text files will be found in the same folder as the original, with file names including "long results", "short results" and "result counts per page". If all that is needed is the hit count estimates, then these can be found in the short results file, shown below. The first number is the hit count estimates from the first results page and the last number is the actual number of URLs returned by the search engine. In this case the initial hit count estimates are wildly inaccurate but the last column seems more likely to give reasonable estimates (see the search engine chapter below, however).

450	"Link analysis an information science approach"	152
7840000	"Mining the web discovering knowledge from hypertext data"	729
90700000	"Information politics on the web"	310

If more information is needed than just the number of matching URLs then this can be extracted from the "long results" file by using Webometric Analyst features for producing reports based upon search engine results.

3. *Creating the reports* A standard set of summary reports can be produced from the raw search engine results in the "long results" file. This lists the matching URLs returned, some text extracted from the matching page and also repeats the search used. The number of URLs returned is never over 1,000, which is the maximum returned by search engines. To create a set of standard summary reports, select "Make a Set of Standard Impact Reports from a Long Results File" from the "Reports" menu, select the long results file and follow the instructions. This generates a set of web pages summarising the searches. To view the results, double-click on the file index.htm in the new folder created to load it into a web browser. This web page lists the main summary statistics and gives links to the more detailed results. Figure 9.4 contains an extract from the overview.html page, which summarises the main results.

Overview of search results

Table listing the URLs of pages matching the queries submitted. In addition, it contains the number of domains, sites, STLDs and TLDs containing one or more URL matching the query, as derived from the URL list.

Name	Base query	URLs	Domains	Sites	STLDs	TLDs
-	"Link analysis an information science approach"	152	111	94	26	23
-	"Mining the web discovering knowledge from hypertext data"	729	560	452	80	54
-	"Information politics on the web"	310	234	191	36	33

Figure 9.4. Overview impact summary from a Webometric Analyst report in the page overview.html.

As shown in Figure 9.4, the main table of a Webometric Analyst impact report lists the number of URLs, domains, web sites, Second or Top-Level-Domains (STLDs) and TLDs matching each search, as calculated from the long result files. This page can be reached from the main index.html page by clicking on the "Overview of results" link. The most reliable impact indicator is normally the number of domains rather than the number of URLs due to

the possibility that text or links are copied across multiple pages within a web site. The results suggest that *Mining the web: Discovering Knowledge from Hypertext Data* has the most web impact, based upon its domain count of 560. The results also include full lists of matching URLs, domains, sites, STLDs and TLDs for each query: clicking on the appropriate links on the home page reveals these lists. Figure 9.5 illustrates such a list: the top half of a table of domain names of URLs matching the second query.

Domains of pages matching the base query: "Mining the web discovering knowledge from hypertext data"

The table lists the domains of pages matching the base query: "Mining the web discovering knowledge from hypertext data". The URLs column lists the number of URLs returned by the query with the given domain.

Domain	URLs	%
mse.sem.tsinghua.edu.cn	3	0.4%
www.zbh.uni-hamburg.de	3	0.4%
hpc.isti.cnr.it	2	0.3%
www.vbyte.com	2	0.3%
dblab.ssu.ac.kr	2	0.3%
www.5yiso.cn	2	0.3%
webspace.ulbsibiu.ro	2	0.3%
www.di.unipi.it	2	0.3%
www.ricercaitaliana.it	2	0.3%

Figure 9.5. Part of a table of domains matching a search from the Webometric Analyst report page domains2.htm.

Also included in the web impact report are random lists of URLs matching each search, with a maximum of one per domain. These are intended for use in content analysis classification exercises, as discussed in the web impact assessment chapter. Finally, there is some information about the search at the top of the page, which can be edited in a web editor if necessary, and near the bottom of the home page is a link to a page with a comparative breakdown of TLDs in the STLDs returned for each of the queries.

For information about how to conduct a simple analysis of this data, see the bullet point list at the end of section 2.1.

It is important to note again that the results of an impact query do not give fair impact comparisons if the search engine has stopped giving results at 1,000. In fact, search engines sometimes stop after about 950 unique results and so it is reasonable to be safe by only relying upon the results if all the URL counts are below 925. See the advanced chapter for query splitting techniques for searches with more results than this.

Webometric Analyst Link Impact Reports

A Link Impact Report (LIRE) can be created in a few steps with Webometric Analyst. LIREs are not based upon hyperlinks but instead use URL citations because Bing does not allow searches for links any more. An URL citation is a mention of the URL of a web site in a page, usually of another web site. For example if a BBC web page mentioned the URL of the University of Wolverhampton, www.wlv.ac.uk, then this would be an URL citation. The URL has to be visible in the text of a page to count and it may or may not be also hyperlinked.

URL citations are like hyperlinks in the sense that they connect two pages through an URL and hence are a reasonable substitute for hyperlinks.

For a LIRe, first, download the program from the web site lexiurl.wlv.ac.uk. Second, create a plain text file with one domain name or URL per line. The easiest way to create a plain text file is to use Windows Notepad (normally accessed via Start|All Programs|Accessories|Notepad). Each line should contain just the domain name of the web site and not any additional file or path information and there should not be any extra spaces in the file or blank lines. The only exception is that if a web site shares its domain name with another web site then the full URL of the home page of the site should be given rather than the shared domain name. Once the file is created, start Webometric Analyst and select the Link Impact Report (LIRe) option from the first Wizard screen and click OK.

The second wizard screen requests the plain text file with the domain names and/or URLs: click OK and select the file in the dialog box that appears. Once the file is selected, Webometric Analyst will start gathering data and after several minutes or half an hour will display a report in the form of a set of interlinked web pages. The main table, accessed by clicking the *Overview of results* link, contains the number of URLs, domain names, web sites, STLDs and TLDs of URL citation links to each web site. For the rows in the table containing web sites with just a domain name, the figures are for URL citation links to anywhere in the web site but for web sites with additional path information the figures are for links to just the specified URL. All data in the report is derived from Bing searches.

A few less common types of Link Impact Report can also be created but for this the classic interface will be needed rather than the Wizards. The instructions below explain how to create a standard Link Impact Report with the Classic Interface – this method can be customised for alternative LIRes.

For information about how to conduct a simple analysis of this data, see the bullet point list at the end of the introductory part of section 3.3.

Link Impact Reports – classic interface example

The instructions in the classic example section above apply almost without change to link impact reports. A link impact report is an analysis of the web pages that link to any one of a set of URLs or Web sites. To create a link impact report, a list of URLs or domain names can be fed into Webometric Analyst and it will download a list of pages that link to them, via search engine searches, and then produce a summary report based upon the URLs of these pages. The main difference is that the searches used are not phrase searches like "Link analysis an information science approach" but are link searches like "linkanalysis.wlv.ac.uk" -site:wlv.ac.uk, as described in the link analysis chapter.

Webometric Analyst for network diagrams

Webometric Analyst can support all of the stages of gathering and processing data in order to produce a network diagram for the links within a collection of web sites. A network diagram can be created in a few steps with Webometric Analyst Wizard. First, download the program from the web site lexiurl.wlv.ac.uk. Second, create a plain text file with one domain name per line. The easiest way to create a plain text file is to use Windows Notepad (normally accessed via Start|All Programs|Accessories|Notepad). Each line should contain just the domain name of the web site and not any additional file or path information and there should not be any extra spaces in the file or blank lines. For example the file might contain the following two lines.

```
cybermetrics.wlv.ac.uk
lexiurl.wlv.ac.uk
```

Once the file is created, start Webometric Analyst and select the Network Diagram option from the first Wizard screen and click OK.

The second wizard screen requests the plain text file with the domain names: click OK and select the file in the dialog box that appears. Once the file is selected, Webometric

Analyst will start gathering data and when this is finished will display a network diagram, but this may take several minutes. The circles in the network diagram will have areas proportional to the number of pages in the web site that they represent and the arrows in the network diagram will have widths proportional to the number of links that they represent. All data on the diagram is derived from Bing searches.

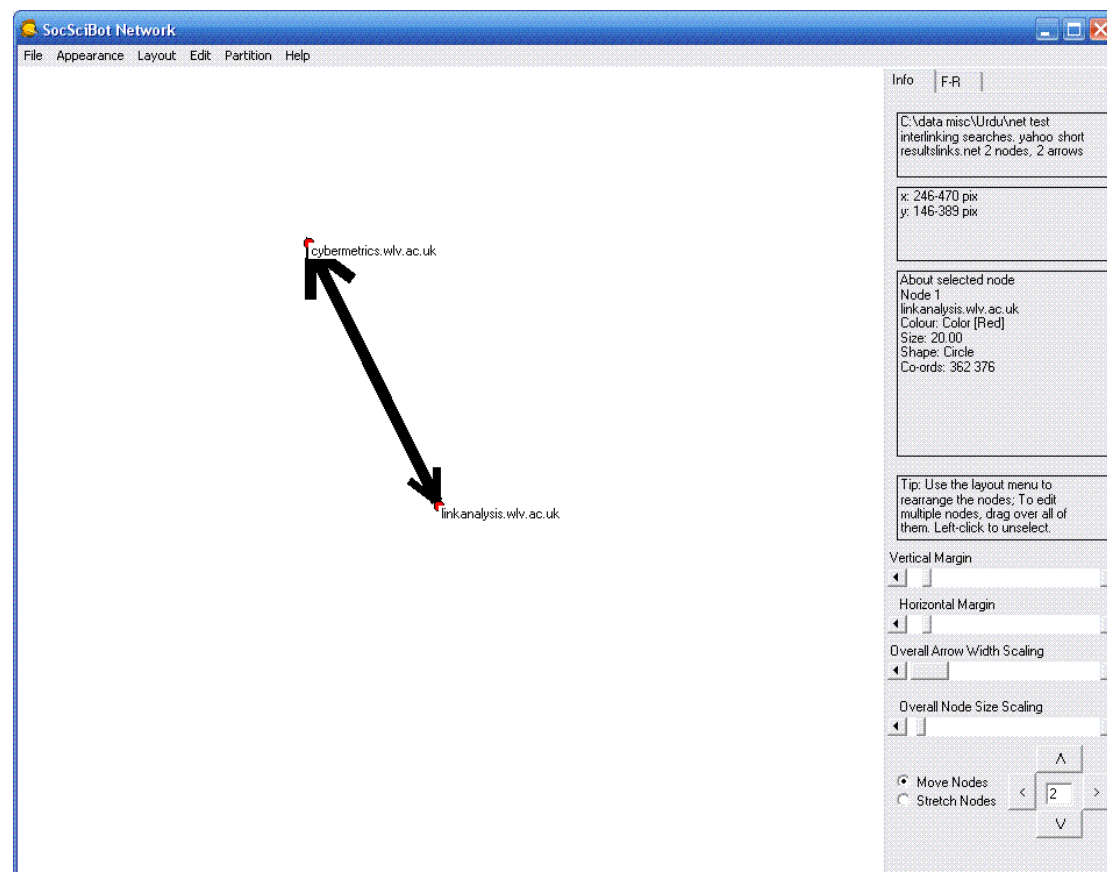


Figure 9.6. A simple network diagram created by Webometric Analyst.

Different network diagram variations can be created than the standard one. To create a co-inlink diagram instead of a direct link diagram, follow the same instructions as above but in the first step check the “Show Advanced Options” box and then an additional Wizard dialog box will eventually appear. This will give the option to run co-inlink searches (using URL citations) instead of direct link searches. Check this option and wait for the co-inlink diagram to be created.

Rearranging, saving and printing network diagrams

The network diagrams produced in Webometric Analyst (and SocSciBot, which uses the same graphing tool) can be rearranged, reformatted, saved and printed. This section briefly describes the most common of these actions.

Once the network has been displayed, it is important to arrange it so that it is readable and the patterns are easily visible to a human reader. To achieve this, the circles should be moved around so that they do not overlap or intersect with lines and so that the lines themselves intersect as little as possible. It is also helpful to position circles close together if they interlink and further apart if they do not, as far as possible. This rearrangement can be achieved manually or with the help of the Fruchterman-Reingold positioning algorithm. There are also many options to alter the appearance of the diagram in the main menus and right-click menus. For example, the right-click menus contain options for changing the colour or the border colour of the circles in the diagram. Most of the right-click formatting options are

mostly applicable by selecting one or more web sites by dragging the mouse across them and then using the right click button to access a menu of options.

The network diagram and positions can be saved from in SocSciBot Network by Selecting Save As from the File menu and saving as a SocSciBot Network file type. Alternatively, the network can be saved as a Pajek network, although in this format less information may be saved.

A network can be printed using the Print option in the File menu. There are various ways in which a network can be included another document, such as a Word file. The options below list various ways, in increasing order of image quality.

