Hyperlink Analysis

Mike Thelwall
University of Wolverhampton, UK

INTRODUCTION

Links between Web pages can be used as a source of information with which to identify types of Web based communities for which interlinking is common, and to investigate community-based linking practices. These communities could be of individuals, organizations or even information sources. Hyperlink analysis, also known as link analysis, provides a set of techniques to aid data gathering, filtering and processing, as well as providing methods to help interpret results.

Hyperlink analysis originated in citation analysis because of the structural similarity between links and citations (Ingwersen, 1998). Whereas links connect Web pages, citations typically connect academic journal articles. Citations have been analyzed for many years by researchers in order to analyze research communities in the scholarly literature (Borgman & Furner, 2002) amongst other things, and the first people to develop link analysis drew from citation analysis experience. Early contributors included Rousseau (1997), who dubbed links *sitations*, and Ingwersen (1998), who used citation counts to estimate the impact of collections of Web pages, using the assumption that more useful pages would tend to receive more links. The creators of Google also harnessed links to deliver more useful search results based upon a similar assumption (Brin & Page, 1998). One of the earliest academic applications of link analysis was to identify "communities" of information around a particular topic (Larson, 1996). Since then many others have used links to help identify Web communities (Flake, Lawrence, Giles, & Coetzee, 2002), or to describe them (Foot, Schneider, Dougherty, Xenos, & Larsen, 2003).

BACKGROUND

Although various forms of link analysis are performed, a social science approach will be given, using links to interpret online community structures rather than to identify them. A social science link analysis has the following broad stages.

1. Identification of Web pages or sites relevant to the topic or community under investigation.
2. Collection of data about links to, between, or from the chosen Web sites.
3. Identification of patterns in the link data.
4. Interpretation of findings, for example through an investigation into why the links were created or what functions they perform.

The fourth stage is particularly important because early research tended to make simplifying assumptions about links, such as that they communicate information or form a social bond between Web page owners. In fact links can be created for very diverse reasons, even in relatively formal areas such as university Web sites (Bar-Ilan, 2004; Thelwall, 2003b). This means that researchers must take care to avoid making assumptions about the *causes* of any link patterns identified.

IDENTIFYING WEB PAGES

In order to investigate Web link patterns, a set of Web pages relevant to the research question must be identified. Examples of collections from previous research include university Web sites (Bar-Ilan, 2004; Thelwall, 2003a), academic department Web sites (Li, Thelwall, Musgrove, & Wilkinson, 2003; Thomas & Willet, 2000), academics' home pages (Kretschmer, 2003), politicians' home pages and political Web sites (Foot et al., 2003; Park, Barnett, & Kim, 2001), business Web sites (Park, Barnett, & Nam, 2002) and pages with information relevant to a given topic (Larson, 1996). The pages are typically found through a combination of search engine searches and the use of pre-existing link lists. Another approach is to obtain an offline list of relevant organizations or people, and then to search for their Web sites online.

Virtual communities of Web sites can also sometimes be identified with the help of links. For example, in an analysis of Web publishing relating to U.S. presidential candidates, the mere fact that a Web page links to the site of one of the candidates could be taken as evidence of its involvement in the debate (Foot et al., 2003). Of course, Web sites mentioning the candidates name may also be involved, whether they link or not, but using links rather than text as a primary source of evidence for community membership can be a practical and necessary step, given the enormous number of pages (e.g., from news sites) that cover U.S. elections.

COLLECTING LINK DATA

In some investigations the links to be investigated are identified manually by researchers visiting and checking the pages individually (e.g., Foot et al., 2003). For some types of investigation this is not possible because there are too many to check. To resolve this problem, the advanced search interfaces of commercial search engines may be used. For example, at the time of writing the link: URL command in Google would return a list of pages indexed by Google that linked to the URL. Similarly the command linkdomain:domain in AltaVista would return a list of pages that are linked to any page with the correct domain name. Using commands like these, researchers can estimate how many links there are to a page or Web site without having to visit all pages that might possibly contain a link. Note, however, that search engines do not index the whole Web and so their link counts will typically be underestimates.

Search engine link commands can also be narrowed down to report the number of links between a specific pair of Websites, but not all search engines provide this functionality. To illustrate an application of this type of data, one study used AltaVista to count the links between the academic Websites of the various countries of the European Union (Musgrove, Binns, Page-Kennedy, & Thelwall, 2003; Thelwall & Tang, 2003). This data was used to show that countries with a common language tended to interlink more and that English language pages were commonly used for international links, even between pairs of non-English speaking countries.

An alternative to search engines for link data is the personal Web crawler. A number of researchers have used crawlers to download collections of Web sites, extracting links automatically from the downloaded sites (Garrido & Halavais, 2003; Thelwall, 2001b). This is more labor intensive and uses more computing resources than commercial search engine queries, at least from the researcher's perspective, but can give more reliable and comprehensive results (Thelwall, 2001b).

PATTERN IDENTIFICATION

At the heart of all link analysis investigations is some kind of search for patterns. Some broad categories of investigation are discussed below and summarized in Table 1. In this table an *inlink* is a link to a site, an *outlink* is a link from a site, whereas *interlinking* refers to linking between two or more Web sites. This terminology reflects the perspective of the analysis; for example an inlink to one site is also an outlink from another. It also reflects an assumption that links within a Web site are typically less interesting than links between Web sites. Internal site links are often for navigational purposes and inter-site links are therefore more likely to reflect significant communication intentions.

Descriptive analysis uses as raw data counts of links to or from a set of Web sites. Comparing the inlink or outlink counts across the set of sites can be used to identify the sites that stand out. For example, the site most highly linked to may be particularly important in a network, and may be regarded as a network "authority". For interlinking data, measures from social network analysis can be applied to identify significant sites, such as those that seem to be the most central in the network (Björneborn, 2004; Park, 2003). Statistically-oriented researchers have gone one step further, fitting mathematical laws to link count statistics (Huberman, 2001; Rousseau, 1997), but this is not necessary for most investigations. At the opposite extreme purely qualitative techniques, such as interviews, can be used to investigate the context of link creation (Hine, 2000; Kim, 2000).

Modeling is a mathematical approach to *explain* link patterns in terms of underlying phenomena. For example, modeling has been used to demonstrate that research productivity can be used to explain why some university Web sites receive many more links than others (Thelwall & Harries, 2004). It has also provided evidence of a snowball effect in link attraction: sites that are heavily linked to are far more likely to attract new links than those that have not yet been linked to (Pennock, Flake, Lawrence,

Table 1. Four types of link pattern identification

Analysis	Data	Examples
Descriptive	Inlinks outlinks interlinking	Compare link counts for a set of sites
Modeling	Inlinks Outlinks interlinking	Producing mathematical models to explain link patterns such as geographic trends
Network diagrams	Interlinking	Displaying links between set of Web sites or pages, using a variety of graphing techniques
Spectral analysis	Inlinks outlinks	Comparing the types of sites that link to or are linked from, a set of Web sites