Scientometrics (2013) 95:503–528
DOI 10.1007/s11192-013-0964-1

Journal clustering of library and information science for subfield delineation using the bibliometric analysis toolkit: CATAR

Yuen-Hsien Tseng · Ming-Yueh Tsay

Received: 17 May 2012 / Published online: 24 February 2013

Abstract A series of techniques based on bibliometric clustering and mapping for scientometrics analysis was implemented in a software toolkit called CATAR for free use. Application of the toolkit to the field of library and information science (LIS) based on journal clustering for subfield identification and analysis to suggest a proper set of LIS journals for research evaluation is described. Two sets of data from Web of Science in the Information Science & Library Science (IS&LS) subject category of Journal Citation Reports were analyzed: one ranges from year 2000 to 2004, the other from 2005 to 2009. The clustering results in graphic dendrograms and multi-dimensional scaling maps from both datasets consistently show that some IS&LS journals clustered in the management information systems subfield are distant from the other journals in terms of their intellectual base. Additionally, the cluster characteristics analyzed based on a diversity index reveals the regional characteristics for some identified subfields. Since journal classification has become a high-stake issue that affects the evaluation of scholars and universities in some East Asian countries, both cases (isolation in intellectual base and regionalism in national interest) should be taken into consideration when developing research evaluation in LIS based on journal classification and ranking for the evaluation to be fairly implemented without biasing future LIS research.

Keywords Document clustering · Bibliographic coupling · Journal classification · Research performance evaluation · Freeware

Y.-H. Tseng
Information Technology Center, National Taiwan Normal University, Taipei, Taiwan
e-mail: samtseng@ntnu.edu.tw

M.-Y. Tsay (✉)
Graduate Institute of Library, Information and Archival Studies, National Chengchi University, Taipei, Taiwan
e-mail: mytsay@nccu.edu.tw

Introduction

Journals are footprints by which the development of knowledge in a discipline or profession can be followed (Bush et al. 1997). They are popular unit of various bibliometric/scientometric analyses (Larivière et al. 2012). Applications of journal analysis include field delineation (Zhang et al. 2010) for discipline classification, core journal identification (McCain 1991) for strategic journal selection in library collection or scientometric analysis, and journal ranking (Nisonger and Davis 2005) for evaluation of journal quality or research outputs. These analyses are important in both knowledge exploration and policy development that may affect various stakeholders. As an example, in a study for visualizing library and information science (LIS) concept spaces, Åström (2002) concluded that journal selection does affect how research fields can be perceived and defined. This conclusion can be viewed in the reverse direction, i.e., the delineation (perception or definition) of a research field corresponds to the selection of journals to be included in the field. This view has been implicitly assumed in evaluation of faculty regarding tenure, promotion, and annual raise decisions, when the evaluation utilizes the ranking of selected journals in a discipline as a proxy indicator of research quality in that discipline.

As an example, in a recent research evaluation policy widely adopted in Taiwan, research funding, academic awards, and position promotion of individual researchers (or national research grants for institutes) rely on the ranking of the journals in which the individuals publish their research output (or how well the institutes perform in a specific research field). Since journals may rank differently in different research fields, to which field a journal is categorized may have significant impact on the evaluation. The task of journal classification or clustering to delineate the research fields in an objective and unbiased way is inevitable for this policy to be fairly implemented.

More specifically, when competing for government funding or applying for research projects/position promotion/academic awards, individual scholars or universities need to exhibit how well they perform in terms of academic publications, which often refer to the journal classification system and impact factor ranking of Thomson Reuters' Journal Citation Reports (JCR) database. Publications in higher ranked journals in a subject category receive higher rewards. It turns out that the journal classification and ranking in JCR become a high-stake issue that affects the research evaluation of many scholars and universities in Taiwan. This is also true for those in social science and, consequently, in the LIS field, where international journal articles have now served as the primary bases for research evaluation, rather than other types of publications, such as books or conference presentations. As a consequence, the journals listed in the Information Science & Library Science (IS&LS) subject category of JCR, the most matched subject category for LIS, dictate the research directions of LIS scholars in Taiwan. This similar trend applies to other fields all over Taiwan.

In JCR 2009 edition, there are 66 journals included in the IS&LS category, 77 journals in 2010, and 83 journals in 2011. The growing size and particularly the inclusion of more sub-disciplinary journals in the IS&LS subject category affect the rankings of subfield journals. In consequence, the performance of individuals who devoted to a particular subfield is affected. Without making substantial changes in their research topics, those individuals who used to publish in a certain set of journals are affected in either positive or negative way. As such, their evaluation based on this policy can be biased. A remedy would be to identify stable subfields in the JCR's subject category and research output could be assessed within the sampling of subfield journals.

In this work, we apply scientometric techniques for subfield delineation analysis of LIS based on JCR's IS&LS journal classification and Thomson Reuters' Web of Science (WoS) data. The result could suggest possible improvement of the evaluation system mentioned above. The analysis techniques used include journal clustering, multi-dimensional scaling (MDS) mapping, and some indicators that help identify subfield characteristics. These techniques have been implemented in a freeware toolkit named CATAR for recurrent use and verification.

In the last few decades, there are abundant techniques developed for scientometric analysis. A widely adopted methodology is to compute the similarities among the bibliographic data for clustering based on co-citations, bibliographic coupling, or co-word analysis, e.g., see Yan and Ding (2012). The resulting clusters can be mapped in a two- or three-dimensional space for information visualization and exploration (White and McCain 1997; Noyons and Van Raan 1998). In addition, various indicators and cross-tabulation analysis can then be applied to these clusters for more insightful information (Buter and Noyons 2001; Noyons et al. 1999a). Researchers capable of algorithmic design and implementation have adapted or extended this methodology to fit their analysis tasks. However, most adaptions and extensions are made in an ad hoc way (Moya-Anegon et al. 2006), making verification, comparison, or reuse of them by others infeasible (Börner et al. 2010). Fortunately, some of them have been implemented and packaged in software tools freely available on the Web, which include CiteSpace (Chen 2006; Chen et al. 2010), Sci^2 Tool (Sci^2 Team 2009), VOSviewer (Van Eck and Waltman 2010), BibExcel (Persson 2009), and Sitkis (Schildt and Mattsson 2006). These tools can be applied to many applications of scientometric tasks. But customization or adaptation still is inevitable as scientometric analysis varies from tasks to tasks. Our work follows this idea of making the software implementation freely available so as to facilitate its re-use and application to similar analysis tasks pursued by others.

As such, the objectives of this article include:

(1) To describe a set of bibliometric clustering and mapping techniques that are suitable to certain scientometric analyses and to make the technical implementation public in a software toolkit for re-use in other similar tasks.
(2) To apply the toolkit to the field of LIS for subfield identification and analysis so as to suggest a proper set of journals for ranking, in the hope that such suggestion would lead to more justifiable research evaluation in Taiwan.

The rest of the article is organized as follows. First, related studies are briefly reviewed. The details of the implemented techniques are described next. Application of the toolkit to the IS&LS journals is then presented, followed by the discussion of the implications of the results. This paper is concluded with the strength and limitations of the toolkit and directions for future improvement.

Literature review

The delineation of subfields of LIS has been studied with various scientometric approaches. The frequently used techniques include co-citation analysis for pairing items (such as authors, journals, articles), agglomerative hierarchical clustering (AHC) for grouping items in dendrograms, and MDS for visualizing them in 2- or 3-dimensional maps. The following paragraphs review a number of studies concerning the LIS subfield identification. Their

data sets, methods, tools, and major results are described, so as to provide information for comparison with our work.

Åström (2002) analyzed author co-citation and keyword co-occurrence based on 1135 records from 1998-2000 published in four and five highest ranked information science (IS) and library science (LS) journals, respectively, from JCR to generate the MDS maps with the help of the BibExcel tool. The results revealed three clusters: hard information retrieval (IR), soft IR, and bibliometrics from 52 most cited authors' co-citation map, and three clusters: LS, IR, and bibliometrics from the co-occurrence map of 47 most frequently occurring keywords as well as from the keyword and author combined map. Åström speculated that the absent of LS in the author co-citation map was probably caused by the LS authors' publication patterns, where the other frequent publication channels, such as books and regional journals, are not covered by JCR, and therefore LS authors were under-represented in the citation based ranking. Compared with the analysis of White and McCain (1998) that emphasized on 12 IS journals identifying two sub-disciplines: IR and studies of aspects of literature and communication, Åström's results showed that journal selection does affect how research fields can be perceived and defined.

Åström (2007) conducted another time-slice co-citation analyses based on 21 LIS journals selected from 55 journals covered by the IS&LS category in JCR 2003 version. In this analysis, all general LIS journals were manual identified and the specialized ones were excluded. Based on the most cited documents out of 13,605 articles over the three 5-year periods ranging from 1990 to 2004, the document co-citation maps found two stable subfields: informetrics and information seeking and retrieval (ISR). With the popularity of world wide web, webometrics has emerged as a dominating research area in both informetrics and ISR.

Janssens et al. (2006) applied a series of full-text analysis techniques and "traditional" methods such as MDS and AHC to 938 articles or notes published between 2002 and 2004 in five LIS journals to visualize the salient research topics. The five journals representing the LIS field are: *Information Processing and Management* (IPM), *Journal of the American Society for Information Science and Technology* (JASIST), *Journal of Documentation* (JDoc), *Journal of Information Science* (JIS), and *Scientometrics*. Their optimum solution to cluster the LIS articles resulted in six clusters: two in bibliometrics, one in IR, one containing general issues, and the other two in webometrics and patent study which were identified as small but emerging clusters.

