DTD 5

ARTICLE IN PRESS

LIBCOL-00559; No. of pages: 12; 4C:

Available online at www.sciencedirect.com

SCIENCE DIRECT®

Library Collections, Acquisitions, & Technical Services xx (2005) xxx–xxx

Library Collections, Acquisitions, & Technical Services

How thematic maps can assist collection management: A qualitative assessment of Journals' thematic focus

Fidelia Ibekwe-SanJuan

Department of Information-Communication, University of Lyon 3, 4, cours Albert Thomas, 69008 Lyon, France

Abstract

We present a method for mapping the content of a text collection. This method uses linguistic analysis to relate terms extracted from the texts and clusters them into thematic topics mapped onto a 2D space. While the graphic display of domain topics is useful for several information-driven tasks, the focus of the paper is more on the comparison of journal ranking by productivity (number of published papers in the collection) and by content representativity (ranking by number of terms and clusters). The results show that the two rankings are not identical, thus pointing to possible discrepancies between pure productivity and terminological density.

Keywords: Journal collection management; Content analysis; Data analysis; Thematic maps; Query refinement

1. Introduction

The issue of journal representativity vis-à-vis fields of knowledge is a crucial one for library collection management. Identifying the leading journals in a field and thus the journals to subscribe to has been a constant preoccupation for librarians and information scientists as a whole. This problem was addressed as early as 1934 by the world famous Bradford's law. Bradford found in essence that about 10% of the journals publishing in a field are responsible for producing 90% of the articles in that field. To recover the missing 10% of the articles,

E-mail address: ibekwe@univ-lyon3.fr.

1464-9055/$ - see front matter © 2005 Published by Elsevier Inc.
doi:10.1016/j.lcats.2005.08.004

about 90% of journals are needed. A lot of research has been carried out around modeling Bradford' law to suit different situations. In the same vein, the *Journal Citation Report* (*JCR*) computes impact factors of journals to measure the actual use by scientists of works published by certain journals. Quoting the Institute for Scientific Information (ISI), Giles C.L writes "the impact factor is a measure of the frequency with which the average article in a journal has been cited for a particular year (actually averaged over 2 years) and is calculated dividing the number of citations to articles published in two previous years by the total number of articles published in those years. This produces a normalized parameter so that small and large journals can be compared."

Among the target users of impact factors are librarians who have to "manage and maintain journal collections and budget for subscriptions." The *JCR* covers more than 7500 most highly cited, peer-reviewed journals in approximately 200 disciplines, 3300 editors across 60 countries.

Undoubtedly, tools like *JCR* and Bradford's law are important at the macrolevel for selecting core journals in the disciplines covered by a library and thus for collection management scheme based on journals representativity. However, for content-level analysis of journal representativity per topic (specialties within disciplines), a microlevel and fine-grained approach is needed. Such an approach can actually "enter into" the texts of articles published by journals and map out the core topics. This can be utilized in specialized collection management where identifying core journals is not the issue (they would already have been identified using *JCR* or Bradford's law) but librarians or information scientists actually need to understand from what angle and on what specific topics the subscribed journals make publications on. This could be a further criteria for ranking journal relevance for specific users needs (highly specialized libraries or libraries with different categories of users, needing different levels of expertise).

We propose to this end, a thematic mapping system developed by Ibekwe-SanJuan and SanJuan [10], which takes as input raw texts from a journal collection and returns topic maps represented in a 2D space, which can be used to synthesize the contents of the text collections.

After a review of related works on automatic theme mapping in the Related work section, an overview of the TermWatch system is given in the System overview section. The Mapping domain topics from a collection of IR journals section shows the application of TermWatch to a collection of bibliographic records in the information retrieval field. The Conclusion section explores how the clusters obtained can be mapped onto the source journals of the texts in order to gauge their representativity with regard to the specific topics identified through the clusters. As parts of this research have been published elsewhere [10,11], this paper will focus on a new dimension: possible application of thematic mapping to assist library journal collection management.