1. Paint Bitmap: press the Print Scr button on the keyboard, load Microsoft Paint (Start|All Programs|Accessories|Paint) and press Control-V. This should copy the screen into the Paint program, where it can be edited down to the correct size. Once edited, the file can be saved (as a bitmap .bmp for the highest resolution, or as a GIF .gif for the smallest file size) and then incorporated into a document (e.g., in Word, using Insert|Image|From File).
2. Medium resolution TIF. Within SocSciBot Network, select File|Print and then choose the Microsoft Office Document Image Writer printer driver from the Printer dialog box (in the Printer Name section). Use this to print a .TIFF version of the network, which can be inserted into a document, as for option 1.
3. High resolution TIF. A high resolution TIF printer driver is normally needed to get a higher resolution file than 300dpi. Once this is installed (it will probably have to be bought) follow the instructions for 2 above, except selecting the new printer driver.

Network diagram – classic interface example

A few less common types of link and co-link diagram can be created but for this the classic interface will be needed rather than the Wizards. The instructions below explain how to create a standard network diagram with the Classic Interface and this method can be customised for alternative network diagrams.

This section illustrates the steps needed to create a network diagram via the classic interface with a small example. The same technique is possible for any collection of web sites as long as they all have their own unique domain name, there are not too many sites in the list and there are some links between the sites.

1. Create a plain text file using Windows Notepad or a similar text editor containing a list of the domain names of each web site, one per line. For example, the file might be called smalllist.txt and contain the following list of domains.


```
www.oxford.ac.uk
www.harvard.edu
www.wlv.ac.uk
```
2. Use Webometric Analyst's ability to generate a list of searches between all pairs of sites by selecting *Make title link, URL citation link, co-title and co-URL citation searches for a set of URLs [with titles] (i.e., direct links and site co-inlinks)* from the *Make Searches* menu and selecting the list of domains just created (e.g., smalllist.txt). This will create a new text file (e.g., called smalllist directURLciteSch.txt) containing the necessary searches. The searches can be seen by opening the file.
3. Click on the *Run All Searches in a File* button and select the file with searches in (e.g., called smalllist.searches.txt) and wait for the searches to finish. This may take several minutes but for a large network it could take several hours.
4. Once the searches are finished, the hit count estimates in the short results file can be used as the raw data for a network diagram. To convert the short results into a format that can be read by the Pajek or SocSciBot Network programs, select *Convert link or colink short results file to Pajek Matrix* from the *Utilities* menu and select the short results file (e.g., smalllist.searches. bing short results.txt). The new file created will be in the necessary format (e.g., smalllist.searches.

Bing short results.net). This can be loaded into Pajek, if installed on the computer, or can be viewed in SocSciBot Network.

5. To use SocSciBot Network to display a network diagram, select SocSciBot Network from the File menu and the visualisation screen will appear. From the new SocSciBot Network screen, select Open from the File menu and select the network file (e.g., smalllist.searches.shortresults.matrix.net). The network will then be displayed on screen in a random format.

Co-link network diagrams

As described in the co-link section of the link analysis chapter above, co-link diagrams are often more revealing than link diagrams because they present an external perspective on a collection of web sites and can reveal structure even if the web sites in question do not interlink significantly. If a URL citation co-link network diagram is needed instead of a link network diagram then, as described above, the Webometric Analyst wizard can be followed as above for link networks but with the modification of checking the *Show advanced options* box and selecting co-inlink networks instead of link networks. If following the non-wizard steps then in Step 3, select the new file of co-inlink searches (containing coURLciteSch.txt, ignoring the other new files) for the searches. Note that the lines drawn in a network diagram for co-links should not have arrows on because co-links do not have a direction.

Webometric Analyst additional features

Webometric Analyst contains numerous utilities and options that can be used for non-standard or enhanced analyses. These typically involve a combination of menu item functions. Some of the key functions are described below but more will be added over time.

1. *Creating network diagrams interactively.* Network diagrams can be built site-by-site or page-by-page in SocSciBot Network, using its inbuilt menu and right-click menu options to download pages and add to the diagram the pages that they link to. For example, start SocSciBot Network from Webometric Analyst's *Show SocSciBot Network* menu option in the *File* menu and start by deleting the default diagram by selecting *Delete the Whole Network* from the *Edit* menu. Now add a first page by selecting *Add New Node* from the *Edit* menu and entering the domain name cybermetrics.wlv.ac.uk as the node name. To add web pages linked to this one, select the new circle by dragging the mouse over it, and select *Crawl (recursively download) from selected node(s) and add all links to diagram* from the right-click menu. Enter the crawl depth 1 (just follow links from the page), leave the search text blank and wait for it to crawl and update the diagram.
2. *Gaining over 1,000 results per search.* If there are more than 1,000 results for a search then this is a problem due to the search engine 1,000 result URLs maximum. In the professional version of Webometric Analyst it is possible to gain extra matches using the "query splitting" technique. This option is not available in the free version because it generates many additional searches, so makes extra demands on the search engine used. Please email the author if you are a researcher and need this extra capability.

10. (S2) Web Crawling: SocSciBot

This chapter introduces the web crawler software that can download and process a set of web sites in order to extract key summary statistics or to visualise any interlinking between the sites. Web crawlers are an alternative to automatic search engine searches that are appropriate when conducting an in-depth analysis of a set of web sites or when more reliable data is needed than provided by search engines. Web crawler data is more reliable than that from search engines in the sense that the scope of the crawls is determined by the researcher, whereas the extent of coverage of any particular web site by a commercial search engine is unknown. Moreover, the data from webometric crawlers like SocSciBot 4 (available free at socscibot.wlv.ac.uk) is designed to be as accurate as possible for webometric purposes, whereas search engine results are optimised for fast and effective information retrieval and are not optimised for accuracy.

Web crawlers

A web crawler is a program that can be fed with a single URL and then can download the web page, identify the hyperlinks within that web page and add them to its list of URLs to visit. The web crawler can then repeat the above process for each new URL in this list and keep going until it runs out of new URLs or reaches some other pre-defined limit (e.g., crawling a maximum of 15,000 URLs). Small-scale personal web crawlers typically visit one site at a time, starting at the home page URL and identifying and downloading web pages within the same site. Once finished, the crawler will usually have identified and downloaded all pages within a web site that can be found by following links from the home page. It is important to note that the crawler can only find new pages by following links and so it may miss pages that are not linked to. In a small, well-designed web site a crawler should be able to find all pages, however.

Figure 10.1 illustrates the “findability” issue for web crawlers. In this diagram, circles represent pages in a web site and arrows represent hyperlinks between them. A web crawler starting at the home page A will be able to follow its links to find pages B and E. It will then be able to follow the link on page B to find page C but it will not be able to find pages D and F because none of the pages found link to them.

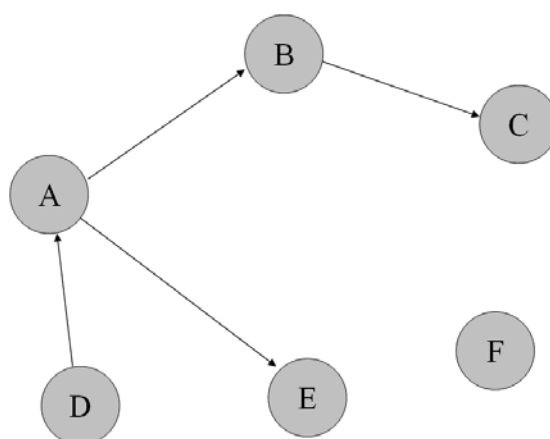


Figure 10.1. A simple web site link structure with circles representing pages and arrows representing hyperlinks between them. Pages D and F are not findable by crawling from A.

Web crawling can be used for relational link mapping. Suppose that a network diagram or other analysis of the links between a set of web sites is needed. The necessary data could be gathered by any web crawler if it was fed with the home page of each of the sites and tasked with crawling the sites as fully as possible. A webometrics web crawler like SocSciBot (described below) can perform the downloading and also extract data about the links between the web sites crawled.

In the sections above on link relationship mapping, the use of advanced searches in the search engine Bing or Google was described as a method for finding links. A web crawler would, in theory, return the same results but in practice web crawler data are more reliable for two reasons. First, the researcher can be sure that the web crawler has visited all sites in the study whereas a search engine may not have visited them all since search engines only cover a fraction of the web. Second, search engines tend not to report all results that they know about, hiding duplicate and near-duplicate pages. In consequence the figures returned by a web crawler will be a more accurate reflection of the number of links that it has found. There are some advantages in using search engines instead of web crawlers, however, particularly for collections of very large web sites. Commercial search engines may cover more pages in large web sites than individual web crawlers because large web sites may become fragmented over time and hence no longer fully crawlable by a web crawler. A search engine may be able to find the fragments by knowing the URLs of some pages in them from previous visits to the site when it was better connected (Thelwall, 2001). A second advantage is that crawling large web sites can take a long time and some web sites are too large to be practical to crawl with webometric web crawlers (e.g., www.microsoft.com, any large U.S. university web site).

Overview of SocSciBot

SocSciBot is a web crawler designed for webometrics research. It has been used to collect and analyse data on sets of web sites for at least fifty research articles. It has three main features: a crawler, a link analyser, and a text analyser. In version three, these three features operated as separate programs but they are combined in version 4. SocSciBot must be used in two separate phases for any research project: crawling and analysis. The two phases should not run concurrently because data from the crawling can interfere with the results of the analysis. The following illustrates the three stages for any SocSciBot investigation.

- *Create a new project and give it a name.* SocSciBot can crawl multiple web sites and analyse them together but the web sites must be collected into the same “project” in order for this to be possible. Hence, the first step with using SocSciBot, once it has been downloaded, is to create and name an empty project to contain all of the crawls. New projects can be created by entering a name in the SocSciBot Wizard Step 1 that appears when SocSciBot starts.
- *Crawl all the web sites to be analysed together.* Once the project has been created, it can be populated by crawling the web sites to be analysed within the project. New crawls can be added to a project by first selecting the project by name in SocSciBot Wizard Step 1 and then following the instructions in Wizard Step 2 to register the web site to be crawled.
- *Analyse the crawled web sites.* Once the crawls of all the web sites are complete, the downloaded data can be analysed through the link analyser, the text analyser or both. The links from web sites in a project can be analysed by first selecting the project by name in SocSciBot Wizard Step 1 and then choosing the link analysis option in Wizard Step 2 to access a standard set of reports on the links.

Network diagrams of sets of web sites with SocSciBot

This section describes how to use SocSciBot to crawl a set of web sites and then to construct a network diagram of the links between them. It uses SocSciBot 4.1, which is available free online from socscibot.wlv.ac.uk. The screenshots shown below are likely to change slightly in later versions of the program.

Step 1. Initiate a crawl of the web sites that are to be analysed. This can be done in two ways. The easiest way is to run a simultaneous crawl of all sites. To achieve this, start by creating a plain text file containing a list of URLs of the home pages of the sites to be crawled, one per line. For instance the file might be created with Windows Notepad (normally accessed via Start/Programs/Accessories/Notepad), called `startlist.txt` and contain the following text.

```
http://linkanalysis.wlv.ac.uk
```

http://cybermetrics.wlv.ac.uk
 http://socscibot.wlv.ac.uk

Note that the file should not be created with a word processor because such programs add unwanted data to the files they create.

Step 2. Start SocSciBot and initiate the web crawls. Before starting SocSciBot, it is a good idea to download and install the free network analysis program Pajek, as SocSciBot looks for this only the first time it starts. Once SocSciBot starts for the first time it asks for a location in which to save the data. Make sure that this is a place where you have permission to write or save data. This should preferably be in a hard disk on your own computer rather than a network drive as the latter can cause problems. SocSciBot then asks for a new project name (or to select an existing project, but the latter option can be ignored for now). Enter a project name such as “First test project” and press the “Start new project” button, as shown in Figure 10.2.

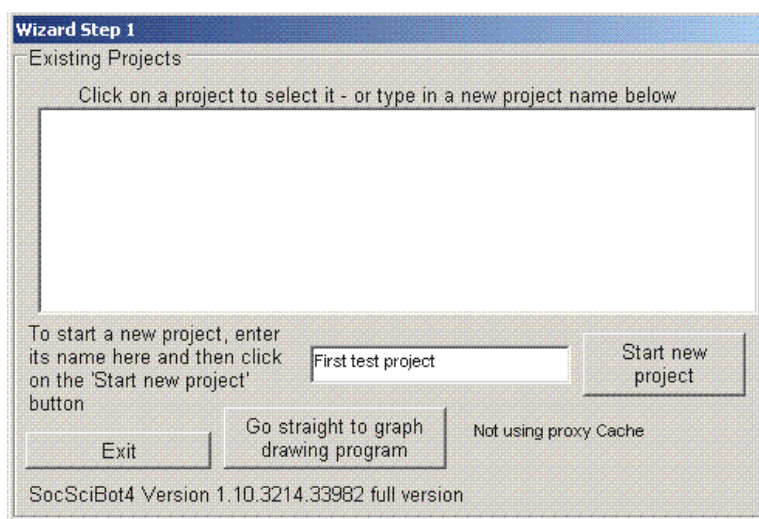


Figure 10.2. SocSciBot Wizard Step 1 showing a new project called “First test project” about to be created.