Moya-Anegon et al. (2006) choose 17 out of 24 journals listed as having the greatest impact in JCR 1996 edition to visualize the field structure of LIS. The rejected journals have editorial scopes related to the application of IS to a specific technique or area of knowledge (e.g., medicine, geography, telecommunications), with LIS as a secondary interest. From the cited references of the journal articles published in the period 1992–1997, 77 most cited authors and 73 most cited journals were selected for co-citation analysis. The results were mapped based on self-organization map, MDS, and AHC, using an ad hoc program for preprocessing and the Statistica software package for analysis. The author co-citation analysis (ACA) resulted in six subfields: scientometrics, citationist, bibliometrics, soft IR (cognitive), hard IR (algorithmic), and communication theory, while the journal co-citation analysis (JCA) led to four domains: IS, LS, science studies, and management. The science studies contains the only LIS journal: scientometrics and other non-LIS journals such as Nature, Science, Cell, etc. This domain from JCA roughly corresponds to the subfields of scientometrics, citationist, and bibliometrics from ACA, and the IS domain relates to the subfields of Soft IR and Hard IR. There is no clear correspondence between the communication theory from ACA and the management from

JCA, and most notably, there is no subfield from ACA corresponding to the LS domain from JCA. Moya-Anegon et al. (2006) pointed out that nearly all the most cited authors have works published in only IS and Science Studies. Just as what Åström (2002) speculated, the authors in LS were not cited enough to exceed the most cited threshold and therefore the LS subfields were not seen from ACA.

Ludo Waltman et al. (2011) used JASIST as a seed journal and selected other 47 journals which are most strongly related with JASIST based on co-citation data to form the field of LIS for further analysis. Among these 48 selected journals, none of them are from the journals in the manage information systems (MIS) cluster like the management identified from JCA by Moya-Anegon et al. (2006). Based on the 12,202 publications in the period 2000–2009, the 48 journals were grouped into three subfields: LS, IS, and scientometrics by VOSviewer using journal bibliographic coupling (JBC).

Milojevic et al. (2011) used co-word analysis to explore the cognitive structure of LIS based on 10,344 articles published between 1998 and 2007 in 16 journals which were selected from the ranked list complied in the perception study of Nisonger and Davis (2005). The most frequently occurring 100 terms were extracted from the article titles, and their co-occurrence and AHC were produced by use of WordStat. Three main branches were found: LS, IS, and bibliometrics/scientometrics, each with 10, 6, and 3 sub-branches, respectively. Clustering of the 16 journals based on the 100 terms by use of AHC and MDS led to the same three branches.

The above studies identified subfields in LIS with various numbers of representative LIS journals. The journals are first manually determined and then the subfields are analyzed. Although the identified subfields are good sources for our reference, their objectives are different from ours in this work, as we want to enumerate as many LIS relevant journals as possible, or to identify relevant ones from a given set of journals, so as to assign each relevant journal to an LIS subfield, for the task of research output evaluation based on journal ranking. The above studies use manually selected or less complete set of journals for clustering. The incomplete set does not cover each relevant journal to a subfield; the manually selected set has the consensus problem, as different studies used different numbers of journals as surrogates to study the LIS field or subfields.

The ranked list of 71 LIS journals compiled by Nisonger and Davis (2005) in a survey according to the opinions of 56 deans of schools with American Library Association-accredited LIS programs and directors of 120 ARL libraries could serve as a compete set of LIS relevant journals or a source for journal subfield analysis, in addition to those covered by JCR's IS&LS category. However, as Milojevic et al. (2011) mentioned: (1) it is based on the opinions of one group (deans and directors) over a broader range of stakeholders; (2) it is focus on what is perceived to be important in traditional library and information schools. In addition, this journal list based on LIS administrators in the US is less likely to be accepted by other disciplines in Taiwan when cross-disciplinary ranking is required using JCR's subject categories.

The series studies conducted by (Ni and Ding 2010; Ni and Sugimoto 2011; Ni et al. 2012) are most relevant to ours. From 61 journals covered by IS&LS category of JCR 2008 edition, Ni et al. (2012) selected 58 journals for clustering based on four methods: venue-author-coupling, journal co-citation analysis, co-word analysis, and journal interlocking. The three rejected journals are published in languages other than English, which causes difficulty for co-word analysis and journal interlocking verification. Their MDS and AHC analyses resulted in four to five subfields. The consistent subfields derived from these four methods include: MIS, IS, LS, and specialized (and communication) clusters. The MIS cluster consists of mainly eight journals and six of the eight journals ranked at top 10 in

2008 JCR (Ni and Ding 2010). However, the dissimilarity between the MIS journals and other subfield journals in these four methods is distinguishable as the four MDS maps consistently show that the MIS journals are separated from the other clusters. According to the result, journals in the IS&LS category in JCR are not firmly connected with LIS research, and proper re-organization of LIS journals in JCR is suggested by Ni and Ding (2010) and Ni and Sugimoto (2011).

Although Ni et al. (2012) have studied the LIS journal clustering with four methods and wide coverage of IS&LS journals, a complete journal set is necessary for journal ranking as suggested above and additional methods could be tried for complementary information as suggested by Chang and Huang (2012). In this work, we cluster a complete set of IS&LS journals, regardless of their languages, using JBC not only to complement the work of Ni et al. (2012), but also to find the evidence that the IS&LS category include subfields that may affect the research assessment from a different perspective.

The reviewed studies above used various clustering and mapping tools, such as WordStat or Statistca. To follow the above studies using other time-span data requires building the same processes by repeating the data conversion, preprocessing, or even customized computation. As with more freeware built in various fields to facilitate research, scientometric free tools have been released from time to time. BibExcel, developed by Persson (2009) in Windows environment, could be used for citation/co-citation analysis, bibliographic coupling, clustering, and exporting data to Pajek or NetDraw for network visualization and analysis, based on WoS records or other formats with similar field structures. Sitkis, developed by Henri Schildt (Schildt and Mattsson 2006) with Java and Microsoft Access, could import WoS data for citation analyses and clustering, generate statistics based on authors' countries or universities, and export UCINET-compatible data files for other network analyses. VOSviewer, developed by Van Eck and Waltman (2009) in Java, unifies mapping and clustering in one operation (Waltman et al. 2010) and originally supports science mapping based on given matrix data only. It now directly supports import of WoS data and subsequent mapping operations. CiteSpace, developed by Chen (2006) in Java, enables import of WoS, PubMed, and other scientific data for co-citation analysis, network visualization and clustering, automatic cluster labeling, and analyzing trends and pivotal points in scientific development. Sci^2 Tool (Sci^2 Team 2009), adapted from Network Workbench (Börner et al. 2010) for scientometric purposes, supports the temporal, geospatial, topical, and network analyses directly from WoS, Scopus, and other scientific records based on direct linkage, bibliographic coupling, co-citation, and co-word analysis in different levels of data aggregation (such as authors, institutes, and countries). Leydesdorff's software (http://www.leydesdorff.net) is a set of more than 20 executable command-line files that support different analyses, such as co-word, co-author, author bibliographic coupling, JBC, author co-citation, collaboration of nations/institutes, etc. A more detailed description and comparison of nine science mapping tools (including commercial software) could be found in the work of Cobo et al. (2011).

It is noted that these tools may be updated aperiodically and thus their functions could be improved from time to time. For example, although the latest released version of VOSviewer adds co-word analysis only, it is actually able to do more bibliographic coupling and co-citation analyses for documents, journals, authors, and organizations, as demonstrated in the 2nd Global TechMining Conference in September 2012 (Van Eck and Waltman 2012).

The above tools streamline individual data transformation processes and provide diverse bibliometric/scientometric analyses. As indicated by Cobo et al. (2011), there is no single tool able to do all the analyses. If the users' needs fit their design purposes, however, these tools could greatly improve analysis efficiency. They could also be used in parallel for

cross verification for the same analysis tasks, therefore improving the validity, providing different perspectives, or identifying abnormity of the results. In our cases for scientometric analysis, we needed multistage clustering (Tseng et al. 2007) and some indicators such as trend index (Tseng et al. 2009), which have not yet been supported in existing tools. Therefore, we developed our own, as described in the next section.

Methods

Scientometrics is itself a science concerned with measuring and analyzing science (Leydesdorff 2001; Moed 2005; Van Raan 1997). Among its various approaches to reveal quantitative features and characteristics of a research field, bibliometric clustering and mapping analysis based on existing scientific publications is often used. For years, scientometricians have developed effective processes to analyze these data. Various examples can be found in classical literature such as those in Carpenter and Narin (1973), Small and Koenig (1977), and Noyons et al. (1999b). General workflows of bibliographic clustering and mapping for scientometric studies have been summarized by Börner et al. (2003). The general processes may remain the same; the detailed steps could vary from task to task. As shown in the previous section, analytic variations for specific tasks in each study still emerge, and so do the tools that streamline these analytic processes. The analytic process described here follows the general workflows, with a set of information techniques developed to add new features or insights to this process. Specifically, our implementation for scientometric analysis by way of bibliometric clustering and mapping takes the following steps:

1. *Data collection* defining the scope of scientific publications and collecting the corresponding document corpus for analysis.
2. *Text segmentation* identifying the title, authors, citations, and other fields of each article in the corpus.
3. *Similarity computation* calculating the similarity between each pair of documents based on their common features (such as keywords or references) for document clustering.
4. *Multi-stage clustering (MSC)* recursively grouping similar documents/clusters into larger clusters based on the above similarities until a manageable number of topics or sub-fields emerged from the collection.
5. *Cluster labeling* generating cluster descriptors for ease of cluster interpretation.
6. *Visualization* creating a 2-dimensional map based on the MDS technique for visually revealing the relations among the resulting clusters.
7. *Facet analysis* cross-tabulating detected sub-fields with other facet data such as authors, institutes, countries, citations, and publication years to know the most productive or influential agents, and other worth-noting events.