2. Related work

Evaluating the state of the art of research in a scientific or technical field has been the object of research since the early sixties. This has led to the emergence of bibliometrics in

1969 then to scientometrics in 1977 and later to informetrics. Today other objects of metrics have appeared: cybermetrics or webometrics. The two major methods used in these studies are the co-citation [17,20] and co-word analyses [3]. Co-citation analysis remains the most popular measure of author–journal contribution to a specific field [18]. After generating a matrix of co-occurrence of citations or of keywords, the underlying methods use a clustering algorithm to reduce the information space and obtain clusters of frequently co-occurring authors, journals, or keywords. These clusters are then mapped onto a 2D space in order to depict their layout and understand the scientific structure of the discipline surveyed. These methods have proved their utility at the macrolevel where entire disciplines are mapped out in order to perceive the social networks and leading actors of the field. They are, however, not targeted to fine-grained content analysis of sub-specialties, thus not very successful at the microlevel. One of the reasons is that clustering being based on occurrences, they need high occurrence thresholds in order to obtain meaningful results.

Clustering techniques are also used in the Information retrieval community (IR) and can be traced to Salton [16], Jardine and Van Rijsbergen [7], and Sparck Jones [19]. The underlying assumption is widely known as the "cluster hypothesis," which postulates that "closely associated documents tend to be relevant to the same requests." The basic approach to clustering in IR consists in partitioning a collection of documents into many small clusters or groups. The intent being later to map user queries to the most similar cluster. This is particularly useful in a context where users do not know a priori which search words to use or do not know the contents nor the indexing vocabularies of the database, as is the case with very large databases or the Internet. Clustering has also been used to address the specific issue of query expansion. Query expansion consists in formulating new query terms using the relevant set of documents. Thus, there is an underlying notion of cluster in this activity: it is hypothesized that the relevant "cluster" of documents "contains terms which can be used to describe a larger cluster of relevant documents [2]." Some techniques like the latent semantic indexing model have been introduced to this end [4]. Another domain in IR, which makes use of clustering, is the presentation of results of a query. Hearst [8] reviewed methods of text categorization or of clustering that enhance the presentation of retrieval results. The aim of these studies is not to explain the layout of research topics but to present groups of "similar" documents in answer to a user's request. To enhance this presentation, considerable interest is being given recently to the use of graphic display interfaces offering 2D or 3D facilities to enable users identify the situation of the relevant documents. Recently, clustering methods are being applied to information search on the Internet [22] and also to gene expression data in the bioinformatics field [21].

The aim of the TermWatch system [10] system is similar to that of co-citation analysis and co-word analyses. However, the thrust here is on text clustering through prior linguistic processing. Consequently, the system builds on recent advances in computational linguistics and particularly on computational terminology to enhance the input to the clustering scheme. Typically, the end user, a domain specialist wishes to know what are the major topics contained in a huge corpus, what topics are evolving and how each topic is related to one another. He or she needs a global view, a map of the domain research topics embodied in the corpus. Additionally, he or she may want to see factual information (authors, laboratories, countries) on each of the topic nodes. The novelty of TermWatch over existing thematic mapping systems is that clustering is based not on co-occurrence of text units but on linguistic relations among them. It focuses on mapping the text content whereas dominant bibliometrics and scientometrics focus on factual data (author co-citation counts, country's or laboratory's publication counts, etc.). In these cases, there is no major difficulty in extracting the units to be counted.

3. System overview

The TermWatch system is a joint research program between two associate professors from two French universities, University of Lyon 3, and University of Metz (LITA). TermWatch has three major components: a term extractor, a linguistic relations miner, and a clustering module.

3.1. Term extraction module

This module extracts terminological units directly from the text collection to be analyzed. The terms extracted reflect the different topics addressed in each text, and thereby the different topics in the whole text collection. Terms should be taken here in their terminological sense (i.e., text units that refer to domain concepts or objects). Our term extraction rules rely on the recent research in the computational terminology field [1]. Most terms appear as noun phrases (NPs) although some verb and prepositional phrases can be terms. We currently extract only terminological NPs, which are multiword expressions that can appear as compounds (information retrieval system) or as syntagmatic NPs with prepositional attachments (special terminology of information science). Term extraction is performed in using the LTPOS tagger developed by the University of Edinburgh. LTPOS is a probabilistic part-of-speech tagger based on Hidden Markov Models. It has been trained on a large corpus and achieves an acceptable performance. It uses the Penn Treebank tag set, which ensures portability of the output with many other systems. Since LTCHUNK, a component of this system only identifies simplex NPs without prepositional attachments, we wrote contextual rules based on the output of the chunker to identify complex terminological NPs.