Step 3. Select a multiple site crawl or to enter the URL of a single web site home page to crawl that site. Select the “Download multiple sites” option, as shown in Figure 10.3, (the other option is discussed below) and click the *Crawl Site with SocSciBot* button.

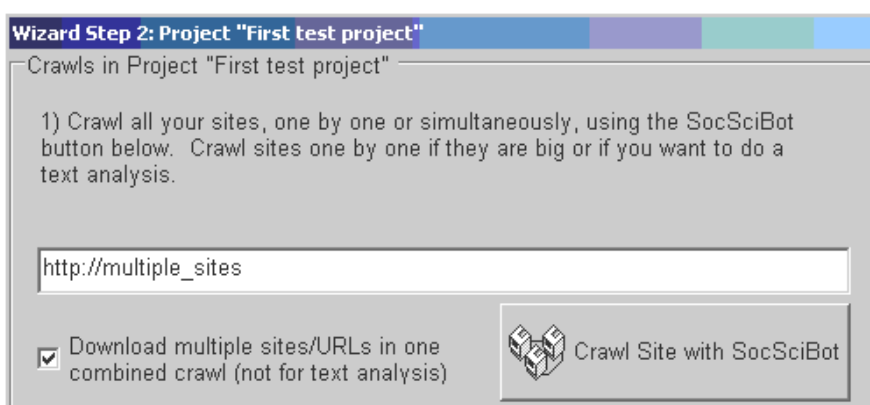


Figure 10.3. Registering to crawl multiple sites in a combined crawl.

Step 4. Load the list of home pages of sites to be crawled. To achieve this, first select the option: “I have a list of home pages of web sites that are not too big and I want all the web sites crawled completely in one go” option and then click on the “List of URLs to crawl” button and select the file startlist.txt created in step 2. The home pages and web site domain names should now appear in a list box at the bottom of the screen (see Figure 10.4). The

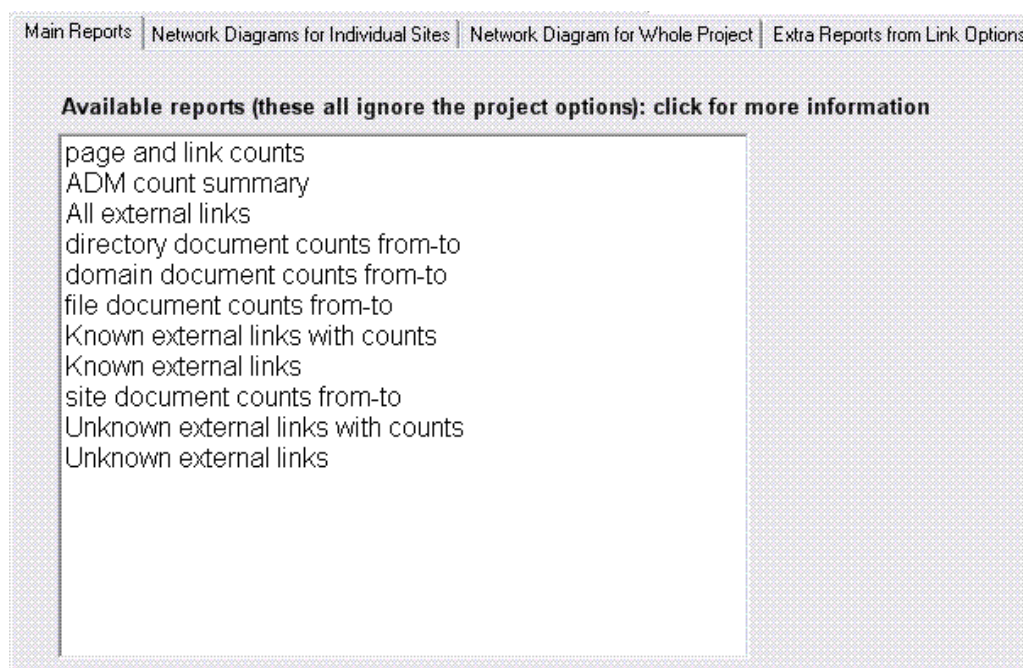


Figure 10.5. One of the lists of SocSciBot link analysis reports.

Step 7. Producing a network diagram. For this analysis, the focus is on producing a network diagram. Select the *Network Diagram for Whole Project* tab and click on the *Re/Calculate Network* button. Once this calculation is complete, the network data is ready to be drawn by SocSciBot or loaded into a network analysis program like Pajek. To view the network in SocSciBot, click the (rather odd) report name “single.combined.full” to load the data into SocSciBot network. This network is initially displayed at random but the domains can be moved to improve the readability of the diagram (e.g., placing interlinking sites close together and minimising the crossing of lines). Alternatively, the Fruchterman-Reingold algorithm can be used to automatically arrange the sites, but the options panel settings for this algorithm (the *F-R* tab on the right of the screen) may need to be used in order to produce a reasonable result. See the section above, *Rearranging, saving and printing network diagrams*, for more information.

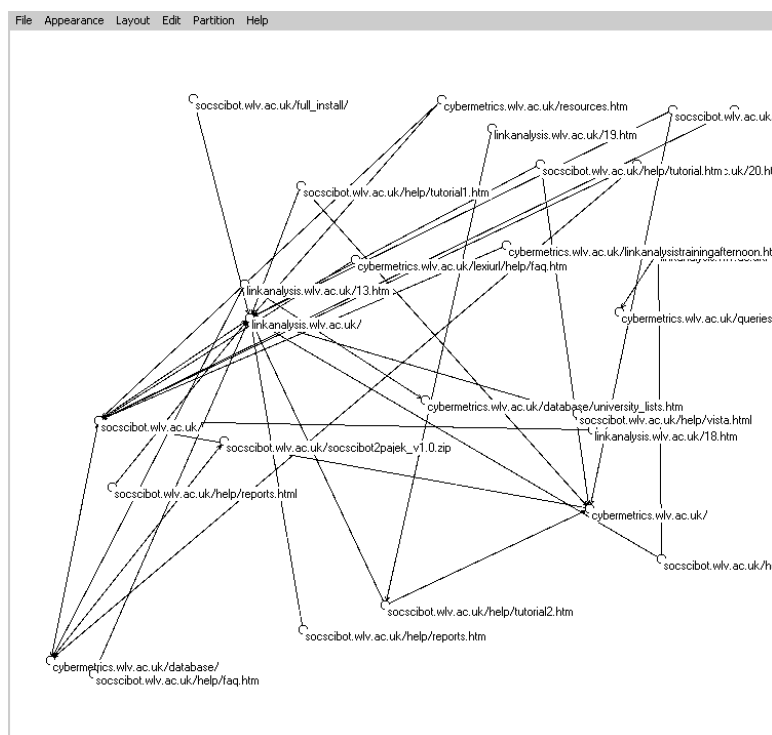


Figure 10.6. Interlinking pages in a SocSciBot Network diagram.

Step 8. Producing a site interlinking network diagram. To show links between sites rather than links between pages, several steps must be taken. In the SocSciBot Tools Main Reports screen, choose *Select types of links to include in reports* from the *Link Type Options* menu. Select the Domain aggregation level (rather than the default page aggregation level) and click OK. Now select again the *Network Diagram for Whole Project* tab and click on the *Re/Calculate Network* button. This will generate new data for a link diagram where the circles are domains rather than pages. To see this diagram, click on the report name “single.combined.full” again. As shown in Figure 10.7, all three domains contain pages that link to the other two domains.

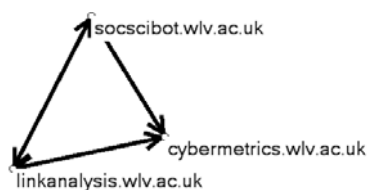


Figure 10.7. Interlinking domains in a SocSciBot Network diagram.

Note that network diagrams can be printed directly from the File/print menu item but if a very high resolution version (e.g., 600dpi) is needed for an academic publication then a high resolution document-producing printer driver must be installed first and then the diagram can be printed in high resolution via the print menu if the new printer driver is used.

If anything goes wrong with the above operations then please see the SocSciBot web site for help or post a report on the socscribot4.blogspot.com blog. If step 3 fails, however, because there are too many web sites or they must be crawled separately for some reason then start a new project and do not select the multiple crawls option but enter instead the homepage URL of the first site to be crawled. Select “crawl” to go to the main crawl screen, then select “Crawl Site” to start the crawl and then wait to be finished. For each site to be crawled follow the same process, but make sure that all crawls are allocated to the same project. It is normally possible to run four crawls simultaneously on the same computer. Once all the crawls have finished, step seven above can be followed.

Other uses for web crawls

The data gathered by SocSciBot can also be used for other purposes than producing a network diagram. Below are some examples of uses:

- The impact of each web site created by links within the other web sites crawled can be seen in the link counts report. This is accessible via the Main Reports tab in the SocSciBot Tools Main Reports screen and the information is in the ADM count summary report. This report contains the number of links to and from each site to and from each other site, as counted by page, directory, domain name or web site.
- Network statistics are also available from the Main Reports tab in the SocSciBot Tools Main Reports screen. These include the total number of pages, directories and domains linking between each pair of sites (the *from-to* reports). If more social network analysis statistics are needed then the data matrix used to create the networks can be copied into a specialist program such as UCINet or extracted from Pajek.
- Network diagrams of individual sites can be produced but first the counting level must be changed to “pages” or directories rather than sites or domains in the example above (again using the Link Type Options menu), and the site self-links option should be checked in the same menu option box. Click on the *Network diagrams for individual sites* tab then click on the *Re/Calculate Network* button. Clicking on the domain name of a site will then produce a network diagram of the pages in the site in SocSciBot Network.

11. (A1) Search Engines and Data Reliability

Commercial search engines are important for webometrics because they are used to supply the raw data for many studies, for example via Webometric Analyst. In addition search engines are frequently investigated for the variability and coverage of their results because they are so widely employed by web users to find information (Bar-Ilan, 2004). This chapter gives an overview of how search engines work and summarises some webometric research about search engine reliability and coverage. The chapter is more theoretical and less directly practical than the preceding chapters but gives important background information to help understand and interpret the results of webometric techniques using commercial search engines.

Search engine architecture

The overall design or architecture of a search engine incorporates several components with completely different tasks. In essence there are three main parts: the web crawler, the indexed page database and the web page ranking engine (for a more complete characterisation, see: Brin & Page, 1998; Chakrabarti, 2003).

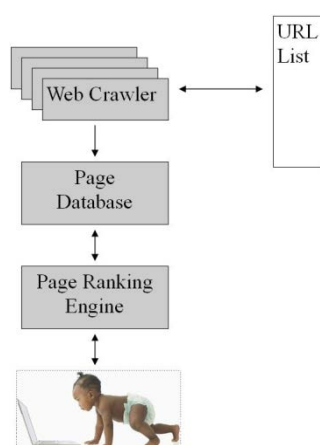


Figure 7.1. The three components of a commercial search engine that co-ordinate to get results to a web user.

The web crawler has the task of finding web pages to populate its search engine's huge database. It does this by building a list of URLs of all known web pages and periodically visiting these pages to check for and download updated versions. When a page is downloaded, all hyperlinks within the page are extracted and added to the URL list, if they are not already there (Figure 7.1). In theory the URL list may have begun life as a single URL or as a short list of URLs but in any major search engine it will contain billions of URLs.

The crawler's URL list does not contain the URL of every existing web page. This is because URLs are not added to it automatically when new web pages are created but only when the URL is submitted to the search engine (as occasionally happens) or when a link to the page is found by the crawler in a downloaded page. In consequence, entire web sites may not be found by a search engine because they are not linked to by other sites. A study in 1999 estimated that the search engines of the day covered up to 17% of the web (Lawrence & Giles, 1999), a surprisingly low figure. The problem is exacerbated by pages that are password-protected, banned from crawling by their owners, or in a format that search engines cannot fully read or interpret (Java, Flash): the so-called "invisible web" (Sherman & Price, 2001). A consequence of this partial coverage for webometric studies is that *search engine results should never be interpreted as an accurate reflection of the web itself*.

The index of a search engine has the task of matching user queries to web pages in the search engine's database. It is able to use standard information retrieval techniques to do this efficiently, retrieving matching pages within tenths of a second from the billions in the

database. Part of a search engine's algorithm may take short cuts to return fast results, such as retrieving only 10% of the matching pages before creating the first page of results. In such a case the algorithm may have to estimate the total number of matching pages in the remainder of the database in order to produce the hit count estimates displayed on the results pages. An additional complicating factor is that there may be multiple databases, perhaps with the most important or most frequently updated pages having their own areas. This can make the reported hit count estimates unreliable.