The details of each step are described below. The rationales behind the design and techniques for each step are discussed, and a number of examples are given whenever appropriate. These analytic steps and technical details are adapted from our previous work (Tseng 2010; Tseng et al. 2007) for patent processing. In this work, we package the techniques in each adapted step for scholar bibliographic data into a software tool called *Content Analysis Toolkit for Academic Research* (CATAR), which can be downloaded from http://web.ntnu.edu.tw/~samtseng/CATAR/. CATAR can not only perform journal clustering for sub-field identification as required by the analysis in this work, but also can it

conduct general document clustering for topic analysis from a set of free-text documents. Thus, the following descriptions are not limited to the journal-clustering task, but include other tasks that CATAR can provide.

Data collection

A set of documents needs to be defined and collected once the objective of the analysis is determined. This is very crucial as meaningful results depend on the proper input that fits the objective. Since Thomson Reuters' Web of Science (hereafter WoS) and JCR are the two databases that are referred to for demonstrating individual's performance in most research evaluation, we focus our analyses on WoS and JCR data.

Text segmentation

WoS provides rich search features to define the document set and allows users to download all the records from the search result. The downloaded files are pure texts, with each publication record containing about forty fields. CATAR processes these files to identify individual records, extract about a dozen fields from each record, parse information within fields (such as author's address for institute and department information), normalize the parsed data (e.g., institutes names are capitalized for matching and counting), and finally store them into a database system for ease of management (e.g., duplicate removal) and verification based on preliminary statistics (e.g., whether the number and the year range of the downloaded records are as expected). These data pre-processing procedures ensure the uniqueness of each record and the correctness of the results that follows from the next steps.

Similarity computation

Document clustering starts from defining the document features and computing similarity between documents based on their features. To derive sub-disciplines for a research field from scientific publications based on document clustering, we regard journals from the discipline as documents and the union of the cited references in each article of a journal as the document's features. Similarities between each pair of journals are then calculated based on the common features normalized by the individual features each journal possesses. Specifically, a similarity based on the Dice coefficient (Salton 1989) between two journals X and Y was used:

$$\mathrm{Sim}(\mathrm{X},\ \mathrm{Y}) = 2 \cdot |R(\mathrm{X}) \cap R(\mathrm{Y})| / (|R(\mathrm{X})| + |R(\mathrm{Y})|)$$

where $R(\mathrm{X})$ denotes the concatenation of the references cited by the articles in journal X, $|R(\mathrm{X})|$ denotes the number of elements in $R(\mathrm{X})$ (i.e., the number of total references in journal X), and $R(\mathrm{X}) \cap R(\mathrm{Y})$ is the common elements of the sets of $R(\mathrm{X})$ and $R(\mathrm{Y})$. The value of this similarity ranges from 0 to 1, denoting from most dissimilar to most similar.

This kind of similarity forms the basis of bibliographic coupling in scientometrics, where it is believed that the more the same references two articles cite, the more likely the two articles are about the same topic. As an example, if article X cites 10 references and Y cites 15 references, and if there are 5 common references among them, then the similarity between X and Y is $2 \times 5/(10 + 15) = 10/25 = 0.4$, or X and Y are bibliographically coupled with a measure of 0.4. Although bibliographic coupling was originally proposed

for finding similar articles, the same idea can be applied to group journals with similar topics of interest.

One obvious advantage of using cited references instead of using articles' keywords as the journal's features is to avoid the problems of polysemy or synonyms associated with some keywords. The polysemy problem occurs when a certain keyword represents two or more different concepts in different contexts, which may overestimate the similarity between two articles (or journals). The synonym problem occurs when a concept is expressed in different keywords, which may underestimate the similarity between two articles (or journals). In addition, some journals and their articles may be published in a language different from the others, which would deteriorate the keyword match problem. On the contrary, if two articles cite the same reference, such as the normalized one from WoS: GARFIELD E, 2003, J AM SOC INF SCI TEC, V54, P400, they should have some topics in common, regardless of the wording and language they used.

However, for those documents without normalized references, CATAR provides co-word analysis as an alternative. Free text words from the title and abstract of each document are lowercased, stemmed, and stop word removed. The resulting words, together with those key phrases extracted from the same document based on the techniques proposed by Tseng (1998, 2002), are used as normalized document features. Pair-wise document similarities are then computed based on the above Dice coefficient for later clustering.

Multi-stage clustering (MSC)

With the pairwise similarities described above, the knowledge structure underlying the documents can be detected by clustering algorithms. CATAR uses an algorithm called complete-linkage hierarchical clustering (Salton 1989), which is commonly used in scientometrics studies [see Fernandez-Cano and Bueno (2002), Jarneving (2007), Ahlgren and Jarneving (2008), and Ahlgren and Colliander (2009) for examples]. The basic idea of this AHC algorithm regards each document as a singleton cluster at first. It then groups the most similar pair of clusters (with similarity larger than a user-specified threshold) into a larger cluster. The same grouping rule applies again to the remaining clusters and newly created ones, where the similarity between any two clusters is defined as the minimum similarity between any pairs of documents each resides in the opposite cluster. This process repeats until no clusters can be merged. In this way, each of the documents is assigned to a cluster automatically. The overall result, if visualized in a graph, is called a dendrogram as shown in an example in Fig. 1, which draws 17 documents and their resulting grouping in a tree like diagram. If the threshold was set to 0.07 in the example, then there would be six clusters (trees), containing (D_1, D_2, D_3), (D_4, D_5, D_6), (D_7, D_8, D_9), (D_{10}, D_{11}, D_{12}, D_{13}), (D_{14}, D_{15}), and (D_{16}, D_{17}), individually. CATAR produces three types of dendrograms: one is in pure HTML text with most detailed information about the clustering results, the second is like the one in Fig. 1 rendered by cascading style sheets and JavaScript, and the third is rendered by Bézier curve graphics. The last two use the free software developed by Robin W. Spencer available from http://scaledinnovation.com/ for aesthetic and compact representation of the clustering results.

The advantage of AHC is that the information of the most similar pairs and groups are all retained in the dendrogram. Another advantage is the relative ease of determining the threshold to group the documents, because the dendrogram provides a scaffold and visual clues to decide the threshold. To quantitatively deciding the most proper threshold, however, the Silhouette indexes (Ahlgren and Jarneving 2008; Rousseeuw 1987;

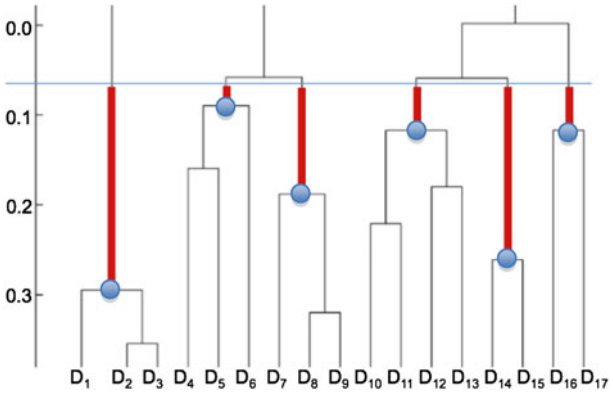

Fig. 1 A dendrogram showing the result of the hierarchical clustering. When a threshold, say 0.07, is set, the dendrogram can be split into several clusters as marked by the *dots* at their roots

Janssens et al. 2006) for each clustering, resulting from a sequence of varying thresholds, are computed. The optimal clustering was determined based on the threshold that leads to the best level of Silhouette index.

Despite the above advantages, when there are large numbers of documents for clustering, direct AHC often does not yield ideal cluster sizes and manageable cluster numbers for manual analysis. CATAR uses a MSC strategy which has been demonstrated in Tseng et al. (2007) as an effective remedy to cope with this situation. In each MSC stage, the strategy first eliminates the outliers (clusters having low similarities with others) and treats each remaining cluster as a virtual document (e.g., the references of the cluster is again the union of all the references of the documents in the cluster). It then clusters the virtual documents by the AHC described above. The same process repeats stage by stage until reasonable clusters emerge, with proper thresholds specifiable by users for each stage. In this way, documents are grouped into categories, which are further clustered into subfields, which in turn can be grouped into fields or domains, as shown in Fig. 2. Although this is not always the case, it represents an expected ideal knowledge structure for the document set.

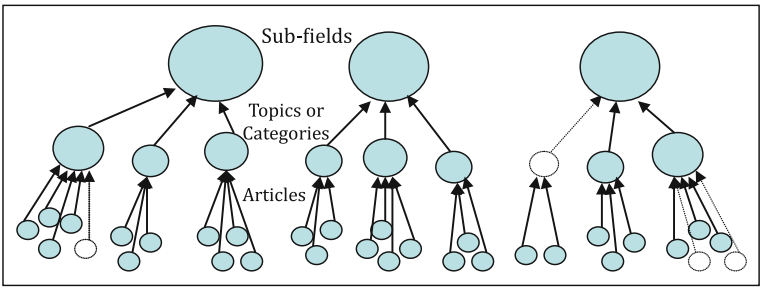

Fig. 2 A conceptual sketch of the multi-stage clustering approach, where *dashed white circles* denote outliers

Cluster labeling

Once the documents are organized into clusters, analysts need to browse the titles or even abstracts to know their content. To help analysts spot the topic for each cluster without much effort, a text mining approach (Tseng 2010) is used to generate cluster descriptors automatically. First, a stop list of non-semantics bearing words, e.g., the, of, and, etc., (van Rijsbergen 1979) is created to filter words in the titles and abstracts. Second, important terms are extracted from each article's text fields (i.e., title and abstract) based on an algorithm that extracts maximally repeated word sequences (Tseng et al. 2007). The correlation coefficient is then computed between each term T and each cluster C using the following equation:

$$\mathrm{Co}(T,C) = \frac{(\mathrm{TP} \times \mathrm{TN} - \mathrm{FN} \times \mathrm{FP})}{\sqrt{(\mathrm{TP}+\mathrm{FN})(\mathrm{FP}+\mathrm{TN})(\mathrm{TP}+\mathrm{FP})(\mathrm{FN}+\mathrm{TN})}}$$

where TP (true positive), FP (false positive), FN (false negative), and TN (true negative) denote the number of documents that belong or not belong to C while containing or not containing T, respectively, as shown in Table 1.

