Scientific Web Intelligence

Mike Thelwall
University of Wolverhampton, UK

INTRODUCTION

Scientific Web Intelligence (SWI) is a research field that combines techniques from data mining, Web intelligence, and scientometrics to extract useful information from the links and text of academic-related Web pages using various clustering, visualization, and counting techniques. Its origins lie in previous scientometric research into mining off-line academic data sources such as journal citation databases. Typical scientometric objectives are either evaluative (assessing the impact of research) or relational (identifying patterns of communication within and among research fields). From scientometrics, SWI also inherits a need to validate its methods and results so that the methods can be justified to end users, and the causes of the results can be found and explained.

BACKGROUND

The term *scientific* in SWI has a dual meaning. The first meaning refers to the scope of the data—it must be academic-related. For example, the data may be extracted from university Web sites, electronic journal sites, or just pages that mention or link to academic pages. The second meaning of scientific alludes to the need for SWI research to use scientifically defensible techniques to obtain its results. This is particularly important when results are used for any kind of evaluation.

SWI is young enough that its basic techniques are not yet established (Thelwall, 2004a). The current emphasis is on methods rather than outputs and objectives. Methods are discussed in the next section. The ultimate objectives of typical developed SWI studies of the future can be predicted, however, from research fields that have used off-line academic document databases for data mining purposes. These fields include bibliometrics, the study of academic documents, and scientometrics, the measurement of aspects of science, including through its documents (Borgman & Furner, 2002).

Evaluative scientometrics develops and applies quantitative techniques to assess aspects of the value of academic research or researchers. An example is the Journal Impact Factors (JIF) of the Institute for Scientific Information (ISI) that are reported in the ISI's journal citation reports. JIFs are calculated for journals by counting citations to articles in the journal over a fixed period of time and dividing by the number of articles published in that time. Assuming that a citation to an article is an indicator of impact (because other published research has used the article in order to cite it), the JIF assesses the average impact of articles in the journal. By extension, good journals should have a higher impact (Garfield, 1979), so JIFs could be used to rank or compare journals. In fact, this argument is highly simplistic. Scientometricians, while accepting the principle of citations as a useful impact proxy, will argue for more careful counting methods (e.g., not comparing citation counts between disciplines) and a much lower level of confidence in the results (e.g., taking them as indicative rather than definitive) (van Raan, 2000). Evaluative techniques also are commonly used for academic departments. For example, a government may use citation-based statistics in combination with peer review to conduct a comparative evaluation of all of the nation's departments within a given discipline (van Raan, 2000). SWI also may be used in an evaluative role, but since its data source is only Web pages, which are not the primary outputs of most scientific research, it is unlikely to ever be used to evaluate academics' Web publishing impact. Given the importance of the Web in disseminating research (Lawrence, 2001), it is reasonable, however, to measure Web publishing.

Relational scientometrics seeks to identify patterns in research communication. Depending on the scale of the study, this could mean patterns of interconnections of researchers within a single field, of fields or journals within a discipline, or of disciplines within the whole of science. Typical outputs are graphs of the relationships, although dimension-reducing statistics, such as factor analysis, also are used. For example, an investigation into how authors within a field cite each other may yield an author-based picture of the field that usefully identifies sub-specialisms, their main actors, and interrelationships (Lin, White & Buzydlowski, 2003). Knowledge domain visualization (Börner, Chen & Boyack, 2003) is a closely related research area but one that focuses on the design of visualizations to display relationships in knowledge domains. Relationship identification is likely to be a common outcome for future SWI applications. An advantage of the Web over academic journal databases is that it can contain more up-to-date information, which could help produce more current domain visualizations. The disad-

vantage, however, is that the Web contains a wide variety of information that is loosely related to scholarly activity, if at all, even in university Web sites. The challenge of SWI and the rationale for the adoption of Web intelligence and data mining is to extract useful patterns from this mass of mainly useless data. Successful SWI will be able to provide an early warning of new research trends within and among disciplines.

MAIN THRUST

SWI uses methods based upon Web links (Web structure mining) and text (Web content mining). A range of relevant mining and structure mining techniques is described in the following section.

Academic Web Structure Mining

Modeling

Early academic Web structure mining sought to assess whether counts of links to university or department Web sites could be used to measure their online impact. This originated in the work of Ingwersen (1998). In brief, the results of this line of research indicated that links between university Web sites, unlike citations, almost never represented knowledge transfer within the context of research. For example, few of these links point to online journal or conference articles. Nevertheless, it seems that about 90% of links are related in some way to academic activities (Wilkinson et al., 2003), and counts of links to universities correlate significantly with measures of research productivity for universities (Thelwall & Harries, 2004) and departments in some disciplines (Li et al., 2003; Tang & Thelwall, 2003). These results are consistent with Web publishing being a natural by-product of research activity (people who do more research tend to create more Web pages), but the chances of any given Web page being linked to does not depend upon the research capabilities of its author, on average. In other words, more productive researchers tend to attract more links, but they also tend to produce more content, and so the two factors cancel out.

A little more basic information is known about academic Web linking. Links are related to geography (closer universities tend to interlink more) (Thelwall, 2002). Links are related to language (universities in countries sharing a common language tend to interlink more, at least in Europe, and English accounts for at least half of international linking pages in European universities in all countries except Greece) (Thelwall, Tang & Price, 2003).

Data Cleansing

An important but unexpected outcome of the research previously described was the need for extensive data cleansing in order to get better results from link-counting exercises. This is because, on a theoretical level, link counting works best when each link is created independently by human experts exercising care and judgement. In practice, however, many links are created casually or by automated processes. For example, links within a Web site are often for navigational purposes and do not represent a judgment of target-page quality. Automatically-generated links vary from the credit links inserted by Web authoring software to links in navigation bars on Web sites. The following types of link normally are excluded from academic link studies.

- All links between pages in the same site.
- All links originating in pages not created by the hosting organization (e.g., mirror sites).

Note that the second type requires human judgments about ownership and that these two options do not address the problem of automatically-generated links. Some research has excluded a portion of such links (Thelwall & Aguillo, 2003), but an alternative more automated approach devised to solve this problem is changing the method of counting.

Several new methods of counting links have been devised. These are deployed under the umbrella term of Alternative Document Models (ADMs) and are, in effect, data cleansing techniques (Thelwall & Wilkinson, 2003). The ADMs were inspired by the realization that automated links tended to originate in pages within the same directory. For example, a mini Web site of 40 pages may have a Web authorizing software credit link on each page but with all site pages residing in the same directory. The effect of these links can be reduced if links are counted between directories instead of between pages. In the example given, the 40 links from 40 pages would be counted as one link from a directory, discarding the other 39 links, which are now duplicates. The ADMs deployed so far include the page ADM (standard link counting), the directory ADM, the domain ADM, and the whole site ADM. The choice of ADM depends partly on the research question and partly on the data. A purely data-driven selection method has been developed (Thelwall, 2005a), designed to be part of a much more automated approach to data cleansing; namely, Multiple Site Link Structure Analysis (MSLSA).