3.2. Linguistic relations miner

In order to cluster the extracted terms, this module searches for meaningful linguistic relations between them. The idea is that clustering can be performed based on other dimensions than the co-occurrence one. This dimension being linguistic will ensure the semantic coherence of the terms gathered into one cluster. To this end, we studied a variety of linguistic operations, which have come to be known in the terminology community as "variations." Systems aiming to extract domain terms need to address the variation issue in order to capture the actual state of a domain's terminology. This is particularly useful in several applications like acquisition of domain terminology and update, automatic indexing, question answering, information extraction, knowledge representation, and scientific and technological watch. Variations are local morphosyntactic and semantic operations affecting the form and structure of an existing term, thus yielding new terms, which are close to the initial one. Variations cover a wide spectrum of linguistic phenomena occurring at different linguistic levels, thus making their identification impossible without integrating computational linguistics techniques.

At the morphological level, we have spelling variants (specialization/specialization; centre/center; programme/program); inflection variants (academic library/academic libraries) including derivational morphological variants with prefix and suffix addition (tumor promoter/tumor promotion); abbreviations (www/World Wide Web); and compounding process (online Web access/on line Web access/on-line Web access).

Syntactic variants involve structural or formal changes in a term (information retrieval, retrieval of information, efficient retrieval of information), the addition of new modifier or head words in an existing term, that is, syntactic variants of "academic library" found in the corpus are *Canadian academic library privilege*, *changing culture in academic library*, *electronic communication in academic library*, *greater utilization of academic library service*, *Hellenic academic library link*, *service in Malaysian academic library*, *directors of academic library*, *future of academic library*. These relations can be distinguished according to the grammatical function of the word affected: head variation involves the addition or substitution of a new head word in a term as in "academic library" and "directors of academic library" whereas modifier variations implies that only modifier words are affected as in "academic library" and "Canadian academic library." Modifier and head roles are determined by the position of constituent words in a term.

Although morphological and syntactic variants also hold semantic relations, there are explicit semantic variants, which can be realized by surface linguistic markers. For instance, in the following sentences, the sequence "such as" signals a hypernym/hyponym (generic/specific) relation between the NP found on its left (nonlinear systems) and the following one (robotic manipulations). Likewise, the sequence "known as" creates a synonymy relation between "mathematical operation" and "convolution." These relational markers have been studied by Hearst [8], Morin and Jacquemin [15].

(1) The main motivation for this design was to control some known nonlinear systems, such as robotic manipulators, which violate the conventional assumption of the linear PID controller.
(2) This combination is performed by a mathematical operation known as convolution.

Given that all the semantic relations existing between domain terms may not be realized through surface linguistic markers, it is necessary to complete the semantic relation mining using an external resource such as WordNet [5]. WordNet is a general-purpose lexical taxonomy with synonymy, hypernym, and association (see also) relations between words. Synonymous words are gathered into the same "synsets" (classes of words used in the same sense). In our system, WordNet is used to look up word–word relationship between terms when the two terms share common lexical elements but differ by one word. For instance, WordNet, enabled us to establish a relationship between “automatic categorization” and “automatic classification” because “categorization” and “classification” were found in the same WordNet synset. At the moment, morphosyntactic relations and WordNet semantic relations have been implemented in this module.

All the relations mined between terms allow us to build a graph of term variants, which serve as input to the clustering algorithm.

3.3. Clustering module

TermWatch implements a clustering approach, Classification by Preferential Clustered Link (CPCL) presented in Ibekwe-SanJuan [13]. It works in two stages. A first level of clustering consists in grouping together terms sharing the same headword and semantic relations (either given by an external resource like WordNet or harvested through other lexico-syntactic patterns). This results in connected components. For instance the following terms were put into the same component “information department, information science department, Sheffield University’s information department.” The result of the component building stage is a monothematic organization, which is not the desired result. What we seek to highlight is the transversal relation between these lone themes (i.e., what associations have the authors been making between these themes?) To highlight these association, we now cluster the connected components into classes using the second subset of variation relations, those that involve a shift in the head noun, thus a shift in the topical focus of the noun phrase (NP) as in “academic library” and “Canadian academic library privilege.” Like in most clustering methods, we need to compute a similarity index in order to build clusters. This coefficient is defined as follows:

$$d(i,j) = \sum_{R \in \mathrm{CLAS}} \frac{N_R(i,j)}{|R|}$$

where $N_R(i,j)$ denotes the number of R variations between two connected components i and j. A notable difference with other clustering algorithms is that we do not compute this index on the list of terms, but on the set of connected components. The user can set the number of iterations at which the algorithm is stopped and the minimal similarity index to be considered or let the algorithm converge and then choose the results of a given iteration.