The correlation method is effective for large number of clusters having short documents in them. But it tends to select specific terms that are not generic enough for clusters having a few long documents (or the virtual documents described above), because it does not take into account the number of occurrence of a term in the cluster (i.e., the sum of the term's occurring frequency in each of those documents inside the same cluster). In other words, it is effective for the smaller clusters resulting from the initial clustering stage, but does not yield proper descriptors for the larger clusters from the higher clustering stage. As a remedy for the higher stage clustering, the multiplication principle verified by Tseng (2010) is used to combine the correlation coefficient and the number of occurrence of a term in the cluster to rank the terms for more effective descriptor generation. The resulting cluster descriptors are shown in the HTML format of the dendrogram, while the more concise dendrograms generated by JavaScript do not have these descriptors for aesthetic reason.

Visualization

To represent the detected knowledge structure, two techniques are used by CATAR: one is the previously introduced MSC, the other is MDS (Kruskal 1997). Based on the pre-calculated similarities between each topic, the MSC method organizes the topics in a hierarchical way. This creates a structure that is readily available to a folder tree or topic tree representation, as demonstrated in Fig. 1. As stated in (Janssens et al. 2006), one of the disadvantages of AHC is that wrong choices (merges) that are made by the algorithm in the early stage can never be repaired (Kaufman and Rousseeuw 1990). Therefore, MDS is often used as a complementary tool to explore the relations of the topics. The MDS

Table 1 Confusion matrix for the number of documents (not) containing term T inside (outside) cluster C

Cluster C	Term T	
	Yes	No
Yes	TP	FN
No	FP	TN

technique computes the coordinates of each topic from the pair-wise similarities in the specified dimensions of Euclidean space, which are usually 2 or 3 for ease of visual interpretation. With these coordinates, a topic map is created by a plotting tool. CATAR uses the MDS program in the RuG/L04 freeware package (Kleiweg 2008) for coordinate computation and the GD module in the Perl programming language (Wall et al. 2000) for plotting. To utilize functionalities of existing visualization tools, CATAR also produces output files readable by VOSviewer (Van Eck and Waltman 2010) and Pajek (Nooy et al. 2012) for additional visualization manipulations, such as interactive zooming, scrolling, and searching.

Facet analysis

Once the subfields or topics have been identified, it becomes easy to cross-tabulate the topics with other facet information, because the data from WoS contain rich fields describing an article. This kind of cross analysis often leads to more information than a single facet analysis can provide. For example, it is possible to know the topic distribution of all productive authors (and thus the major domains of their expertise), instead of knowing only their productivity.

In addition to cross-tabulating the results, some indexes are introduced to help quantify the data in the cross tables. For example, the Herfindahl index, also known as Herfindahl–Hirschman Index, or HHI, (Hirschman 1964; Calkins 1983) is an economic indicator that measures the amount of competition (or monopoly) among some actors in a market. It is equivalent to the Simpson diversity index (Simpson 1949) that measures the diversity of species in an ecological environment. In our application, this index can be used to reveal the level of globalization (or localization) of a journal (a cluster, or a sub-field), which can provide more insights when interpreting the results. The index, called HHI in CATAR, is defined as:

$$\mathrm{HHI} = \sum_{i=1}^{n} S_i^2$$

where S_i denotes the publication share of country i in a journal cluster and n is the total number of countries contributing papers to the journal cluster. HHI is proportional to the average publication share, weighted by individual country's publication share. As such, it ranges from 1/n to 1.0, moving from a huge number of contributing countries to a single dominant country. Interestingly, the reciprocal of the index (e.g., 1/HHI) indicates the "equivalent" number of dominant countries in the cluster (Liston-Heyes and Pilkington 2004). For example, if the HHI of the authors' countries for a journal (or subfield) is 0.5, it means that there are equivalently only two ($1/0.5 = 2$) countries that contribute to the journal (or subfield), indicating that the journal is local to some areas or the subfield was studied only in areas with similar cultures. HHI has been used to measure concentration of received citations (Yang et al. 2010; Evans 2008; Larivière et al. 2009) and word concentration for term selection (Milojevic et al. 2011). The use of HHI to measure dominance of contributing countries in a cluster is a new attempt in scientometric study.

Another useful index for revealing cluster characteristics is the trend indicator verified by Tseng et al. (2009), where the number of publications per year in a topic is listed as a time series and the slope of the linear regression line that best fits the time series data is used to indicate the trend of the topic. Tseng, et al. (2009) shows that it is the rank of the trend index that matters, rather than the trend index itself, in the observation of the trend.

The index is particularly useful if there are a large number of topics to be monitored from a large stream of scientific publications.

Results

As shown in the "Literature review" section, previous studies using journals from JCR's IS&LS category to identify cognitive structures of LIS tend to yield four to five subfields: IS (including hard IR, soft IR, and information seeking), LS (practical vs research-oriented), scientometrics (bibliometrics, informetrics, and webometrics), MIS, and other peripheral topics. The distinctness of the MIS cluster has led many studies to remove these journals from subsequent analyses, arguing that they should not be included in the same JCR class (Larivière et al. 2012).

The inclusion of the MIS journals in JCR's IS&LS category also induces debates in journal ranking based evaluation in Taiwan, where the classification system in JCR was used extensively as a convenient tool for research performance evaluation. As an example, in the evaluation system developed recently at National Taiwan Normal University, authors whose papers published in one of the top 10 % journals ranked in any of 228 subject categories of JCR, based on the journal's impact factor, received two times more reward than those who published papers in lower-rank journals in the same category. Unfortunately, some prestigious LIS journals were not ranked as high as they are in the journal prestige ranking surveyed by Nisonger and Davis (2005). Table 2 lists the top ten journals in IS&LS category of JCR version 2009. The *Journal of the American Society for Information Science and Technology* (JASIST) ranks at 7 among 66 journals and is thus excluded from the top 10 % journals (7/66 = 10.60 %), and so is *Scientometrics* which ranks at 10 among 66. This trend remains the same in 2010 and 2011, where JASIST ranks at 11/77 = 14.29 % and 10/83 = 12.05 %, and *Scientometrics* ranks at 14/77 = 18.19 % and 12/83 = 14.46 % in 2010 and 2011, respectively. However, The journal prestige ranking survey based on 56 deans, directors, or department chairs conducted by Nisonger and Davis (2005) identified JASIST as the top 1 journal and *Scientometrics* at rank No. 7 among 71 journals. The fact that these two journals are outside the top 10 % journals in IS&LS category is due to the inclusion of MIS related journals. In fact, in Taiwan the two fields of LS and MIS belong to different colleges and their research evaluation mechanism in the college level is different. The classification system of JCR that groups them together under one subject category not only baffles most LIS scholars in Taiwan, but also impacts those authors of the prestigious journals by incautiously down-grading their contribution when it comes to performance ranking among colleges.

The above situations motivate us to examine the issues of the journal classification in JCR. In particular, we explore the complete journals of the IS&LS category to understand the characteristics of its constituent journals with a hypothesis that the IS&LS journals are in different subfields and that some of these subfields have distinct intellectual base from the others for a sufficient long period of time. Our intent is to find evidence in support of this hypothesis to help develop a more objective journal classification system, and in turn, more suitable research evaluation policy. To this ends, we apply CATAR to identify the sub-fields in LIS based on JBC.

Two sets of data records from WoS were analyzed. These were records published by the 66 journals listed in IS&LS subject category of JCR 2009 version. The first set (Set 1) includes the records from 2000 to 2004 and the second set (Set 2) includes the records from 2005 to 2009. Among these 66 journals, there are some journals that were not indexed

Table 2 The ranks and impact factors of top ten journals in the subject category of information science and library science in JCR version 2009

Rank	Abbreviated journal title	Articles	Total cites	Impact factor	5-Year impact factor
1	MIS Quart	38	6,186	4.485	9.208
2	J Am Med Inform Assn	105	4,183	3.974	5.199
3	J Comput-Mediat Comm	60	1,279	3.639	N/A
4	J Informetr	33	253	3.379	3.379
5	Annu Rev Inform Sci	10	563	2.929	3.030
6	Int J Comp-Supp Coll	19	229	2.692	3.655
7	J Am Soc Inf Sci Tec	203	5,167	2.300	2.480
8	Inform Manage-Amster	56	3,276	2.282	4.297
9	J Assoc Inf Syst	31	430	2.246	N/A
10	Scientometrics	189	3,508	2.167	2.793

during the earlier period of 2000–2004. Therefore, the total numbers of journals within these two sets are 50 (2000–2004), and 66 (2005–2009), and the corresponding numbers of articles within each set are: 9,546 and 11,471, respectively.

Figure 3 shows the clustering results presented in Bézier curve dendrograms, as described in subsection: "Multistage clustering". The results are derived by complete-linkage clustering with the similarity threshold set to zero. Set 1 leads to six isolated clusters having multiple journals (left column of Fig. 3) and three clusters having single journal (not shown). Set 2 also results in six clusters having multiple journals (right column of Fig. 3) and four clusters with single journal (not shown). For ease of comparison and identifying the transition between the two five-year spans, the clusters having similar journals are placed side by side. As an example, in the first cluster, the journal *Scientometrics* being in the cluster with most journals in management information systems during 2000–2004 has evolved into an independent cluster together with *Journal of Informetrics* and *Research Evaluation* during 2005–2009, signifying the growth of this particular topic in LIS.