The page ranking engine is its most secret part of a commercial search engine. It uses information about the degree to which the matching pages appear to focus on the topic of the query, the importance or authority of the matching pages and information about the user (e.g., their geographic location and search history) to rank the pages. The rank order is presumably designed to give the maximum chance that the user finds relevant information on the first results page so that they do not get frustrated and want to switch to a different search engine. Although the exact features used for ranking are not known for any major search engine, hyperlinks seem to be particularly useful as evidence of the importance of the target page. Google's key PageRank algorithm is based upon the assumption that the most useful or important pages will tend to be linked to often, and that links from high quality pages are particularly good indicators of the value of target pages (Brin & Page, 1998). Newer algorithms like TrustRank can work in parallel with PageRank and seek to identify and reject pages that are likely to be spam because they are linked to by spam pages (Gyongyi, Garcia-Molina, & Pedersen, 2004).

Duplicate and near-duplicate elimination

Duplicate and near-duplicate elimination is a factor that is important for webometrics in terms of understanding both the hit count estimates returned by search engines and the lists of URLs returned as matching a search.

In conjunction with ranking pages, search engines attempt to ensure that they do not return many similar or identical pages in their results (Gomes & Smith, 2003). This is because a page is unlikely to be relevant to a user if they have seen a similar page before and not visited it or have visited it but not found it useful. This duplicate and near-duplicate elimination seems to be done "just in time" – after ranking the results or as part of ranking the results - and can sometimes result in a drastic reduction in the total number of results delivered to the searcher. This is an important cause of unreliability in hit count estimates. One study of this phenomenon for Live Search (Bing) suggested that large hit count estimates on the first results page (e.g., over 8,000) would tend to reflect the number of URLs in the index matching the search, with little duplicate elimination. In contrast, small initial hit count estimates (e.g., under 300) would tend to reflect the number of remaining URLs after all duplicate and near-duplicate elimination (Thelwall, 2008). In-between figures and estimates on later results pages could reflect either or a combination of the two. This complicates the interpretation of hit count estimates in webometrics, especially when comparing large and small values.

Search engine duplicate elimination sometimes excludes multiple pages from the same site after the first two pages. It also seems to work on the basis of page titles and the snippets about the pages displayed in the search engine results: if these are similar for two pages then one of them might be eliminated. Since the snippets tend to be selected by the search engine to contain the search terms, this means that pages that are "locally similar" around the search terms may be regarded as essentially duplicates and held back from the user. Finally, since the duplicate elimination appears to occur in stages rather than all at once, it is possible to find near-duplicate results or multiple pages from the same site displayed on different results pages.

As a result of all the factors discussed above, the typical list of URLs returned by a search engine is likely to be incomplete with an unknown number of missing URLs. In addition, search engines return a maximum of 1,000 URLs per search. If a URL sample is taken from the search engine results, such as the first 10 matches, first 100 matches or all

1,000 matches returned (if there are more than 1,000) then it is important to recognise that this is likely to be a biased sample because of the ranking mechanism. The results returned are probably the most authoritative pages, coming from the most important or authoritative sites, or the pages that match the searches best. Hence, *if URLs are to be processed to extract statistics then it is best to use as large a list as possible to minimise bias.*

Comparing different search engines

The results returned by search engines for many searches can be very different from each other. These differences may be due to different ranking procedures, different coverage of the web by the crawlers or different duplicate elimination processes (Thelwall, 2008). Differing search engine results can be exploited for webometrics in two ways. First, statistics generated from results can be compared between search engines. This comparison can give an indication of how reliable the figures are. For example, if one search engine returned 90% of URLs from the .com domain but another returned 90% from the .edu domain then this would cast doubt on both statistics. The second way of taking advantage of differences is to use multiple search engines is to combine or average their results to minimise the idiosyncrasies of each one.

Research into search engine results

This section briefly describes a range of webometrics research into the performance of search engines in terms of the information returned. This introduces an important type of research and gives further background information about the reliability of search engine results.

Early research into search engines compared the hit count estimates over time and found them to vary significantly from day to day for many searches (Bar-Ilan, 1999). For example the hit count estimates for the same search in a single search engine could fluctuate daily, with periodic step changes in results. Rousseau (1999) proposed a solution to this problem in the form of treating the hit count estimates as a time series and using an average (or a moving average) for webometric investigations. Subsequent research found that the major search engines had become far more stable with their hit count estimates so that there did not seem to be a need for averaging any more (Thelwall, 2001). Nevertheless, search engines did seem to periodically make substantial changes to their databases, such as increasing capacity, and so occasional variations still occurred.

A similar type of study investigated changes in hit count estimates between the different pages of results for the same search. It was found that the major three search engines at the time (Google, Yahoo and Live Search/Bing) all reported changed hit count estimates on separate results pages occasionally, although these estimates did not vary frequently between pages. For example, a search might have a hit count estimate of 5,000 for the first 7 results pages, but 3,000 for the remaining 13 pages. About half of the sets of results pages had at least one change and this change was almost always a downward revision. Overall these findings suggested that duplicate elimination was the process driving the changes and that this elimination was conducted periodically when delivering results but not for every page (Thelwall, 2008).

Bar-Ilan has conducted a series of studies monitoring changes in the information reported by search engines over time. Picking the information science topic of informetrics (the study of information measurements), she ran searches in a range of search engines for this topic and compared the results from year to year. Although she found increasingly more results each year, she also discovered pages missing from the results: either pages apparently never found by a search engine's crawlers or pages previously known to a search engine but apparently subsequently forgotten (Bar-Ilan, 1999; Bar-Ilan & Peritz, 2004; Mettrop & Nieuwenhuysen, 2001). These forgotten pages could have been present in the search engine database but not returned to the searcher as a result of duplication, or they could have been discarded from the database for other reasons. Interestingly, Bar-Ilan was able to show that search engines sometimes did not return a page that matched the queries used even when the page was in the search engine's database and contained relevant information that was not in

any other returned page. This latter point suggests that the near-duplicate elimination process is imperfect in terms of the contents of pages.

Vaughan, Cho and others have researched the extent to which search engine results contain demonstrable biases. From the findings, it seems that search engines have clear but probably unintentional bias in favour of countries that were early adopters of the web but not any language biases (Vaughan & Thelwall, 2004). The bias seems to be due to the link crawling process used by web crawlers to build their URL list of the web: newer web sites are less likely to be found because they are less likely to have been linked to by other web sites already known to the crawler. This is exacerbated by ranking algorithms, which have an age bias: search engines tend to rank older pages more highly than newer pages (Cho & Roy, 2004).

Modelling the web's link structure

Some research in computer science and statistical physics has shed light on the link structure of the web, as covered by search engines. This gives additional background information about the web that is useful to interpret the coverage of web crawlers and search engine ranking.

Research undertaken by the search engine AltaVista (which no longer exists) at the turn of the century analysed interlinking between web pages within a single web crawl and found that they could be split into five different chunks, as illustrated in Figure 7.2 (Broder et al., 2000). The core of the crawl was the “Strongly Connected Component” (SCC). This is a huge collection of 28% of the pages that is particularly easily crawlable in the sense that a crawl starting at any SCC page would eventually reach the whole of the SCC. Hence it seems reasonable to assume that all major search engines would have found virtually all of the SCC. OUT (21%) consists of all pages not in the SCC that are reachable by following links from the SCC. OUT is also very crawlable: a crawl starting in any SCC page would reach all OUT pages but the reverse is not true: OUT pages do not form a good starting point for a crawl but should also be included almost entirely in any major search engine.

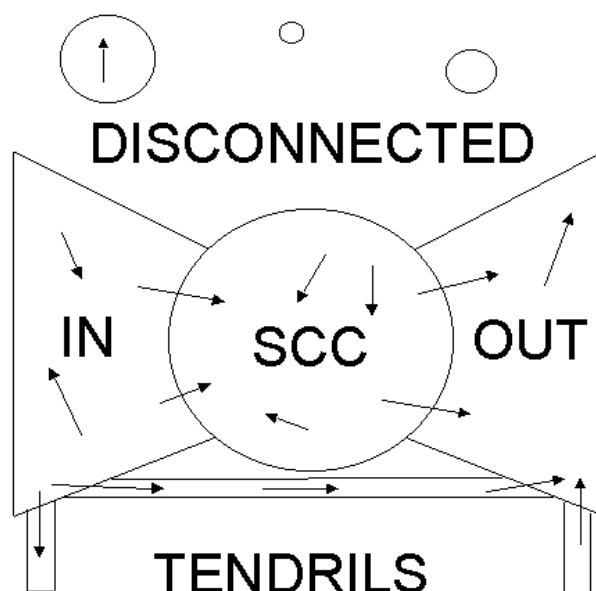


Figure 7.2. The link structure of a major search engine web crawl (Broder et al., 2000).

IN pages (21%) are those that are not in the SCC but from which the SCC can be reached by following links. IN pages are opposite to OUT: they do form a good starting point for a crawl because a crawl starting in IN will capture the whole SCC and out, as well as part of IN. No

crawl starting point guarantees capturing much of IN, however, and so IN is likely to be only partly covered by other major search engines. The TENDRILS collection of pages (22%) is connected to IN, OUT or both in more exotic ways and the DISCONNECTED pages (8%) are not connected in any way to the other components. Like IN these last two components are unlikely to be well covered by other search engines.

The above discussion relates to the “topology” of the web. It reveals how there is likely to be a large core of the web that is well covered by all search engines and that the rest of the web is likely to be covered unevenly.

A second type of research has focused on identifying mathematical laws in link count data. This has helped to popularise terms like “long tail” and “power law”. The main finding is that link counts are extremely unevenly distributed between pages. Whilst most web pages attract only a few links or no links at all, a small number attract millions. In fact the number of links to web pages approximately follows a mathematical relationship known as a power law. Figure 7.3 gives an example of its shape for links to UK university web sites from other UK university web sites. Note that the axes of this graph are logarithmic, which hides the extent to which pages that have attracted only a few links are enormously more common than pages that have attracted many links. For example, pages with 10 inlinks are about 500 times common than pages with 100 inlinks.

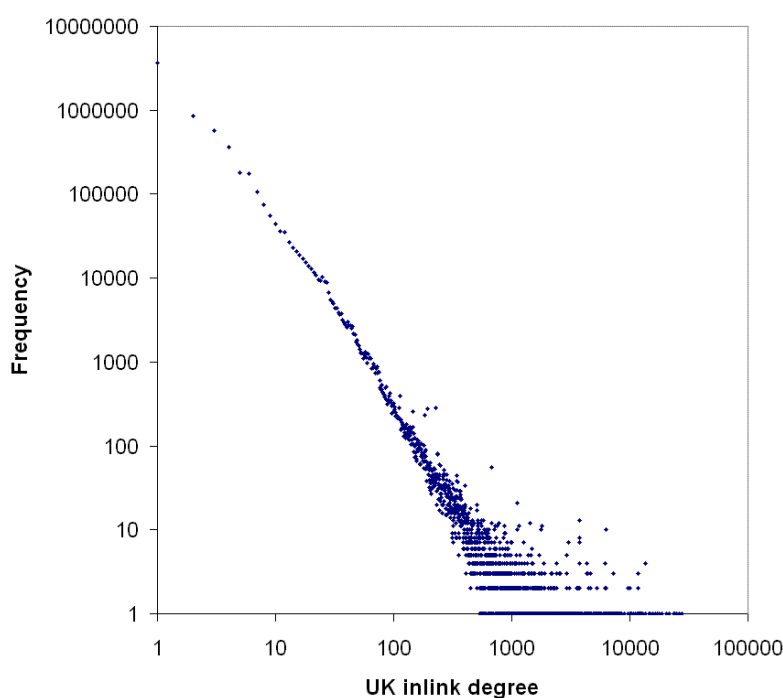



Figure 7.3. The power-law distribution of links to UK academic web sites (Thelwall & Wilkinson, 2003).

Power laws tend to arise when a “rich get richer” phenomenon is at work. This suggests that web pages tend to attract new links partly because they already have some links (Barabási & Albert, 1999). In fact, the more inlinks that a page has, the more new inlinks it is likely to attract. This can be explained by search engine ranking. Pages that have already attracted many inlinks are likely to be near the top of search engine results and so are easily found. Other factors being equal, these will be the best known pages and the most likely to attract any new links created. As mentioned above, a consequence of this is that it is difficult for new sites to get noticed because they start with no links and hence are at a disadvantage compared to established sites that have already attracted many links.

Although the power law rule holds approximately for the web in general, it does not necessarily apply to specific collections of single types of pages. For instance it is a poor fit for university home pages (Pennock, Flake, Lawrence, Glover, & Giles, 2002). For small

groups of pages of a similar type, it may be that organisational factors such as size and industry sector are more important than the rich-get-richer phenomenon.

A statistical property of power laws relevant to webometrics is that their arithmetic means are not useful because of the skewed distribution. The median is a better measure of central tendency because it effectively ignores the huge values.