The results of the clustering are mapped onto an integrated visualization tool, *Aisee* (http://www.aisee.com). The system architecture is given in Fig. 1.

4. Mapping domain topics from a collection of IR journals

The text collection used in this experiment consists of titles and abstracts extracted from 16 scientific journals publishing articles in the IR and related fields (computer sciences).

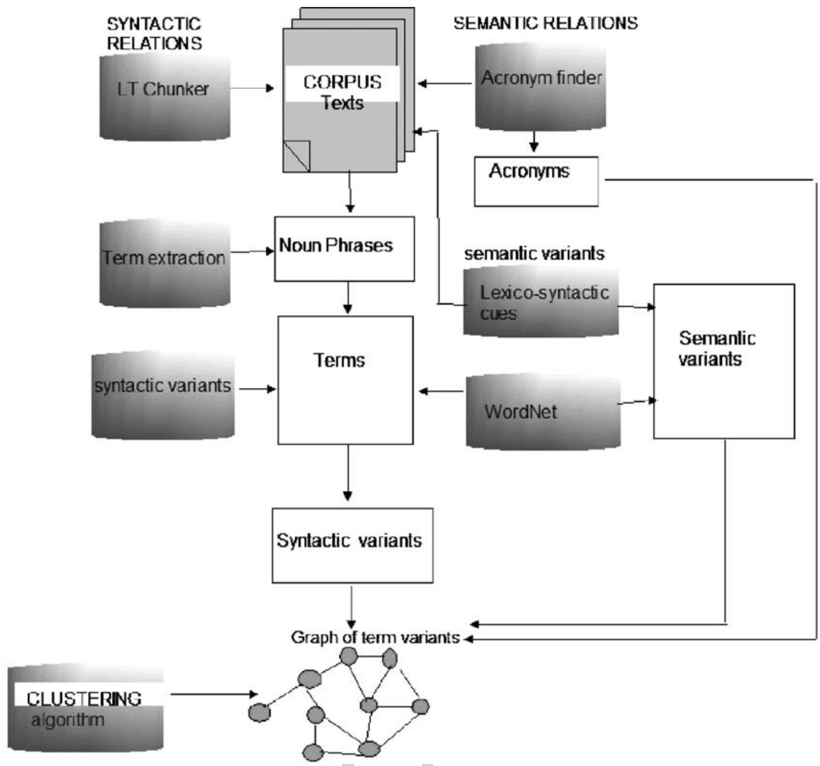

Fig. 1. TermWatch's architecture.

The aim is to map out the research topics addressed in these abstracts over a period of 8 years (1997–2003). The 3,355 titles and abstracts records were thus extracted from the PASCAL multidisciplinary database maintained by the French Institute for Scientific Information (INIST) (http://www.inist.fr). These make up roughly 455,000 words. Although we worked on abstracts rather than on full texts, they were the authors' own texts and shorter texts like abstract are known to be more information dense than full texts. Thus, abstracts represent in our view, adequate surrogates of the full papers. The table below shows the ranking of the journals according to number of bibliographic records. Column 1 is the journal rank, column 2 gives the number of bibliographic records per journal, column 3 the proportion in the entire corpus, column 4 the cumulative, and the last column the journal name (Table 1).