To allow more coherent topical clustering, a higher similarity threshold 0.01 based on optimal level of Silhoutte values was set to obtain another outcome, in which less relevant journals are excluded from being grouped together in the same clusters. This results in eight clusters in Set 1 (2000–2004) and nine clusters in Set 2 (2005–2009), as marked by the dots in Fig. 3, where the cluster ID is labeled at the root of the corresponding cluster tree. From the derived cluster descriptors as described in subsection: "Clustering labeling", the topics of these clusters can be identified, as listed in Table 3. For example, cluster 4 in Set 1 contains two journals (*Research Evaluation* and *Scientometrics*) and has the cluster descriptors as: "patent, bibliometric, scientometric, citation, indicator". To facilitate comparison, the clusters from both sets are aligned in a way that clusters having similar topics are placed in the same row. As can be seen, cluster 2, 4, 6, 8 in both sets contain similar journals. The others are slightly exceptions to this alignment: In the first row of Table 3, there are two clusters in each set. This is due to the fact that cluster 1 in Set 1 contains the journals of cluster 9 in Set 2 and cluster 1 in Set 2 contains the journals of cluster 5 in Set 1, in addition to the fact that cluster 1 in both sets contains similar journals. Similarly, in the fourth row cluster 3 and 7 in Set 1 correspond to cluster 3 in Set 2. Finally, cluster 5 and 7 in Set 2 are new journals with regional or emerging topics.

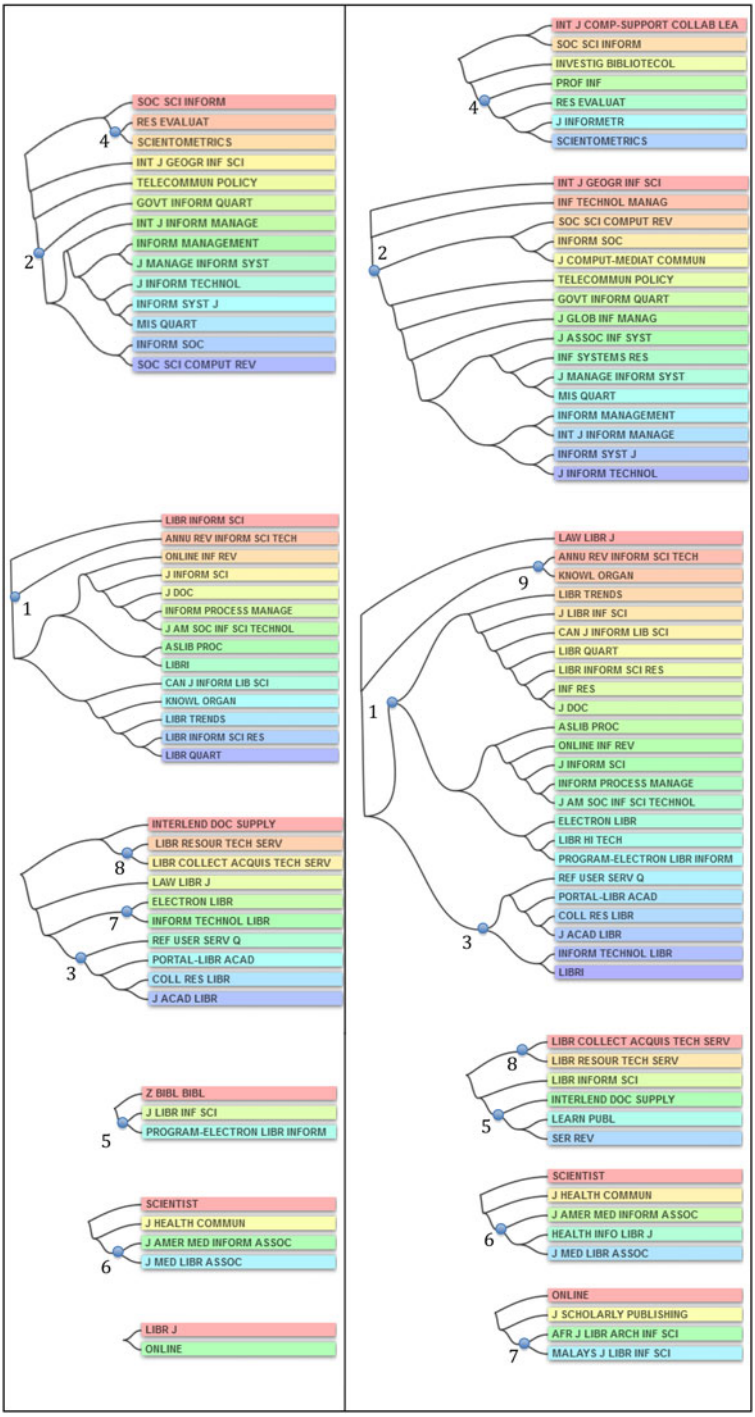

Fig. 3 Clustering results in dendrograms derived by journal bibliographic coupling and complete-linkage clustering with the similarity thresholds set to 0.0 and to 0.01 from Set 1 (*left column*) and Set 2 (*right column*)

Table 3 Major subfields derived from the journals of IS&LS of JCR

Subfield label	Set 1 (2000–2004) Cluster ID (no. of journals): descriptors	Set 2 (2005–2009) Cluster ID (no. of journals): descriptors
IR	1 (13): Retrieval, search, library, information retrieval, digital	1 (15): Library, retrieval, digital, science, search
	5 (2): Library, public library, ict, overview, higher education	9 (2): Knowledge organization, ontology design, control vocabulary, retrieval, concept
MIS	2 (10): Information system, information technology, software, investment, management	2 (14): Information system, adoption, mobile, technology, e-government
SM	4 (2): Patent, bibliometric, scientometric, citation, indicator	4 (4): Patent, scientific, bibliometric, citation, indicator
AL	3 (4): Library, academic, information literacy, academic library, reference	3 (6): Library, academic library, information literacy, librarian, student
	7 (2): Library, academic library, computer game, portal, digital	
ML	6 (2): Clinical, patient, medical, informatic, medical informatic	6 (3): Clinical, health, care, public health, health care
CD	8 (2): Catalog, collection development, technical service, library, acquisition	8 (2): Technical service, library, catalog, collection development, academic library
OA		5 (3): Journal, open access, publish, library, publisher
RL		7 (2): Nigerian, library, ghana, university, nigerian university

From the above alignment, there are six major subfields in the subject category of IS&LS of JCR. They are manually labeled in Table 3 based on the cluster descriptors as: (1) IR; (2) MIS; (3) scientometrics (SM); (4) academic library (AL); (5) medical library (ML); (6) collection development (CD); and two small subfields: (7) open access (OA) and (8) regional library (RL). To further study their topical similarity, the journals scattered over a two-dimensional map based on bibliographic coupling similarity using MDS mapping as described in subsection: "Visualization" are shown in Figs. 4 and 5 for Set 1 and Set 2, respectively. As can be seen, the geographical relations in the two sets are similar. The journals in IR scatter over the upper right hand side of the map, mingled with journals in SM (such as *Scientometrics*). Major journals in MIS locate in the left-hand part of the map and are isolated from the others. Journals in the other subfields aggregate in the lower right-hand corner, indicating their closer similarity in topics within LS.

From Figs. 3, 4, 5 and Table 3, it can be seen that the journals in MIS indeed are very dissimilar from those in LS. The classification system that groups them together by JCR has profound impact on individual scholars and institutes when it comes to research evaluation relying on this classification system.

As another evidence, Table 4 illustrates the most productive departments for the first four clusters, which is a standard output of CATAR as described in subsection: "Facet analysis". As can be seen, those authors who publish most in the MIS subfield (cluster 2) all come from the departments of management information systems or business schools. Admittedly, there are authors from MIS related departments publishing papers in IR (such as Department of Information Management) and SM (e.g., School of Management). This shows that some of the research topics in the MIS subfield are highly related to those in IR

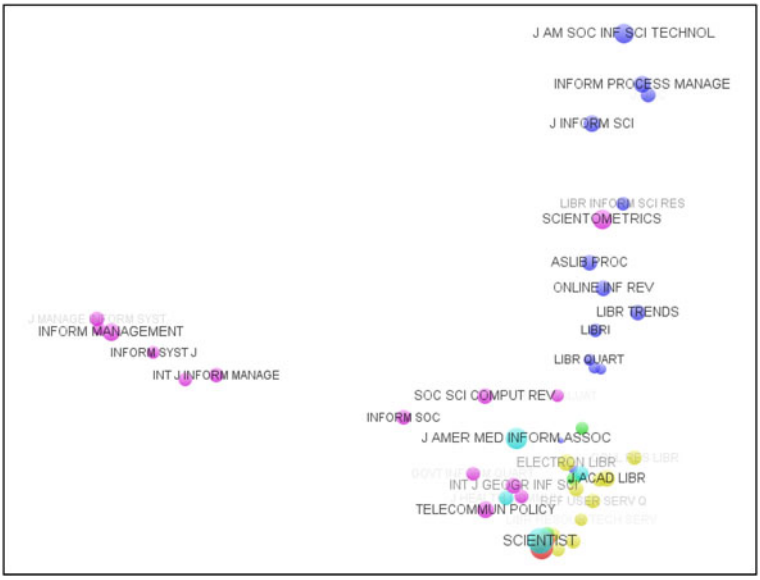

Fig. 4 Topical map of journals from Set 1 based on bibliographic coupling similarity using MDS mapping. This map is created by CATAR and rendered by VOSviewer