12. (A2) Tracking User Actions Online

This chapter gives a brief introduction to methods for tracking user actions online that do not rely upon data collected from search engines or web crawlers. This is webometrics in the sense of analysing web data quantitatively but some might not regard it as webometrics because the raw data used is not derived from web pages but from other sources of information about how web pages are accessed. For more information on issues in this section, see a review article about web metrics (Heinonen & Hägert, 2004), the computer science field of web usage mining (Srivastava, Cooley, Deshpande, & Tan, 2000), and special information science user-tracking topics such as deep log analysis (Huntington, Nicholas, & Jamali, 2007) and log or download analysis for scientific impact measurement (Brody, Harnad, & Carr, 2006; Moed, 2005).

Single site Web analytics and Web server log file analysis

The most common method of evaluating the impact of a web site or analysing the activities of visitors to a single web site is web log file analysis. Web server log files typically contain detailed information about the page accesses of a web site's visitors. These files are normally created by the web server program that operates the site but can also be created by third party programs like Google analytics if the web site owner places appropriate code in each page. In either case the raw data files typically contain information revealing the pages each user visited, which web site directed them to the site, the approximate length of time spent on the site and the approximate geographic location of the visitors. Web analytics software, sometimes called web server log file analysis software, can process this information to report useful summary statistics, such as the daily, weekly or monthly total number of unique visitors, most common web sources of new visitors to the site and most popular pages within the site. This is useful to get an overall impression of how the site is being used and how this usage is changing over time. For example, the Google Analytics program surprisingly revealed that the most popular page within Link Analysis web site linkanalysis.wlv.ac.uk was the page supporting chapter 23 (see Figure 12.1), and delving deeper into the report revealed that the most common source of visitors for that page was a Wikipedia page about networks.

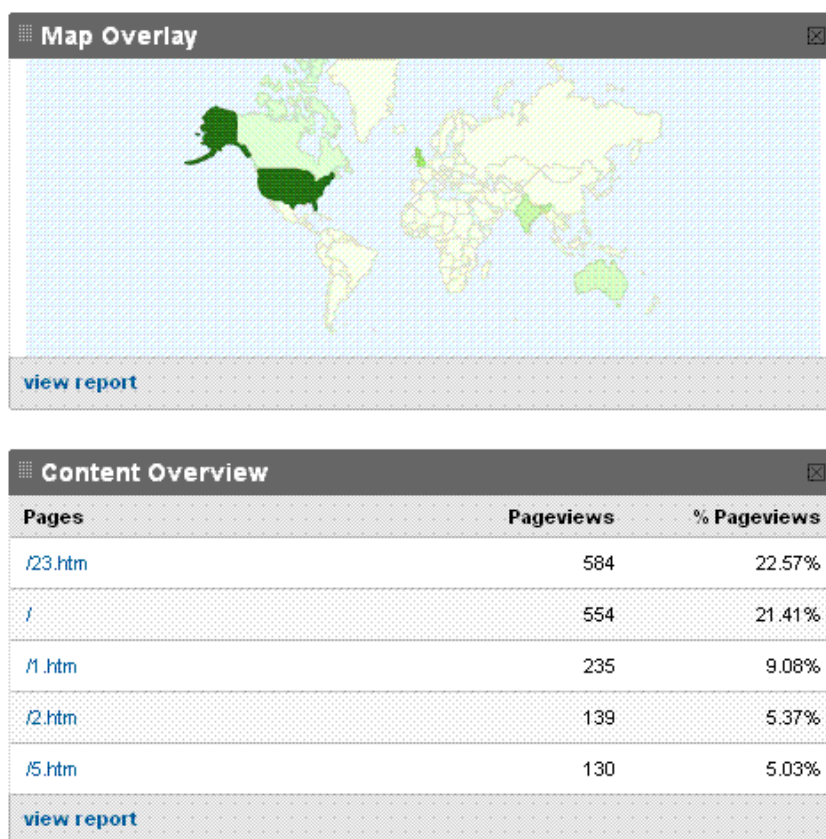


Figure 12.1. Detail from the Google Analytics report for linkanalysis.wlv.ac.uk showing the geographic origins of visitors (top) and the most commonly visited page (bottom).

The knowledge of which web site a user visited previously gives useful information about which sites link to the site analysed and how much traffic these source sites send. If the source site is a search engine like Google then the previous page will be a search results page and the URL of this results page encodes within it the original search terms. As a result, web analytics software is able to report the search terms that led users to a site, which gives useful insights into *why* they visited it.

Much web analytics information is more detailed and more robust than the equivalent information from web impact reports. For example the most popular pages in a web site could be estimated in a web impact report by finding which pages were most commonly targeted by links from other web sites whereas the web log files would contain accurate visitor statistics. Nevertheless, web analytics are typically unsuitable for a web impact evaluation because they can only be applied with the permission of web site owners so cannot be used to demonstrate impact relative to similar or competitor sites. If possible, both should be used together for any site as the information that they give is complementary even though it partly overlaps.

Multiple site Web analytics

A number of companies track the activities of large numbers of web users, either through a program installed on user computers or through access to anonymised activity logs through arrangements with individual Internet Service Providers. Three currently prominent such companies are HitWise, Comscore and Alexa. These give out some information free in the form of press releases or web reports but earn revenue by selling more comprehensive data. Alexa regularly produces lists of the most visited web sites in the world and in individual countries. These are useful sources of data to track and confirm the importance of rapidly emerging web sites.

A book written by HitWise gives many examples of the kinds of insights that this kind of data can give into online human behaviour (Trancer, 2008). One of the more

impressive conclusions was that increased activity in social network sites coincided with reduced visits to pornographic sites. It seems that significant numbers of web users were using social network sites instead of pornography or during times previously occupied by pornography use. A similar analysis showed that generic gambling like poker was connected to major sporting events because a proportion of gamblers would use online gaming sites when sports betting was unavailable.

Web usage data would be ideal for online analytics reports because it contains the same kind of information about web sites as server log files but would include data about competitors' web sites. Unfortunately, however, it is likely to be expensive to purchase and thus impractical for most purposes. For relatively popular sites, it might be possible to use Alexa's free webmaster web traffic statistics (see www.alexa.com and sign up), which would be a valuable additional statistic to report.

Search engine log file analysis

The log files of commercial search engines contain a record of all user searches and can be used by the search engines to find out things like the most popular search terms, which links tend to be clicked in response to given keyword searches and how many results pages tend to be viewed before users give up. This information is normally private although some has been released to selected researchers and search engines sometimes make public some summary data, such as the most popular search terms or the fastest increasing search terms (e.g., www.google.com/trends/hottrends). In addition, users can produce time series trend graphs of reasonably popular search terms through Google trends (www.google.com/trends).

The results of search engine log file analysis have been written up extensively elsewhere, most notably in a book (Spink & Jansen, 2004) but some key results can be summarised here (based on data from the three search engines Excite, AltaVista, and AllTheWeb). Searchers tend to submit simple searches with an average of about 2.5 keywords per query, although this average seems to be slowly increasing and searchers reasonably often construct some kind of complex query (about a quarter of the time), such as a phrase search or one involving Boolean operators. Very approximately three quarters of searches stop at the first results page, showing the importance of the top search results.

13. (A3) Advanced Techniques

This chapter contains a selection of advanced techniques in webometrics. Some are extensions or applications of the topics in previous chapters and the remainder are new techniques.

Query splitting

The major search engines return a maximum of 1,000 results for any query. This means that if you submit a search and keep clicking on the “Next Page” link to see more results, then the results will stop when about 1,000 have been displayed. This limit is a problem for webometric techniques like web impact reports and link impact reports, which rely upon lists of URLs and may be compromised if significantly incomplete lists are returned.

Query splitting is a technique to get additional URLs matching a search beyond the maximum of 1,000 normally returned (Thelwall, 2008). The basic idea is very simple: to submit a set of modified versions of the original query and then combine the results of each modified version. This works as long as any result that matches a modified query also matches the original one. This matching can be ensured by always starting with the original query and modifying it by adding extra terms. For example, suppose that the query *jaguar* returns the maximum 1,000 URLs but Google reports that there are about 1 million matching pages. To get additional matches, new queries could be made like *jaguar animal* or *jaguar car*. Any result that matched either of these would contain the word *jaguar* and would therefore match the original search. The chances are that some of the 1,000 URLs matching each of the two searches above would not be the same as the results returned for the original search and so combining the results of all three is likely to give many more than 1,000 URLs, and up to a maximum of 3,000 URLs, all of which match the original search.

Whilst the above method could be applied with a human selecting the terms to add, it is also possible to automate it using automatically submitted queries. For example, the query splitting technique selects a term that occurs in about 10% of the titles and snippets of the first 1,000 results and makes two queries: one adding this term to the original query and the other subtracting it. Suppose that the 10% word chosen for the query *jaguar* was *cat*. Then the two new queries would be *jaguar cat* and *jaguar -cat*. An advantage of this process is that the two new queries cannot have any URLs in common and so if both return 1,000 results then this would give 2,000 different URLs in total. This query splitting can be repeated again on the two new searches to give up to 4,000 new URLs. For instance, suppose that the 10% word for *jaguar cat* was *tiger*. Then the two new queries would be *jaguar cat tiger* and *jaguar cat -tiger*. Similarly, for the query *jaguar-cat* suppose that the 10% word was *engine*. Then the two new queries would be *jaguar -cat engine* and *jaguar -cat -engine*. Figure 13.1 gives an additional example.

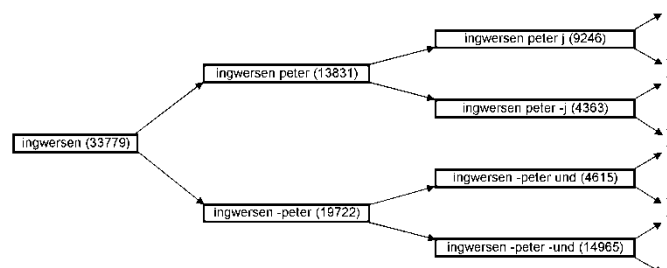


Figure 13.1. An illustration of the start of query splitting for the query “Ingwersen”. The number of matches for each query is in brackets (Thelwall, 2008).

In theory the query splitting process could be repeated until all searches returned less than 1,000 matches but in practice search engines have limits on the complexity of searches processed (e.g., a maximum of 10 terms or 150 characters per search) and so this is not always possible. If it is possible then the final set of queries should in theory give all results known to the search engine but this is not likely to be the case because of issues related to duplicate elimination in the results (see *Duplicate and near-duplicate elimination* within section 7.1). In consequence the final results set should be made by combining all results, including those from the original query and from all intermediate steps. Even this combined list is unlikely to be exhaustive, however.

Virtual memetics

Virtual memetics is a set of methods to track online the evolution over time of a single meme, i.e., a single transmitted unit of cultural information. The method was developed to track the international online spread and morphing of a joke but is applicable to any text-based information that is widely circulated via the web (Shifman & Thelwall, 2009). It is based on analysing a sample of relevant web pages and using date-specific searches to track evolution over time.

The first step of virtual memetics is common to many webometrics methods: constructing an appropriate search engine query to match pages containing the meme (e.g., a joke) and building a list of URLs or matching pages either via an automatic query submission program like Webometric Analyst or using manual searches. Query splitting (see the section on this) may be needed if there are more results than the maximum 1,000 returned by a search engine.

The second step is to download a random sample of a manageable collection of web pages from the list (e.g., 300) with a maximum of one page per site.

The third step is to automatically or manually analyse the sample for variations in the meme. The idea behind this is to capture key variations in the meme made by individuals as it circulates. Whilst automatic methods might include clustering by text similarity, the manual method would essentially be a human clustering by similarity. Ultimately, however, unless the statistical clustering results are particularly clear-cut, it will be a human decision to distinguish between relatively minor meme variations and more substantial variations that deserve to be named and analysed as a whole. For example, in the initial study using this technique, there were joke variations labelled sexist, feminist and Indian, amongst others.

The fourth step is to track the evolution over time of the variations by identifying evidence about their earliest appearance online. This cannot be perfectly achieved since email cannot be tracked over time and web pages disappear but there are three different methods that can jointly be used to give an indication of when a meme or meme variation first occurred: Google blog search results can be sorted by date so one useful statistic that can easily be obtained is the earliest date in which the meme variation appeared in a blog post indexed by Google blog search (blogsearch.google.com). Similarly, a Google Groups search (groups.google.com) can be used to identify the earliest occurrence of each variation in a Google group. There are many other similar search services but these often have a time limitation of 6 or 12 months and so they may not be useful. A final method is to use date-specific searches (available in the advanced search section of some search engines or as a filtering option after running a search). This allows searches for web pages indexed by the search engine that are older than a given date and have not been modified since that date. Putting together these three methods, the estimated creation date would be the earliest of the three dates. This cross-checking does not give a guarantee that the date is correct but seems like a reasonable method to produce an approximation. The end result of the steps so far would produce a timeline of the major meme variations and their estimated dates of emergence. Figure 13.2 gives an example of a timeline obtained from this method.