As we can see, the journal that contributed most to the text collection is *Information Sciences*, followed closely by *JASIST*. This is the ranking obtained when using quantitative indicator (number of published papers) as the sole measure of journal representativity vis-à-vis a scientific field. We now look at the fine-grained content analysis of the journals contents as mapped out by TermWatch. We will map the clusters obtained onto the journals to see if the same ranking by productivity is maintained. Table 2 below gives some clustering details obtained from this collection. Because clustering is an iterative process, the user can choose the level of iteration at which to stop the process depending on cluster

Table 1
Collection of 16 journals from the IR and related fields

1	831	25%	831	25%	*Information Sciences*
2	688	21%	1519	45%	*Journal of the American Society for Information Science and Technology*
3	283	8%	1802	54%	*Information Processing and Management*
4	272	8%	2074	62%	*Journal of Information Science*
5	267	8%	2341	70%	*Information Systems Management*
6	175	5%	2516	70%	*Journal of Documentation*
7	176	5%	2692	80%	*Information Systems*
8	116	3%	2808	84%	*Information Systems Security*
9	108	3%	2916	87%	*Library and Information Science Research*
10	108	3%	3024	90%	*Online Information Review*
11	87	3%	3111	93%	*Journal of Internet Cataloging*
12	70	2%	3181	95%	*Information Retrieval and Library Automation*
13	67	2%	3248	97%	*Knowledge Organization*
14	44	1%	3292	98%	*Journal of Information Science and Engineering*
15	34	1%	3326	99%	*International Forum on Information and Documentation*
16	29	1%	3355	100%	*Information Retrieval*
	3355	100%			

granularity (size). In this experiment, we chose the results of the second iteration because classes and their layout seemed meaningful. The 674 classes of variable sizes were thus obtained containing a total of 5632 terms.

This clustering output is an improved version of the one already carried out on the IRcorpus and published in Ibekwe-SanJuan and SanJuan [10,11]. In this experiment, we refined the definitions of the variation relations (cf. Section 3.2) and changed their roles during clustering.

4.1. Graphic display of collection thematics

We show here below the output of the system viewed through a graphic display package, Aisee. Aisee interprets clusters built by TermWatch and aligns them according to their centrality (number of outgoing links). Thus, in the cropped image below (we only show the

Table 2
Details of the clustering

Clusters obtained from the IR corpus	
Number of iterations	1
Number of components	1595
Number of clusters	674
Size biggest cluster	135
Size smallest cluster	4
Total terms in clusters	5632

central part of the image), the most central topic is "information retrieval" this is not surprising because the collection is built on this topic.

However, the selection of the texts was not based on keywords but on journal titles. There was therefore no guarantee that "information retrieval" will be found as the most active term with many linguistic relations (variants) in the corpus. It could have been considered as a "meta term" by authors and as such, not used in their abstracts because these journals were more or less about information retrieval. Surprisingly, this turned out not to be the case. The fact that this term actually appeared with a lot of variants shows that researchers actually use the macroterm together with more specific qualifiers to refer to their works or to applications of their studies. Unfolding a cluster shows the most active term variants. Unfolding the "information retrieval" cluster showed that it dealt with objects and methods of information retrieval systems, hence the presence of variants like "content-based image retrieval systems, NLP information retrieval systems, bibliographic retrieval systems, modern text retrieval systems, natural language information retrieval systems, online information retrieval systems..." Thus, the label is the most generic term while the cluster contents point to more specific and current research concerns.

Surrounding this most central clusters are other clusters like "semantic similarity measure," which deals with different similarity measures used in information retrieval like Cosine, Jaccard, angle-based similarity measures, collocation-based similarity measure, distance similarity measure, etc. The cluster labeled "vector space" refers to research on vector space model of information retrieval. The cluster "wide web sites" deals with different types of Web sites (academic, commercial, etc.). "Natural language" cluster portrays research on natural language query processing. The cluster "online information sources" concerns studies dealing with different online resources as shown by variants like "electronic consumer health information, Web information sources, commercially produced online information sources, distributed information sources, sources of bibliographic information..." "Online catalog" contained variants like "Web on-line catalog, operational online catalogs, commercially available Web browsers, needs of online catalog users, on-line catalog searching, next generation of online catalogs, next generation of retrieval systems" showing clearly the theme reflected by the cluster (Fig. 2).

The topographic layout of clusters offered by TermWatch is useful for grasping rapidly the contents of a large text collection. This is particularly important for science and technology watch, that is, understanding the interactions between domain topics and following their evolution through time stamps [10] but also for query refinement.