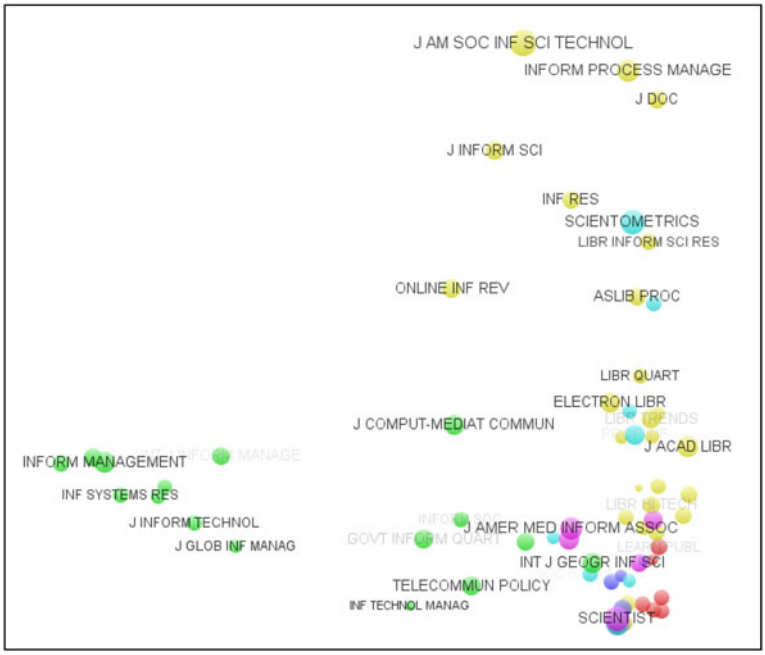

Fig. 5 Topical map of journals from Set 2 based on bibliographic coupling similarity using MDS mapping. This map is created by CATAR and rendered by VOSviewer

Table 4 The top ten productive departments for the first four clusters from Set 2

1. (IR)		2. (MIS)		3. (AL)		4. (SM)	
3,273 Docs.	15	2,126 Docs.	14	915 Docs.	6	1,040 Docs.	4
Dept of Comp Sci	159	Sch of Business	114	University Lib	29	Universiteit Hasselt/Antwerpen[b]	22
Dept of Info Sci	134	Dept of Info System	79	Sch of Lib & Info Sci	25	Dept of Lib & Info Sci	20
Dept of Info Studies	134	Sch of Management	77	Chicago[a]	19	Dept of Info Sci	20
Sch of Lib & Info Sci	114	Dept of Communication	74	Urbana[a]	17	Institute of Research Policy Studies	18
Dept of Lib & Info Sci	89	Coll of Business	65	Williamsburg Regional Lib	16	Sch of Management	17
Sch of Info Studies	73	Coll of Business Administration	56	Lib[b]	16	Center for Sci & Tech Studies (CWTS)	14
Sch of Info	51	Dept of Info Management	46	Sch of Info Studies	16	Sch of Comp & Info Tech	14
Coll of Info Sci & Tech	49	Dept of Management	34	Dept of Info Sci	15	Barcelona[b]	14
Dept of Info Management	48	Dept of Management Info System	34	Graduate of Sch Lib & Info Sci	14	Facultat de Biblioteconomia I Documentació	13
Sch of Lib & Info Studies	47	Sauder Sch of Business	29	Sch of Lib & Info Studies	13	Evaluation Office	13

Coll college, *Sch* school, *Dept* department, *Sci* science, *Lib* library, *Info* information, *Tech* technology, *Comp* computer, *Comm* communication

[a] Incomplete department address data in WoS records, but mostly from University of Illinois

[b] Incomplete department address data in WoS records

Table 5 Ten most productive countries in IS&LS from 2005 to 2009

Country	Years					Total	Slope	Rank
	2005	2006	2007	2008	2009			
USA	968	959	974	963	920	4,784	−9.2	99
UK	215	262	269	229	180	1,155	−10.3	100
Spain	46	78	115	141	149	529	26.9	1
Canada	93	116	103	91	108	511	0.5	30
China	51	71	66	103	133	424	19.6	2
Germany	42	53	74	81	78	328	10.0	4
Netherlands	37	40	55	86	86	304	14.4	3
Australia	55	60	65	57	64	301	1.5	24
Taiwan	38	55	49	69	73	284	8.4	5
South Korea	29	36	43	48	62	218	7.8	6

and SM (authors in MIS publish papers in IR and SM journals). However, based on the bibliographic coupling, MIS journals have different intellectual bases (sources for citations) from those in IR and SM (dissimilar in bibliographic coupling).

The above analysis focuses on the identification of subfields in LIS and results in the finding that supports the exclusion of MIS-related journals from LIS for individual research evaluation in the LIS field. In fact, CATAR reports more information in the same analysis useful for other tasks, as described in subsection: "Facet analysis". Specifically, for each actor identified from the publication records, CATAR ranks the actors in various ways based on their occurrence to identify which are the most productive or of most impact. These actors include authors, institutes, countries (based on author's affiliation), journals, cited referenced (CR), cited authors (from CR), and cited journals (also from CR), etc. As an example in an attempt to assess country-level research, Table 5 shows the ten most productive countries where the number of publications for each year and all years, from 2005 to 2009, are listed. The slope column shows the slope of the linear regression line that best fits the publication numbers during these 5 years, indicating the productivity trend of the country. The last column shows the rank of the trend slope among the 106 countries or areas that publish papers in the IS&LS journals. As can be seen, the USA produced far more papers than the other countries, but this trend is declining slightly. Taiwan is ranked No. 9 in terms of total publications and No. 5 in terms of growth rate. Spain has the highest publication growth rate among the 106 contributing countries, partly due to the inclusion of a Spanish journal (with 246 articles) on the verge of the SM cluster: Profesional De La Informacion (PROF INF in Fig. 3).

Table 6 shows the top ten most productive countries during 2005–2009 with their citation impact based on the times cited (TC) field in the WoS records. In this table, NC denotes a normal count of publications for a contributing country. That is, each country was counted once regardless of how many countries contribute to a particular publication. While FC denotes a fractional count, meaning that each country was counted $1/n$ times if n countries contributed to the same publication. Although FC does not reflect true share of contribution, it demonstrates the share of publications over all the publications in the data set for a country. TC in the table is similar to NC, accumulating the times cited m for each country regardless of the number of contributing countries of the publication which received m citations. While FTC denotes fractional count of times cited similar to FC. FTC

Table 6 Ten most productive countries and their average citation impact in IS&LS from 2005 to 2009

Country	NC	TC	CPP	FC	FTC	FCPP
USA	4,784	19,955	4.17	4,476.3	18,323.4	4.09
UK	1,155	4,279	3.70	1,020.5	3,658.9	3.59
Spain	529	1,289	2.44	485.2	1,155.2	2.38
Canada	511	2,470	4.83	427.4	1,965.8	4.60
China	424	1,928	4.55	324.5	1,301.7	4.01
Germany	328	938	2.86	287.7	748.1	2.60
Netherlands	304	1,985	6.53	234.5	1,482.6	6.32
Australia	301	1,249	4.15	247.1	936.6	3.79
Taiwan	284	1,450	5.11	256.7	1,317.3	5.13
South Korea	218	877	4.02	186.1	725.1	3.90

accumulates m/n times for a contributing country if n countries contribute to a publication that received m citations. Citations per publication (CPP) in the table is the ratio of TC over NC, and FCPP (Fractional CPP) is FTC divided by FC, both of which show the average citation impact of a country's publications. As can be seen from Table 6, among the top ten most productive countries, the Netherlands has the highest rank in terms of CPP and FCPP, and Taiwan comes the next.

The above information can be replenished by CATAR's clustering analysis based on JBC as described above. Table 7 shows the most productive countries in each of the first 8 clusters from Set 2 in Table 3. The first column of the second row in each cluster indicates the number of articles and the second column indicates the number of journals in the cluster. The third and fourth rows are the HHI and inverse HHI indexes, respectively, indicating the diversity of the cluster in terms of contributing countries. As Table 7 shows, both AL and CD clusters have low inverse HHI, implying that 1.71 and 1.78 countries on average contribute to these two clusters, respectively. In other words, these two clusters may be of only regional interest. In contrast, the SM cluster has the highest inverse HHI (11.74), indicating that the scientometrics research topic attracts global interest. The IR and RL clusters also have high inverse HHIs (over 7.0). The IR cluster covers important and evolving research topics, and are studied widely by various countries. The RL cluster contains only two journals: *African Journal of Library Archives and Information Science* and *Malaysian Journal of Library & Information Science*. Their editorial bases are distant in regional location; but both journals share much intellectual base such that they are clustered together. This national diversity characteristic of clusters provides additional information for research assessment—clusters with regional characteristic may be omitted for those countries not in the same region to reduce biased assessment.

Discussions

According to Leydesdorff (2008), the JCR subject categories are classified by the Thomson Reuters staff based on a number of criteria, including the journal's title, its citation patterns, etc. However, these classifications match poorly with the classifications derived from the database itself on the basis of analysis of the principal components of the networks generated by citations (Leydesdorff 2006).

Table 7 Cross tabulation of the journal clusters from Set 2 with the most productive countries

1. (IR)		2. (MIS)		3. (AL)		4. (SM)	
3,273 Docs.	15	2,126 Docs.	14	915 Docs.	6	1,040 Docs.	4
HHI	0.13	HHI	0.22	HHI	0.59	HHI	0.09
1/HHI	7.94	1/HHI	4.63	1/HHI	1.71	1/HHI	11.74
USA	1,101	USA	1,113	USA	672	Spain	272
UK	525	UK	253	Canada	41	USA	113
Canada	183	Canada	151	South Africa	18	Belgium	89
China	145	China	95	China	16	China	78
Taiwan	144	Taiwan	90	UK	15	UK	70
Spain	142	South Korea	87	Spain	15	Netherlands	70
Australia	113	Netherlands	85	Turkey	10	Germany	45
Finland	81	Australia	83	Australia	9	Hungary	40
South Korea	79	Germany	56	South Korea	7	Taiwan	30
Netherlands	74	Singapore	53	New Zealand	7	India	30
5. (OA)		6. (ML)		7. (RL)		8. (CD)	
327 Docs.	3	783 Docs.	3	84 Docs.	2	186 Docs.	2
HHI	0.23	HHI	0.39	HHI	0.14	HHI	0.56
1/HHI	4.34	1/HHI	2.59	1/HHI	7.11	1/HHI	1.78
USA	120	USA	508	Nigeria	24	USA	141
UK	91	UK	132	Malaysia	15	China	9
China	16	Canada	49	Botswana	8	Canada	6
France	10	Australia	28	Ghana	7	Australia	5
Australia	10	Netherlands	21	India	5	Uk	4
Germany	9	France	9	South Africa	4	Spain	3
Netherlands	8	Italy	8	UK	3	Turkey	3
Canada	6	Israel	7	Kenya	2	South Korea	2
Finland	5	Switzerland	7	Iran	2	Taiwan	2
Italy	4	Germany	7	Tanzania	2	Pakistan	2

The IS&LS subject category of JCR may not aim at including journals in LIS only. As its category name indicates: it is a subject category containing two closely related fields—IS and LS, the range of which is slightly different from the range of LIS[1] alone. However, since there is no other subject category that matches LIS better, the IS&LS category are now used for research evaluation in the LIS field in Taiwan as a convenient measure, which leads to the problems mentioned above.