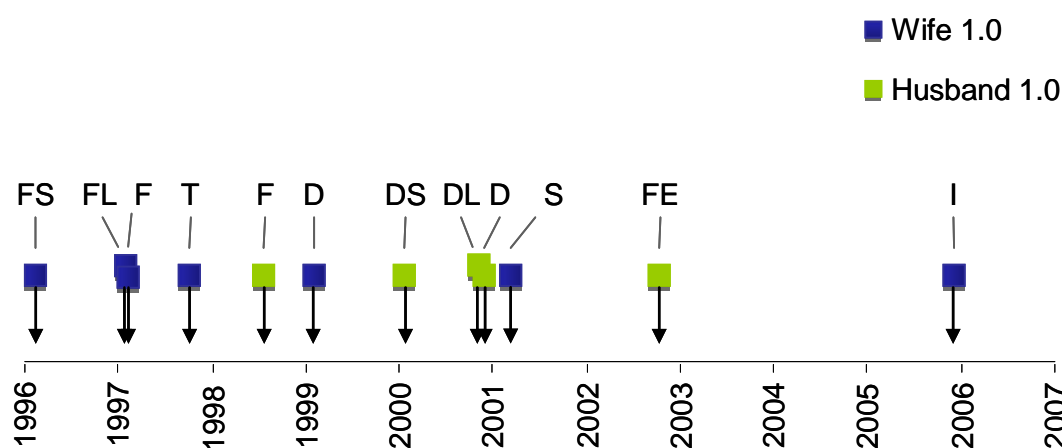


Figure 13.2. A timeline of the earliest appearances of major variations of two memes online extracted using the virtual memetics method. Each code represents a different major variation (e.g., FE is the feminist variant) (Shifman & Thelwall, 2009).

A fifth and optional step is to identify and track translations of the meme and meme variations. This can be achieved using Steps 1-4 above on translations of the jokes. In addition, for extra coverage of non-English languages it may be useful to also search for additional matches with language-specific or regional search engines or interfaces for the major search engines. Finally, comparing the variations of the meme in its original language with its translated international variations may reveal key international similarities or variations.

Web Issue Analysis

Web issue analysis (Thelwall, Vann, & Fairclough, 2006) is the tracking of the spread of an issue online. For example, one study attempted to identify all online mentions of the United Nations Integrated Water Resources Management concept in order to discover how international it had become. The techniques of issue analysis start with those for a web impact assessment, but also include some additional linguistic techniques. Hence, a web issue analysis is based on one or more web searches for the issue, with matching URLs or web sites counted and a content analysis conducted to validate and help interpret the results, and a TLD breakdown used to suggest the international spread of the topic.

The additional linguistic method that makes web issue analysis different from web impact assessment is the identification of words that occur relatively often in web pages mentioning the issue. This can be useful to identify concepts that may have been missed by the content analysis. Word counting for this purpose is impractically time-consuming and so it can only be done with the support of a text analysis computer program (e.g., the text analysis component of SocSciBot). In addition, a control group collection of texts should be similarly analysed so that a comparison of the two results can identify words that are relatively more common in the issue-relevant pages. This then produces two lists of words and word frequencies: one for issue-mentioning pages and one for the control group. These two can be compared by identifying relatively highly ranked words in the issue-specific list or by using formal statistical methods to identify unusually high frequency words in the issue-specific list (McEnergy & Wilson, 2001; Thelwall et al., 2006). An alternative method is to use natural language processing techniques to extract nouns and noun phrases from the matching web pages, reporting the most common of these. Figure 13.3 illustrates the results of this for a web issue analysis.

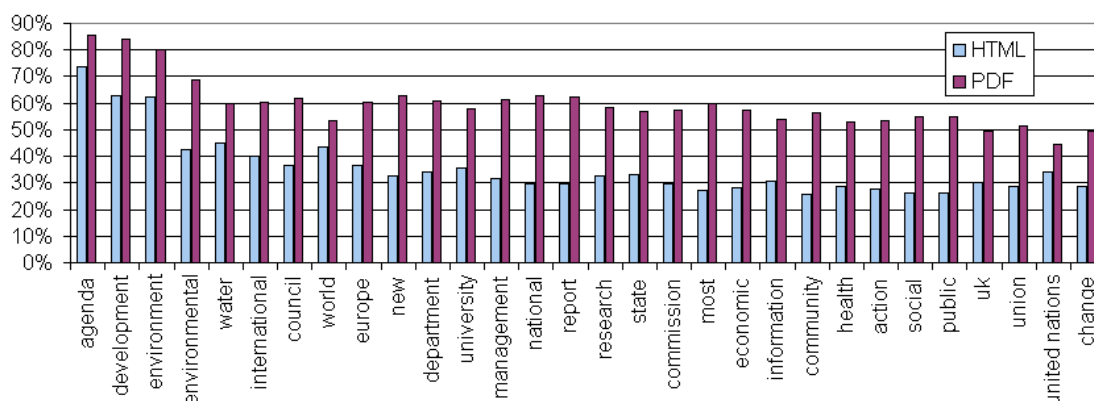


Figure 13.3. Most common nouns in Integrated Water Resources Management web pages, as measured by the percentage of domains (Thelwall et al., 2006).

Data mining social network sites

Some quantitative studies of the web have focussed on specific web sites or types of web site, using specially designed programs to extract detailed information from each page, beyond just the links extracted by web crawlers. This section describes attempts to mine public information from the formerly popular social network site MySpace, using modified web crawlers.

MySpace became apparently the most popular social network site in the world in 2006 but subsequently was overtaken by Facebook in mid-2008. It is particularly suitable for detailed webometric analysis because the profile pages of members over 16 are by default public, although users can choose to make them private. It is possible to randomly sample MySpace members because each one is given a unique ID number, these numbers are given out consecutively, and the numbers can be converted easily into profile page URLs. In fact, because the IDs are given out consecutively, it is even possible to use a random number generator to sample people who joined on any given date. Several studies have taken advantage of the public nature of many MySpace profiles to automatically extract detailed information from them (Hinduja & Patchin, 2008; Liu, 2007).

One early study focused on the proximity of online MySpace Friends by taking a person's Friends and attempting to extract detailed information about their location from the relevant profile section. Initial results suggested that most Friends tended to live close together, for example in the same town (Escher, 2007). Another study took random members and paired them with random MySpace Friends from their Friend lists, and compared their reported attributes and beliefs. The results showed that similar people tended to be Friends disproportionately often, including for similarity in terms of age, religion, country of residence and ethnicity. In contrast, and despite most offline studies of friendship, members were not disproportionately often Friends with others of the same gender (Thelwall, 2009). This study only considered active Friendship in the sense of commenting on Friends' profiles. A related study investigated swearing in the comments written on Friends' profiles, finding, again in contrast to previous offline studies, that strong swearing was equally prevalent in the profiles of male and female UK members. In the US, however, strong swearing was more common in male profiles. In both cases swearing occurred in the majority of profiles and was particularly prevalent in younger members' profiles. Table 13.1 gives an example of the many linguistic types of swearing found. This is interesting because informal friendly swearing is rarely written down and hence MySpace gives an almost unique opportunity to study swearing on a large scale (Thelwall, 2008).

Table 13.1. A classification by profile owner of UK MySpace swearing types from all the UK MySpace users, (427 swear words, 264 male and 163 female profiles; all except 67 were friends' comments) (Thelwall, 2008).

Linguistic type	% of all swearing		Examples (slashes // separate comments)
	M	F	
Predicative negative adjective	0%	1%	..and your myspace page is fucked
Cursing expletive	1%	3%	yeah but. IVE JUST GOT BACK FROM A DEFTONES GIG! so fuck you jim bob! X // so id bollocks to this
Destinational usage	3%	1%	fuck....right....off.... // Chris you're slacking again !!! Get the fuck off myspace lol !! you good anyway ?
Emphatic adverb/adjective OR Adverbial booster OR Premodifying intensifying negative adjective	32%	38%	and we r guna go to town again n make a ryt fuckin nyt of it again lol // see look i'm fucking commenting u back // lol and stop fucking tickleing me!! // Thanks for the party last night it was fucking good and you are great hosts. // sick of being hurt all the bastard time.
General expletive	6%	7%	fuck i didnt know this was u lol // lol....noooooo waaaayyyyyy....fuck !!! // Ahh Fuck it, we're doomed anyway..
Idiomatic set phrase OR Figurative extension of literal meaning	28%	17%	think am gonna get him an album or summet fuck nows // qu'est ce que fuck? // what the fuck pubehead whos pete and why is this necicery mate // Heh long story.. cant be fucked to explain :D
Literal usage denoting taboo referent	3%	3%	Oh n shaggin dead people // ...that does not mean I am the village yokal idiot or that daddy fucked me with a rusty broken pitch fork...
Pronominal form with undefined referent	0%	1%	I Don't Care Who You Think You Are, Where You're From Or How Many Of You Think It'd Be Fun To Start Some Shit With Them, I Will Fuck You Up. Simple As That. // Occupation: No.1 Cunt Kicker-inner.
Personal: Personal insult referring to defined entity	27%	28%	tehe i am sorry.. i m such a sleep deprived twat alot of the time! lol // Maxy is the soundest cunt in the world!!!! // 3rd? i thought i was your main man number one? Fucker // write bak cunt xxx
All	100%		

The techniques discussed above are different from most webometrics research in that they rely upon data mining within the text of the web pages – i.e., extracting specific facts from the pages. It requires specially written programs, which is another difference from most webometrics research. There are many other examples of projects that also extract information from web pages to address research questions. For example, one attempted to automatically detect web page genres from their contents (Rehm, 2002).

Social network analysis and small worlds

The webometric techniques that generate networks lend themselves to analysis by any one of the many academic network analysis techniques, whether from maths (e.g., graph theory), physics (e.g., complex network theory) or the social sciences (e.g., social network analysis). This section discusses the last of these, as most relevant to typical social sciences research.

Social network analysis (Wasserman & Faust, 1994) is a research field concerned with the analysis of networks of individuals or organisations. Its principal belief is that some social systems cannot be properly understood when analysed at the level of individual components but are better analysed as networks of interacting parts. A classic example is communication within a primitive village environment: to find out how information travels it

is necessary to know about each person's friends and acquaintances and their talking patterns. Social network analysis has developed a range of metrics for assessing various aspects of a network. Some of these metrics are centred on individual network nodes: for example the degree centrality of each person in a village might be defined to be the number of friends that they have. Other metrics can be based more upon the structure of a network as a whole. For example, the betweenness centrality of a node is the probability that a random path between two random nodes will pass through the given node. This is high for nodes that play an important bridging function – they may not have many connections but they are important for communication.

Social network analysis can be applied to collections of web pages or web sites with the nodes in the social network being the web pages or web sites and the connection between them being the hyperlinks. The metrics of social network analysis are potentially useful to describe network properties and to help compare networks but caution should be exercised when applying them to web networks because the assumptions made when defining them – typically that the connections between nodes are the key channels of communication – do not necessarily apply to the web. For example, even though hyperlinks can be used to travel between web sites, it is probably more common to switch between sites via a bookmark or a search engine search. Blogspace and discussion forums could well be exceptions to this, however.

One interesting application of social network analysis techniques on the web is Lennart Björneborn's analysis of cross-topic connectors in the UK academic web space (Björneborn, 2004, 2006). He found that although most links between academic web sites in different universities connected similar research areas, some were inter-disciplinary connectors. Computer science web sites were particularly prominent in forming such cross-topic connections.

Folksonomy tagging

The tags generated by users of Web 2.0 content-based systems like YouTube, Flickr and delicious have generated much research interest because they are a new and relatively unfettered method of describing resources that may help information retrieval (Golder & Huberman, 2006). Webometrics research can contribute to the understanding of this informal "folksonomy" tagging by identifying common tags and types of tags. For example, one study extracted tags from millions of YouTube, Flickr and delicious pages in order to find the most frequent tags in each system (Ding, Toma, Kang, Fried, & Yan, 2008). The differences in the results were evidence of significant variations in the way that users employed the tagging facilities.

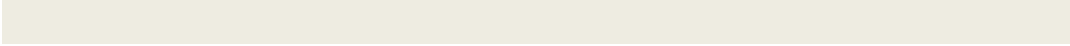
API programming and mashups

As discussed in other chapters, Applications Programming Interfaces (APIs) are pathways provided by search engines and some other software to give programmers simple access to useful facilities, subject to certain restrictions such as daily usage limits (Mayr & Tosques, 2005). By early 2013, APIs of potential use in webometrics were provided only by Bing for general web searches, but also by Flickr, Google Books and YouTube, amongst others, for specialist searches. Given access to programming skills or a program like Webometric Analyst that can tap into the relevant API, this gives webometrics researchers potential access to large amounts of data.