4.2. Ranking journals by term and cluster representativity

The focus of this paper is to determine how the clusters of domain topics mapped TermWatch can assist library collection management. Hence, we will seek to ascertain how the 16 journals are distributed across the 674 clusters by assigning journals to clusters in which they have the highest number of terms. Assuming that the most productive journals as shown in Table 1 will also be the most productive in terms of "terms" and "variation relations," we should obtain the same ranking with respect to a

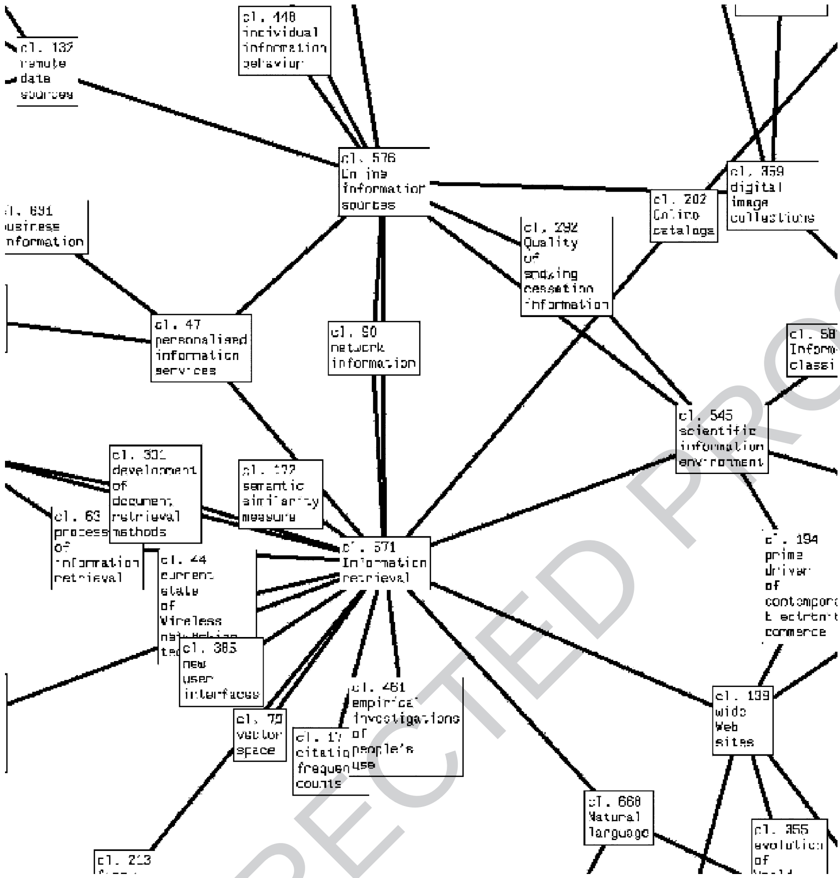

Fig. 2. Part of the thematic map produced by TermWatch on the IRcorpus.

given cluster. We ranked the journals by number of terms they contained, then by number of clusters (Table 3).

As we can observe from the table below, journal representativity by number of terms is roughly correlated with their representativity by number of clusters (except for two positions, 5th and 7th). However, comparison with the journal ranking by number of articles (Table 1) shows some discrepancies. *JASIST* turns out to be the most productive in terms of domain terms and variants whereas it was 2nd by number of articles. Conversely, *Information Sciences* now comes 2nd by representativity in clusters. *Information Processing and Management*, *Journal of Information Science*, *Journal of Documentation*, and *Journal of Information Science and Engineering* maintain their respective positions in the two rankings. On the other hand, “information systems, library and information science research, online information review,” and “knowledge organization” gain two places by arriving at the 5th, 7th, 8th, and 11th positions, respectively, by number of term variants. *International Forum on Information and Documentation* and *Information Retrieval* also gain three places by arriving at the 12th and 13th positions, respectively. *Information Systems Management*, *Information Systems Security*,

Table 3
Journal ranking by number of terms in clusters

Rank	Journal	Number of clusters	Number of terms
1	*Journal of the American Society for Information Science and Technology*	468	3616
2	*Information Sciences*	382	3115
3	*Information Processing and Management*	304	1582
4	*Journal of Information Science*	252	1067
5	*Information Systems*	219	997
6	*Journal of Documentation*	249	899
7	*Library and Information Science Research*	140	517
8	*Online Information Review*	153	488
9	*Information Systems Management*	121	438
10	*Journal of Internet Cataloging*	90	422
11	*Knowledge Organization*	85	227
12	*International Forum on Information and Documentation*	75	164
13	*Information Retrieval*	69	161
14	*Journal of Information Science and Engineering*	58	122
15	*Information Systems Security*	29	83
16	*Information Retrieval and Library Automation*	25	45

and *Information Retrieval and Library Automation* descend by four, five, and four places, respectively. On the whole, seven out of the 16 journals showed consistency in the two rankings while nine journals showed notable differences.