In addition to JCR, there are other journal classification and ranking systems available for research evaluation. Examples include those from SCImago (http://www.scimagojr.com/journalrank.php?category=3309), where a list of 128 journals (2011 edition) ranked by various indicators is provided in the LIS category. However, this list also includes those journals in the above MIS cluster, e.g., *Information Systems Journal*, *Information Systems*

[1] As noted on the Wikipedia web site, there is no generally agreed-upon distinction between the terms "library science" (LS) and "library and information science" (LIS) and to a certain extent they are interchangeable, with the later (LIS) being most often used.

Research, *Information Society*, etc. The fact that the journal classification systems provided by JCR and SCImago do not reflect the ideal classification expected by LIS scholars gives rise to the need of journal clustering for subfield delineation in research evaluation and in other scientometric analysis as well. In this regard, scientometric tools like CATAR may serve for this purpose in support of data-driven, evidence-based, and bottoms-up subfield delineation and analysis.

Our clustering analysis based on JBC using CATAR results in around eight subfields from the IS&LS subject category of JCR. Based on the cluster descriptors shown in Table 3, these subfields approximately cover those that are reviewed. The identified MIS journals are often excluded from being used in LIS cognitive mapping studies as shown in the literature review and can hardly be treated as LIS-relevant subfield as recognized by the LIS scholars in Taiwan based on the colleges they serve. Our analysis demonstrates evidence that MIS is relatively distant from the rest in terms of its intellectual base, in addition to the distinct patterns of journal co-citation, journal interlocking, terminology usage, and co-authorship studied by Ni et al. (2012).

Another finding based on the country concentration in clusters indicated by HHI is that scientometrics attracts most global studies. Although only two to four journals are grouped under this cluster, the related topics are also studied in some of the IR journals such that SM journals are mixed with IR journals in the MDS maps.

The existing analysis toolkits, such as CiteSpace and Sci^2 Tool, can also be used for journal clustering and mapping. However, they do not provide further breakdown analyses such as the HHI index computation to further explore the subfield characteristics. Such customization demands us to develop our own tool for convenience use, and in the results we find that some journal clusters attract studies from global countries and some only attract regional attention.

In summary, research evaluation based on WoS data has received more and more attention in Taiwan and in other areas. An ideal delineation of a field is needed for a solid evaluation policy to be implemented, especially when the policy is based on convenient journal classification and ranking. The journal clustering provided by CATAR reveals different facets of cluster characteristics to help better delineate a research field for evaluation. In the existing journal classification systems, there may be journals that have different intellectual base from the others and there may be journals that are of regional interest only. Both cases can be taken into consideration when sampling proper journals from a candidate set to define a research field for less biased research output evaluation.

Conclusions and future work

This paper describes a series of techniques to cluster LIS journals for subfield identification and analysis. Applications of the analysis were exercised and discussed. The significance of such an analysis and the insights that this approach may reveal have been presented. Furthermore, all the analyses have been implemented in a software toolkit called CATAR, which is free to download for re-use and for verification.

This work is different from previous studies in several ways: (1) It uses the full journal list of IS&LS, instead of a part of it, for clustering such that each journal is assigned to a cluster or subfield. This is important for journal ranking based applications. (2) It uses JBC and demonstrates evidence that MIS is distant from the other subfields in terms of its unique intellectual base, in addition to the distinct patterns based on journal co-citation, journal interlocking, terminology usage, and co-authorship as studied by Ni et al. (2012). In addition, JBC allows journals in different languages to be clustered together as long as

they share the same intellectual bases. (3) Around eight subfields are identified by JBC. Based on their cluster descriptors, these subfields approximately cover all those that are surveyed in the literature review, which includes IS (including hard IR, soft IR, and information seeking), LS (practical vs research-oriented), scientometrics (bibliometrics, informetrics, and webometrics), MIS, and other peripheral topics. (4) The use of HHI for country concentration analysis helps identify global and regional characteristics of journal clusters. Clusters with regional characteristic may be omitted for those countries not in the same region to reduce biased research assessment. (5) Like other tools, CATAR streamlines individual data transformation processes and provides diverse bibliometric/scientometric analyses. It could be used recurrently or in parallel with other tools for cross verification in the same task, therefore improving the validity, providing different perspectives, or identifying abnormity of the results.

The software CATAR is originally designed in an ad hoc way, but has later been packed in a toolkit for ease of reuse. The users need only to download data from WoS, save them in the specified folder, and run the corresponding batch command file. CATAR will then generate all the analysis results from the user's data in a result folder. This kind of design aims to relieve the burden of learning to use a new software tool from the users. In addition, the techniques used by CATAR have been described in details, in the hope that an understanding of the analytic steps and the corresponding techniques will enable better use of the tool and reasonable interpretation of the results. Despite this efficiency, the data formats supported in current version of CATAR is limited (only WoS data is readily readable, others require manual format conversion), the volume of data that can be analyzed are limited by the size of main memory, and the analysis results are scattered over several sub-folders and files, which baffles most novice users. Future work is required to continue the improvement of the functionality of CATAR, to make use of existing tools without reinventing the wheels, and to develop an innovative way to help users understand the rich results from their data for fruitful and insightful interpretation.

Acknowledgments This work is supported in part by the "Aim for the Top University Project" of National Taiwan Normal University (NTNU) sponsored by the Ministry of Education, Taiwan, ROC. This work is also supported in part by the National Science Council (NSC) of Taiwan under the Grant NSC 100-2511-S-003-053-MY2. We are grateful to the anonymous reviewers for their valuable comments and suggestions.

References

Ahlgren, P., & Colliander, C. (2009). Document-document similarity approaches and science mapping: Experimental comparison of five approaches. *Journal of Informetrics, 3*(1), 49–63. doi:10.1016/j.joi.2008.11.003.

Ahlgren, P., & Jarneving, B. (2008). Bibliographic coupling, common abstract stems and clustering: A comparison of two document–document similarity approaches in the context of science mapping. *Scientometrics, 76*(2), 273–290. doi:10.1007/s11192-007-1935-1.

Åström, F. (2002). Visualizing library and information science concept spaces through keyword and citation based maps and clusters. In H. Bruce, R. Fidel, P. Ingwersen, & P. Vakkari (Eds.), *The fourth international conference on conceptions of library and information science (CoLIS4) University of Washington, Seattle, WA, July 21–25 2002* (pp. 185–197). Greenwood Village: Libraries Unlimited.

Åström, F. (2007). Changes in the LIS research front: Time-sliced cocitation analyses of LIS journal articles, 1990–2004. *Journal of the American Society for Information Science and Technology, 58*(7), 947–957. doi:10.1002/asi.v58:7.

Börner, K., Huang, W., Linnemeier, M., Duhon, R., Phillips, P., Ma, N., et al. (2010). Rete-netzwerk-red: analyzing and visualizing scholarly networks using the Network Workbench Tool. *Scientometrics, 83*(3), 863–876. doi:10.1007/s11192-009-0149-0.

Börner, K., Chen, C., & Boyack, K. W. Visualizing knowledge domains. In B. Cronin (Ed.), *Annual review of information science & technology (ARIST), 2003* (vol. 37, pp. 179–255). Medford, NJ: Information Today, Inc.

Bush, I. R., Epstein, I., & Sainz, A. (1997). The use of social science sources in social work practice journals: An application of citation analysis. *Social Work Research, 21*(1), 45–56. doi:10.1093/swr/21.1.45.

Buter, R. K., & Noyons, E. C. M. (2001). Improving the functionality of interactive bibliometric science maps. *Scientometrics, 51*(1), 55–68. doi:10.1023/a:1010560527236.

Calkins, S. (1983). The new merger guidelines and the Herfindahl–Hirschman index. *California Law Review, 71*(2), 402–429.

Carpenter, M. P., & Narin, F. (1973). Clustering of scientific journals. *Journal of the American Society for Information Science, 24*(6), 425–436. doi:10.1002/asi.4630240604.

Chang, Y.-W., & Huang, M.-H. (2012). A study of the evolution of interdisciplinarity in library and information science: Using three bibliometric methods. *Journal of the American Society for Information Science and Technology, 63*(1), 22–33. doi:10.1002/asi.21649.

Chen, C. (2006). CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology, 57*(3), 359–377.

Chen, C. M., Ibekwe-SanJuan, F., & Hou, J. H. (2010). The structure and dynamics of cocitation clusters: A multiple-perspective cocitation analysis. *Journal of the American Society for Information Science and Technology, 61*(7), 1386–1409. doi:10.1002/asi.21309.

Cobo, M. J., López-Herrera, A. G., Herrera-Viedma, E., & Herrera, F. (2011). Science Mapping software tools: Review, analysis, and cooperative study among tools. *Journal of the American Society for Information Science and Technology, 62*(7), 1382–1402.

Evans, J. A. (2008). Electronic publication and the narrowing of science and scholarship. *Science, 321*(5887), 395–399. doi:10.1126/science.1150473.