APIs are typically configured to allow a program to submit a specific request for information in a standard format. For example the Bing web search API allows searches to be submitted together with a range of parameters, such as the number of results to return (up to 50). This request must be encoded and submitted as a URL in a format specified in the Bing API instructions. The information returned by the Bing API in response to this request is structured in an XML document format described in the API documentation. This document can be easily processed by a computer program to extract the desired information.

Most webometrics research using search engines has used APIs to allow a large amount of data to be analysed without requiring extensive human labour. An example of webometrics API-based research that is different to the topics discussed in this book is a study of Flickr that used the Flickr API to randomly select pictures from a Flickr Group for a subsequent human content analysis (Angus, Thelwall, & Stuart, 2008).

Although currently rare in webometrics research, API functionality can be extended through combining different sources of data to produce a *mashup*. This is a program that operates online via multiple data sources or APIs. An example of mashup research is Tobias Escher's combination of MySpace data with Google Maps to chart the positions of all of a MySpace member's friends (Escher, 2007), as discussed above.



14. Summary and Future Directions [modified]

This book introduces a set of big data webometric methods for the social sciences and information science in particular. *Blog searching* can identify fluctuations in public interest in a topic as well as revealing the causes of these fluctuations. This can be supported by Google search volume data to check that any trends found are not peculiar to blogs. *Link impact analysis* can be used to indicate the impact of web sites or pages and *web impact assessment* can be used to indicate the online impact of sets of documents or terms. In addition, *hyperlink-based network diagrams* can be used to identify patterns of interlinking between sets of web sites. All of these techniques are relatively simple to apply and can be used as one of a number of methods to investigate a research problem, or can be applied in a more in-depth manner as a separate research method. The techniques can also be applied to investigate online phenomena or the online reflections of offline phenomena. All of the methods have a number of limitations, however, that mean that their results should be interpreted carefully and cautiously in many cases.

In addition to discussing the above techniques, this book also discusses tools to carry out the studies (blog search engines, SocSciBot and Webometric Analyst) and gives practical instructions for using them. In addition, a range of supporting background information is given, such as on the workings of search engines, to help interpret the research. An important topic for webometrics is the reliability of search engines and this is also a separate area of study within webometrics. Finally, the book also gives a brief introduction into a range of advanced webometric topics that have more specific purposes than the main techniques discussed.

The future of social science big data research seems to be in finding new applications of the established techniques and to develop new methods in response to the emergence of new information sources. This is partly reliant upon the provision of free search interfaces (e.g., Alexa traffic ranking, the Topsy search engine) by companies that have the resources to collect data on a large scale. Such interfaces help to make the techniques easily accessible to the wider research community. Moreover, webometrics research is likely to continue to develop parallel data collection and analysis software, such as SocSciBot, that are not dependent upon free data sources.

In terms of research topics, tagging has been an important new research area (i.e., the study of the keywords assigned to online resources by users), although one that is of primary relevance to information science and computer science rather than the wider social sciences. It may be also possible to extend current social network research but this is dependent upon the privacy models of sites being sufficiently open to allow webometrics research.

An important indication that social science big data research and webometrics are likely to continue to expand in the future is the apparently irrevocable establishment of user-generated content as an important part of the web. It seems inevitable that some of the future big online applications or movements will include the generation of data in a form that could be counted and analysed to generate public opinion or other trends. This will allow social scientists to continue to gain new insights into both online and offline phenomena.

15. Glossary

Co-link Either a co-inlink or a co-outlink. The term co-link is sometimes used when it is clear from the context whether it is a co-inlink or a co-outlink.

Co-inlink, co-inlinked Two web sites that are both linked to by a third web site are called co-inlinked. In practice, co-inlinks are often derived from search engine data and in this case an additional restriction applies: two co-inlinked pages are always linked from a single co-inlinked page in a third web site.

Co-outlink, co-outlinked Two web sites that both link to a third web site are said to co-outlink. In practice, co-outlinks are often derived from search engine data and in this

case an additional restriction applies: two co-outlinking pages always link to a single page in a third web site.

Direct link A hyperlink from one web page to another.

Directory or URL directory Any URL ending in a slash (e.g., <http://www.wlv.ac.uk/admin/>).

Domain or domain name The part of a typical URL following the initial <http://> and preceding the next slash, if any (e.g., www.microsoft.com, www.wlv.ac.uk).

Domain inlink/outlink, see inlink/outlink.

Duplicate elimination and near-duplicate elimination In the context of search engines, this is the phenomenon that causes search engines to hold back some of the pages matching a user's query because pages previously returned appeared to be similar. This similarity appears to be based upon either the pages appearing in the same web site or the titles and snippets from the pages appearing in the results being quite similar.

Hit count estimate The figure reported on results pages by search engines as the total number of matching pages. For example, the 10,000 in "Results 1-10 of about 10,000" is a hit count estimate from a Google search.

Inlink A hyperlink to a web page from another web page, sometimes called a page inlink. A site inlink is a link to a page from a page in a different web site. Similarly a domain inlink is a link to a page from a page with a different domain name. The term inlink is sometimes used when it is clear from the context whether it is a page, domain or site inlink.

Interlinking Two web pages interlink if either one has a link to the other. Similarly two web sites or domains interlink if any page in one has a link to any page in another.

Outlink A hyperlink from a page to a different web page, sometimes called a page outlink. A site outlink is an outlink from a page to a page within a different web site. Similarly a domain outlink is a link from a page to a page with a different domain name. The term inlink is sometimes used when it is clear from the context whether it is a page, domain or site inlink.

Page inlink/outlink, see inlink/outlink.

Site inlink/outlink, see inlink/outlink.

STLD The last two segment of a domain name after the penultimate dot if a standard naming system is used (e.g., com.au, org.uk, ac.nz), otherwise the TLD of the domain name.

Title mention This is the inclusion of a title in a web page, with or without a hyperlink. For example, "I like the BBC News" is a title mention in this page for BBC News. Title mentions can be used as a replacement for hyperlinks in Webometric Analyst instead of URL citations.

TLD Top Level Domain. The last segment of a domain name after the final dot (e.g., com, org, uk).

URL citation This is the inclusion of an URL (or URL without the <http://>) in a web page, with or without a hyperlink. For example, "I like news.bbc.co.uk" is an URL citation in this page for the BBC news web site.

Web site A loose term that at its most general encompasses any collection of web pages that form a coherent whole in terms of content or owning organisation. In webometrics, the term is also used for all collections of web pages sharing the same domain name ending, where this ending includes one dot-separated segment immediately preceding the STLD (e.g., wlv.ac.uk, microsoft.com).

16. References

- Adamic, L., & Glance, N. (2005). The political blogosphere and the 2004 US election: Divided they blog. *WWW2005 blog workshop*, Retrieved May 5, 2006 from: <http://www.blogpulse.com/papers/2005/AdamicGlanceBlogWWW.pdf>.
- Aguillo, I. F., Granadino, B., Ortega, J. L., & Prieto, J. A. (2006). Scientific research activity and communication measured with cybermetrics indicators. *Journal of the American Society for Information Science and Technology*, 57(10), 1296-1302.

- Almind, T. C., & Ingwersen, P. (1997). Informetric analyses on the World Wide Web: Methodological approaches to 'Webometrics'. *Journal of Documentation*, 53(4), 404-426.
- Angus, E., Thelwall, M., & Stuart, D. (2008). General patterns of tag usage amongst university groups in Flickr. *Online Information Review*, 32(1), 89-101.
- Bar-Ilan, J. (1999). *Search engine results over time - a case study on search engine stability*. Retrieved January 26, 2006, from <http://www.cindoc.csic.es/cybermetrics/articles/v2i1p1.html>
- Bar-Ilan, J. (2004). The use of Web search engines in information science research. *Annual Review of Information Science and Technology*, 38, 231-288.
- Bar-Ilan, J., & Peritz, B. C. (2004). Evolution, continuity, and disappearance of documents on a specific topic on the Web: A longitudinal study of 'informetrics'. *Journal of the American Society for Information Science and Technology*, 55(11), 980 - 990.
- Barabási, A. L., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286(5439), 509-512.
- Barjak, F., & Thelwall, M. (2008). A statistical analysis of the web presences of European life sciences research teams. *Journal of the American Society for Information Science and Technology*, 59(4), 628-643.
- Björneborn, L. (2004). *Small-world link structures across an academic web space - a library and information science approach*. Royal School of Library and Information Science, Copenhagen, Denmark.
- Björneborn, L. (2006). 'Mini small worlds' of shortest link paths crossing domain boundaries in an academic Web space. *Scientometrics*, 68(3), 395-414.
- Björneborn, L., & Ingwersen, P. (2004). Toward a basic framework for webometrics. *Journal of the American Society for Information Science and Technology*, 55(14), 1216-1227.
- Borgman, C. L., & Furner, J. (2002). Scholarly communication and bibliometrics. *Annual Review of Information Science and Technology*, 36, 3-72.
- Börner, K., Sanyal, S., & Vespignani, A. (2007). Network science. *Annual Review of Information Science and Technology*, 41, 537-607.
- boyd, d., & Ellison, N. (2007). Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, 13(1), Retrieved December 10, 2007 from: <http://jcmc.indiana.edu/vol2013/issue2001/boyd.ellison.html>.
- Brin, S., & Page, L. (1998). The anatomy of a large scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1-7), 107-117.
- Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., et al. (2000). Graph structure in the web. *Journal of Computer Networks*, 33(1-6), 309-320.
- Brody, T., Harnad, S., & Carr, L. (2006). Earlier Web usage statistics as predictors of later citation impact. *Journal of the American Society for Information Science and Technology*, 57(8), 1060-1072.
- Chakrabarti, S. (2003). *Mining the Web: Analysis of hypertext and semi structured data*. New York: Morgan Kaufmann.
- Chakrabarti, S., Joshi, M. M., Punera, K., & Pennock, D. M. (2002). *The structure of broad topics on the Web*, from <http://www2002.org/CDROM/refereed/338>
- Chen, C. (2004). *Information visualization: Beyond the horizon, 2nd ed.* New York: Springer.
- Cho, J., & Roy, S. (2004). Impact of Web search engines on page popularity. *Proceedings of the World-Wide Web Conference, May 2004*, Retrieved February 4, 2007 from: <http://oak.cs.ucla.edu/~cho/papers/cho-bias.pdf>.
- Crane, D. (1972). *Invisible colleges: Diffusion of knowledge in scientific communities*. London: The University of Chicago Press.
- Cronin, B., Snyder, H. W., Rosenbaum, H., Martinson, A., & Callahan, E. (1998). Invoked on the web. *Journal of the American Society for Information Science*, 49(14), 1319-1328.
- Ding, Y., Toma, I., Kang, S. J., Fried, M., & Yan, Z. (2008). Data mediation and interoperation in social Web: Modeling, crawling and integrating social tagging data. *Workshop on Social Web Search and Mining (WWW2008)*, Retrieved October 20,

- 2008 from:
http://keg.cs.tsinghua.edu.cn/SWSM2008/short%20papers/swsm2008_submission_2005.pdf.
- Escher, T. (2007). The geography of (online) social networks. *Web 2.0, York University*, Retrieved September 18, 2007 from: http://people.oii.ox.ac.uk/escher/wp-content/uploads/2007/2009/Escher_York_presentation.pdf.
- Flake, G. W., Lawrence, S., Giles, C. L., & Coetzee, F. M. (2002). Self-organization and identification of Web communities. *IEEE Computer*, 35, 66-71.
- Foot, K., & Schneider, S. (2006). *Web campaigning*. Cambridge, MA: The MIT Press.
- Foot, K. A., & Schneider, S. M. (2002). Online action in campaign 2000: An exploratory analysis of the U.S. political web sphere. *Journal of Broadcasting and Electronic Media*, 46(2), 222-244.
- Foot, K. A., Schneider, S. M., Dougherty, M., Xenos, M., & Larsen, E. (2003). Analyzing linking practices: Candidate sites in the 2002 US electoral web sphere. *Journal of Computer Mediated Communication*, 8(4), <http://www.ascusc.org/jcmc/vol8/issue4/foot.html>.
- Fruchterman, T. M. J., & Reingold, E. M. (1991). Graph drawing by force-directed placement. *Software: Practice and Experience*, 21(11), 1129-1164.
- Garfield, E. (1970). Citation indexing for studying science. *Nature*, 227(669-671).
- Geisler, E. (2000). *The metrics of science and technology*. Westport, CT: Quorum Books.
- Godfrey, A., Thelwall, M., Enayat, M., & Power, G. (2008). Generating new media and new participation in Iran: The case of ZigZag, *International Association of Media and Communication Research*. Stockholm, Sweden.
- Golder, S. A., & Huberman, B. A. (2006). The structure of collaborative tagging systems. *Journal of Information Science*, 32(2), 198-208.
- Gomes, B., & Smith, B. T. (2003). Detecting query-specific duplicate documents. *United States Patent 6,615,209*, Available: <http://www.patents.com/Detecting-query-specific-duplicate-documents/US6615209/en-US/>.
- Gyongyi, Z., Garcia-Molina, H., & Pedersen, J. (2004). Combating web spam with TrustRank. *Proceedings of the thirtieth international conference on Very large data bases*, 30, 576-587.
- Heimeriks, G., Hörlesberger, M., & van den Besselaar, P. (2003). Mapping communication and collaboration in heterogeneous research networks. *Scientometrics*, 58(2), 391-413.
- Heinonen, J., & Hägert, M. (2004). Web metrics. Retrieved February 19 from: <http://web.abo.fi/~kaisa/HH.pdf>.
- Hinduja, S., & Patchin, J. W. (2008). Personal information of adolescents on the Internet: A quantitative content analysis of MySpace. *Journal of Adolescence*, 31(1), 125-146.
- Holmberg, K., & Thelwall, M. (2009). Local government web sites in Finland: A geographic and webometric analysis. *Scientometrics*, 79(1), 157-169.
- Huntington, P., Nicholas, D., & Jamali, H. R. (2007). Site navigation and its impact on content viewed by the virtual scholar: a deep log analysis. *Journal of Information Science*, 33(5), 598-610.
- Ingwersen, P. (1998). The calculation of Web Impact Factors. *Journal of Documentation*, 54(2), 236-243.
- Jascó, P. (2005). Google Scholar: the pros and the cons. *Online Information Review*, 29(2), 208-214.
- Kamada, T., & Kawai, S. (1989). An algorithm for drawing general undirected graphs. *Information Processing Letters*, 31(1), 7-15.
- Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5), 604-632.
- Kousha, K., & Thelwall, M. (2007). Google Scholar citations and Google Web/URL citations: A multi-discipline exploratory analysis. *Journal of the American Society for Information Science and Technology*, 58(7), 1055 -1065.

- Kousha, K., & Thelwall, M. (2008). Assessing the impact of disciplinary research on teaching: An automatic analysis of online syllabuses. *Journal of the American Society for Information Science and Technology*, 59(13), 2060-2069.
- Latour, B., & Woolgar, S. (1979). *Laboratory life: The social construction of scientific facts*. Los Angeles: Sage.
- Lawrence, S., & Giles, C. L. (1999). Accessibility of information on the web. *Nature*, 400(6740), 107-109.
- Lenhart, A., Arafeh, S., Smith, A., & Macgill, A. R. (2008). Writing, Technology and Teens (4/24/2008). *Pew Internet & American Life Project*, Retrieved May 1, 2008 from: http://www.pewinternet.org/PPF/r/2247/report_display.asp.
- Li, X., Thelwall, M., Wilkinson, D., & Musgrove, P. B. (2005). National and international university departmental web site interlinking, part 2: Link patterns. *Scientometrics*, 64(2), 187-208.
- Liu, H. (2007). Social network profiles as taste performances. *Journal of Computer-Mediated Communication*, 13(1), Retrieved June 5, 2008 from: <http://jcmc.indiana.edu/vol2013/issue2001/liu.html>.
- Mayr, P., & Tosques, F. (2005). Google Web APIs: An instrument for webometric analyses? Retrieved January 20, 2006 from: http://www.ib-berlin.de/%2007Emayr/arbeiten/ISSI2005_Mayr_Toques.pdf.
- Mayr, P., & Walter, A. K. (2007). An exploratory study of Google Scholar. *Online Information Review*, 31(6), 814-830.
- McCowan, F., & Nelson, M. L. (2007). Search engines and their public interfaces: Which APIs are the most synchronized? *WWW 2007*, Retrieved January 8, 2008 from: <http://www2007.org/htmlposters/poster2868/>.
- McEnery, T., & Wilson, A. (2001). *Corpus linguistics*. Edinburgh: Edinburgh University Press.
- Merton, R. K. (1973). *The sociology of science. Theoretical and empirical investigations*. Chicago: University of Chicago Press.
- Mettrop, W., & Nieuwenhuysen, P. (2001). Internet search engines - fluctuations in document accessibility. *Journal of Documentation*, 57(5), 623-651.
- Moed, H., F. (2005). *Citation analysis in research evaluation*. New York: Springer.
- Moed, H. F. (2005). Statistical relationships between downloads and citations at the level of individual documents within a single journal (p). *Journal of the American Society for Information Science & Technology*, 56(10), 1088-1097.
- Neuendorf, K. (2002). *The content analysis guidebook*. London: Sage.
- Newman, M. E. J. (2003). The structure and function of complex networks. *SIAM Review*, 45, 167-256.
- Nicolaisen, J. (2007). Citation analysis. *Annual Review of Information Science and Technology*, 41, 609-641.
- Ortega, J. L., Aguillo, I., Cothey, V., & Scharnhorst, A. (2008). Maps of the academic web in the European Higher Education Area: an exploration of visual web indicators. *Scientometrics*, 74(2), 295-308.
- Ortega, J. L., & Aguillo, I. F. (2008). Visualization of the Nordic academic web: Link analysis using social network tools. *Information Processing & Management*, 44(4), 1624-1633.
- Ortega, J. L., & Aguillo, I. F. (2009). Mapping world-class universities on the web. *Information Processing & Management*, 49(2), 272-279. doi:10.1016/j.ipm.2008.1010.1001.
- Park, H. W., & Thelwall, M. (2006). Web science communication in the age of globalization: Links among universities' websites in Asia and Europe. *New Media and Society*, 8(4), 631-652.
- Pennock, D., Flake, G. W., Lawrence, S., Glover, E. J., & Giles, C. L. (2002). Winners don't take all: Characterizing the competition for links on the web. *Proceedings of the National Academy of Sciences*, 99(8), 5207-5211.

- Rehm, G. (2002). *Towards automatic web genre identification*. Paper presented at the the 35th Hawaii International Conference on System Sciences.
- Rodríguez i Gairín, J. M. (1997). Valorando el impacto de la información en Internet: AltaVista, el "Citation Index" de la Red. *Revista Española de Documentación Científica*, 20(2), 175-181.
- Rogers, R. (2004). *Information politics on the Web*. Massachusetts: MIT Press.
- Rousseau, R. (1999). Daily time series of common single word searches in AltaVista and NorthernLight. *Cybermetrics*, 2/3, Retrieved July 25, 2006 from: <http://www.cindoc.csic.es/cybermetrics/articles/v2002i2001p2002.html>.
- Sherman, C., & Price, G. (2001). *The invisible web: Uncovering information sources search engines can't see*. Medford, NJ: Information Today.
- Shifman, L., & Thelwall, M. (2009). Globalization under the radar: The world-wide diffusion of one Internet joke. *Journal of the American Society for Information Science and Technology*, 60(12), 2567-2576.
- Smith, A. G. (1999). A tale of two web spaces; comparing sites using Web Impact Factors. *Journal of Documentation*, 55(5), 577-592.
- Spink, A., & Jansen, B. J. (2004). *Web search: Public searching of the web*. Dordrecht: Kluwer Academic Publishers.
- Srivastava, J., Cooley, R., Deshpande, M., & Tan, P.-N. (2000). Web Usage Mining: Discovery and applications of usage patterns from web data. *SIGKDD Explorations*, 1(2), 12-23.
- Stuart, D., & Thelwall, M. (2006). Investigating triple helix relationships using URL citations: A case study of the UK West Midlands automobile industry. *Research Evaluation*, 15(2), 97-106.
- Stuart, D., Thelwall, M., & Harries, G. (2007). UK academic web links and collaboration - an exploratory study. *Journal of Information Science*, 33(2), 231-246.
- Thelwall, M. (2001). Exploring the link structure of the Web with network diagrams. *Journal of Information Science*, 27(6), 393-402.
- Thelwall, M. (2001). Extracting macroscopic information from web links. *Journal of American Society for Information Science and Technology*, 52(13), 1157-1168.
- Thelwall, M. (2001). The responsiveness of search engine Indexes. *Cybermetrics*, 5(1), <http://www.cindoc.csic.es/cybermetrics/articles/v5i1p1.html>.
- Thelwall, M. (2002). Evidence for the existence of geographic trends in university web site interlinking. *Journal of Documentation*, 58(5), 563-574.
- Thelwall, M. (2002). An initial exploration of the link relationship between UK university web sites. *ASLIB Proceedings*, 54(2), 118-126.
- Thelwall, M. (2003). What is this link doing here? Beginning a fine-grained process of identifying reasons for academic hyperlink creation. *Information Research*, 8(3), <http://informationr.net/ir/8-3/paper151.html>.
- Thelwall, M. (2007). Blog searching: The first general-purpose source of retrospective public opinion in the social sciences? *Online Information Review*, 31(3), 277-289.
- Thelwall, M. (2008). Extracting accurate and complete results from search engines: Case study Windows Live. *Journal of the American Society for Information Science and Technology*, 59(1), 38-50.
- Thelwall, M. (2008). Fk yea I swear: Cursing and gender in a corpus of MySpace pages. *Corpora*, 3(1), 83-107.
- Thelwall, M. (2008). Quantitative comparisons of search engine results. *Journal of the American Society for Information Science and Technology*, 59(11), 1702-1710.
- Thelwall, M. (2009). Homophily in MySpace. *Journal of the American Society for Information Science and Technology*, 60(2), 219-231.
- Thelwall, M., & Harries, G. (2004). Do better scholars' Web publications have significantly higher online impact? *Journal of American Society for Information Science and Technology*, 55(2), 149-159.
- Thelwall, M., & Hasler, L. (2007). Blog search engines. *Online Information Review*, 31(4), 467-479.

- Thelwall, M., & Prabowo, R. (2007). Identifying and characterising public science-related concerns from RSS feeds. *Journal of the American Society for Information Science & Technology*, 58(3), 379-390.
- Thelwall, M., & Price, L. (2006). Language evolution and the spread of ideas: A procedure for identifying emergent hybrid word family members. *Journal of the American Society for Information Science and Technology*, 57(10), 1326-1337.
- Thelwall, M., Tang, R., & Price, E. (2003). Linguistic patterns of academic web use in Western Europe. *Scientometrics*, 56(3), 417-432.
- Thelwall, M., Vann, K., & Fairclough, R. (2006). Web issue analysis: An Integrated Water Resource Management case study. *Journal of the American Society for Information Science & Technology*, 57(10), 1303-1314.
- Thelwall, M., & Wilkinson, D. (2003). Graph structure in three national academic Webs: Power laws with anomalies. *Journal of American Society for Information Science and Technology*, 54(8), 706-712.
- Thelwall, M., Wouters, P., & Fry, J. (2008). Information-centred research for large-scale analysis of new information sources. *Journal of the American Society for Information Science and Technology*, 59(9), 1523-1527.
- Thelwall, M. & Wouters, P. (2005). What's the deal with the web/Blogs/the next big technology: A key role for information science in e-social science research? CoLIS 2005, *Lecture Notes in Computer Science* 3507, 187-199.
- Thelwall, M., & Zuccala, A. (2008). A university-centred European Union link analysis. *Scientometrics*, 75(3), 407-420.
- Trancer, B. (2008). *Click: What millions of people are doing online and why it matters*. London: Hyperion.
- Vaughan, L. (2006). Visualizing linguistic and cultural differences using Web co-link data. *Journal of the American Society for Information Science and Technology*, 57(9), 1178-1193.
- Vaughan, L., & Shaw, D. (2003). Bibliographic and Web citations: What is the difference? *Journal of the American Society for Information Science and Technology*, 54(14), 1313-1322.
- Vaughan, L., & Shaw, D. (2005). Web citation data for impact assessment: A comparison of four science disciplines. *Journal of the American Society for Information Science & Technology*, 56(10), 1075-1087.
- Vaughan, L., & Shaw, D. (2008). A new look at evidence of scholarly citation in citation indexes and from web sources. *Scientometrics*, 74(2), 317-330.
- Vaughan, L., & Thelwall, M. (2004). Search engine coverage bias: evidence and possible causes. *Information Processing & Management*, 40(4), 693-707.
- Vaughan, L., & Thelwall, M. (2005). A modeling approach to uncover hyperlink patterns: The case of Canadian universities. *Information Processing & Management*, 41(2), 347-359.
- Vaughan, L., & Wu, G. (2004). Links to commercial websites as a source of business information. *Scientometrics*, 60(3), 487-496.
- Vaughan, L. & Yang, R. (2012). Web data as academic and business quality estimates: A comparison of three data sources. *Journal of the American Society for Information Science and Technology*, 63(10), 1960-1972.
- Vaughan, L., & You, J. (2005). Mapping business competitive positions using Web co-link analysis. In P. Ingwersen & B. Larsen (Eds.), *Proceedings of 2005: The 10th International Conference of the International Society for Scientometrics and Informetrics* (pp. 534-543). Stockholm, Sweden: ISSI.
- Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications*. Cambridge, NY: Cambridge University Press.
- Zuccala, A. (2006). Author cocitation analysis is to intellectual structure as web colink analysis is to...? *Journal of the American Society for Information Science & Technology*, 57(11), 1487-1502.