Fernandez-Cano, A., & Bueno, A. (2002). Multivariate evaluation of Spanish educational research journals. *Scientometrics, 55*(1), 87–102. doi:10.1023/a:1016003104436.

Hirschman, A. O. (1964). The paternity of an index. *The American Economic Review, 54*(5), 761.

Janssens, F., Leta, J., Glanzel, W., & De Moor, B. (2006). Towards mapping library and information science. *Information Processing and Management, 42*(6), 1614–1642. doi:10.1016/j.ipm.2006.03.025.

Jarneving, B. (2007). Bibliographic coupling and its application to research-front and other core documents. *Journal of Informetrics, 1*(4), 287–307. doi:10.1016/j.joi.2007.07.004.

Kaufman, L., & Rousseeuw, P. J. (1990). *Finding groups in data: An introduction to cluster analysis*. New York: Wiley.

Kleiweg, P. (2008). Software for dialectometrics and cartography. Retrieved December 31, 2008 from http://www.let.rug.nl/~kleiweg/L04/.

Kruskal, J. B. (1997). Multidimensional scaling and other methods for discovering structure. In K. Enslein, A. Ralston, & H. S. Wilf (Eds.), *Statistical methods for digital computers* (pp. 296–339). New York: Wiley.

Larivière, V., Gingras, Y., & Archambault, É. (2009). The decline in the concentration of citations, 1900–2007. *Journal of the American Society for Information Science and Technology, 60*(4), 858–862. doi:10.1002/asi.v60:4.

Larivière, V., Sugimoto, C. R., & Cronin, B. (2012). A bibliometric chronicling of library and information science's first hundred years. *Journal of the American Society for Information Science and Technology, 63*(5), 997–1016. doi:10.1002/asi.22645.

Leydesdorff, L. (2001). *The challenge of scientometrics: The development, measurement, and self-organization of scientific communications*. Leiden: DSWO.

Leydesdorff, L. (2006). Can scientific journals be classified in terms of aggregated journal–journal citation relations using the journal citation reports. *Journal of the American Society for Information Science and Technology, 57*(5), 601–613.

Leydesdorff, L. (2008). Caveats for the use of citation indicators in research and journal evaluations. *Journal of the American Society for Information Science and Technology, 59*(2), 278–287.

Liston-Heyes, C., & Pilkington, A. (2004). Inventive concentration in the production of green technology: A comparative analysis of fuel cell patents. *Science and Public Policy, 31*(1), 15–25.

McCain, K. W. (1991). Core journal networks and cocitation maps: new bibliometric tools for serials research and management. *Library Quarterly, 61*(3), 311–336.

Milojevic, S., Sugimoto, C. R., Yan, E. J., & Ding, Y. (2011). The cognitive structure of library and information science: Analysis of article title words. *Journal of the American Society for Information Science and Technology, 62*(10), 1933–1953. doi:10.1002/asi.21602.

Moed, H. F. (2005). *Citation analysis in research evaluation*. Dordrecht: Springer.

Moya-Anegon, F., Herrero-Solana, V., & Jimenez-Contreras, E. (2006). A connectionist and multivariate approach to science maps: the SOM, clustering and MDS applied to library science research and information. *Journal of Information Science, 32*(1), 63–77. doi:10.1177/0165551506059226.

Ni, C., & Ding, Y. (2010). Journal clustering through interlocking editorship information. In *Paper presented at the Proceedings of the 73rd ASIS&T Annual Meeting on Navigating Streams in an Information Ecosystem* (vol. 47). Pittsburgh, PA, USA.

Ni, C., & Sugimoto, C. R. (2011). Four-facets study of scholarly communities: Artifact, producer, concept, and gatekeeper. In A. Grove (Ed.), *Annual Meeting of the American Society for Information Science and Technology, New Orleans, Louisiana, USA, October 9–12, 2011* (vol. 48).

Ni, C., Sugimoto, C., & Cronin, B. (2012). Visualizing and comparing four facets of scholarly communication: producers, artifacts, concepts, and gatekeepers. *Scientometrics*, 1–13, doi:10.1007/s11192-012-0849-8.

Nisonger, T. E., & Davis, C. H. (2005). The perception of library and information science journals by LIS education deans and ARL library directors: A replication of the Kohl–Davis study. *College & Research Libraries, 66*(4), 341–377.

Nooy, W. D., Mrvar, A., & Batagelj, V. (2012). *Exploratory social network analysis with Pajek* (structural analysis in the social sciences (no. 34)).

Noyons, E. C. M., Moed, H. F., & Luwel, M. (1999a). Combining mapping and citation analysis for evaluative bibliometric purposes: A bibliometric study. *Journal of the American Society for Information Science, 50*(2), 115–131. doi:10.1002/(sici)1097-4571(1999)50:2<115:aid-asi3>3.0.co;2-j.

Noyons, E. C. M., Moed, H. F., & van Raan, A. F. J. (1999b). Intergrating research performance analysis and science mapping. *Scientometrics, 46*(3), 591–604.

Noyons, E. C. M., & Van Raan, A. F. J. (1998). Advanced mapping of science and technology. *Scientometrics, 41*(1–2), 61–67. doi:10.1007/bf02457967.

Persson, O. (2009). BibExcel. Inforsk, Umeå Univ, Sweden. http://www8.umu.se/inforsk/Bibexcel/. Accessed 25 Oct 2012.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics, 20*, 53–65. doi:10.1016/0377-0427(87)90125-7.

Salton, G. (1989). *Automatic text processing: The transformation, analysis, and retrieval of information by computer*. Reading, MA: Addison-Wesley.

Schildt, H. A., & Mattsson, J. T. (2006). A dense network sub-grouping algorithm for co-citation analysis and its implementation in the software tool Sitkis. *Scientometrics, 67*(1), 143–163. doi:10.1556/Scient.67.2006.1.9.

Sci2 Team. (2009). Science of Science (Sci2) Tool. Indiana University and SciTech Strategies. Retrieved August 12, 2008 from https://sci2.cns.iu.edu.

Simpson, E. H. (1949). Measurement of diversity. *Nature, 163*, 688.

Small, H. G., & Koenig, M. E. D. (1977). Journal clustering using a bibliographic coupling method. *Information Processing and Management, 13*(5), 277–288. doi:10.1016/0306-4573(77)90017-6.

Tseng, Y.-H. (1998). Multilingual keyword extraction for term suggestion. In *21st International ACM SIGIR conference on research and development in information retrieval—SIGIR '98, Australia, Aug. 24–28 1998* (pp. 377–378).

Tseng, Y.-H. (2002). Automatic thesaurus generation for Chinese documents. *Journal of the American Society for Information Science and Technology, 53*(13), 1130–1138.

Tseng, Y.-H. (2010). Generic title labeling for clustered documents. *Expert Systems with Applications, 37*(3), 2247–2254.

Tseng, Y.-H., Lin, Y.-I., Lee, Y.-Y., Hung, W.-C., & Lee, C.-H. (2009). A comparison of methods for detecting hot topics. *Scientometrics, 81*(1), 73–90.

Tseng, Y.-H., Lin, C.-J., & Lin, Y.-I. (2007). Text mining techniques for patent analysis. *Information Processing and Management, 43*(5), 1216–1247.

Van Eck, N. J., & Waltman, L. (2009). VOSviewer. Leiden: Centre for Science and Technology Studies (CWTS) of Leiden University. http://www.vosviewer.com/. Accessed 03 Oct 2012.

Van Eck, N. J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics, 84*(2), 523–538.

van Eck, N. J., & Waltman, L. (2012). Multiple perspectives on bibliometric data: Combining different science mapping approaches using VOSviewer. In *Paper presented at the 2nd Global TechMining conference, Montreal, Quebec, Canada, September 5*.

Van Raan, A. (1997). Scientometrics: State-of-the-art. *Scientometrics, 38*(1), 205–218. doi:10.1007/bf02461131.

van Rijsbergen, C. J. (1979). Information retrieval. Retrieved October 25, 2009 from http://www.dcs.gla.ac.uk/Keith/Chapter.2/Table_2.1.html.

Wall, L., Christiansen, T., & Orwant, J. (2000). *Programming Perl* (3rd ed.). Sebastopol, CA: O'Reilly.

Waltman, L., van Eck, N. J., & Noyons, E. C. M. (2010). A unified approach to mapping and clustering of bibliometric networks. *Journal of Informetrics, 4*(4), 629–635. doi:10.1016/j.joi.2010.07.002.

Waltman, L., Yan, E., & van Eck, N. (2011). A recursive field-normalized bibliometric performance indicator: An application to the field of library and information science. *Scientometrics, 89*(1), 301–314. doi:10.1007/s11192-011-0449-z.

White, H. D., & McCain, K. W. (1997). Visualization of literatures. *Annual Review of Information Systems and Technology (ARIST), 32*, 99–168.

White, H. D., & McCain, K. W. (1998). Visualizing a discipline: An author co-citation analysis of information science, 1972–1995. *Journal of the American Society for Information Science, 49*(4), 327–355. doi:10.1002/(sici)1097-4571(19980401)49:4<327:aid-asi4>3.0.co;2-4.

Yan, E. J., & Ding, Y. (2012). Scholarly network similarities: How bibliographic coupling networks, citation networks, cocitation networks, topical networks, coauthorship networks, and coword networks relate to each other. *Journal of the American Society for Information Science and Technology, 63*(7), 1313–1326. doi:10.1002/asi.22680.

Yang, S., Ma, F., Song, Y., & Qiu, J. (2010). A longitudinal analysis of citation distribution breadth for Chinese scholars. *Scientometrics, 85*(3), 755–765. doi:10.1007/s11192-010-0245-1.

Zhang, L., Janssens, F., Liang, L. M., & Glanzel, W. (2010). Journal cross-citation analysis for validation and improvement of journal-based subject classification in bibliometric research. *Scientometrics, 82*(3), 687–706. doi:10.1007/s11192-010-0180-1.