5. Conclusion

We have presented in this paper, an alternative to the journal collection management problem. This could be through thematic mapping using linguistic and data analysis techniques. The proposed approach, embodied in the TermWatch system enables a librarian to grasp more readily the contents of a collection of journal through an in-depth analysis of their texts. The resulting maps can be used for positioning research topics vis-à-vis one another and contribute also to answering specific search needs of certain categories of users. The journal ranking by thematic content also portrays differences with ranking by pure numerical factor (i.e., journal productivity). This finding suggests that while some journals may publish a considerable amount of papers in a given field, this number may not necessarily be correlated with density of domain terms.

6. Uncited references

[6]
[9]
[12]
[14]

References

[1] D. Bourigault, C. Jacquemin, & M. -C. L'Homme, (Eds.). (2001). *Recent advances in computational terminology, vol. 2*. Amsterdam: John Benjamins.

[2] Baeza-Yates, & Ribeiro-Neto, B. (1999). Query operations. *Modern information retrieval* (pp. 117–139). ACM Press.

[3] Callon, M., Law, J., & Rip, A. (1986). *Mapping the dynamics of science and technology*. Basingstoke: MacMillian Press.

[4] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, R., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science, 41*6), 391–407.

[5] Fellbaum, C. (1998). *WordNet, an electronic lexical database*. MIT Press.

[6] Giles, C. L. *Citation index: Journal impact factors*. Retrieved 23/02/2005 from http://www.neci.nj.nec.com/homepages/giles/html/cites.html

[7] Jardine, N., & Van Rijsbergen, C. J. (1971). The use of hierarchic clustering in information retrieval. *Information Storage and Retrieval, 7,* 217–240.

[8] Hearst, M. (1999). The use of categories and clusters for organizing retrieval results. In T. Strzalkowski (Ed.), *Natural language information retrieval. Text, Speech and Language Technology, vol. 7* (pp. 333–374). Kluwer Academic Press.

[9] Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. *Proceedings of the International conference on Computational Linguistics (COLING'92), Nantes,* 539–545.

[10] Ibekwe-SanJuan, F., & SanJuan, E. (2004). Mining textual data through term variant clustering: The Termwatch system. *Proceedings of the International conference on* "*Recherche d'Information assistée par ordinateur. Avignon*" (pp. 487–503).

[11] Ibekwe-SanJuan, F., & SanJuan, E. (2004). Mining for knowledge chunks in a terminology network. *8th International ISKO conference, University College London, 13–16 July 2004* (pp. 41–47).

[12] Ibekwe-SanJuan, F., & SanJuan, E. (2002). *From term variants to research topics*. *International Journal on Knowledge Organization (ISKO), special issue on Human Language Technology, 29*(3/4), 181–197.

[13] Ibekwe-SanJuan, F. (1998). A linguistic and mathematical method for mapping thematic trends from texts. *Proceedings of the 13th European Conference on Artificial Intelligence, Brighton UK, 23–28 August* (pp. 170–174).

[14] Jacquemin, C., & Bourigault, D. (2003). *Term extraction and automatic indexing*. In R. Mitkov (Ed.), *Handbook of computational linguistics* (pp. 599–615). Oxford University Press.

[15] Morin, E., & Jacquemin, C. (2004). Automatic acquisition and expansion of hypernym links. *Computer and the Humanities, 38*(4), 343–362.

[16] Salton, G. (1968). *Automatic information organization and retrieval*. New York: McGraw-Hill (18 pp.).

[17] Small, H. (1999). Visualizing science by citation mapping. *Journal of the American Society for Information Science, 50*(9), 799–813.

[18] Schiffrin, R., & Börner, K. (2004). *Mapping knowledge domains*. *Publication of the National Academy of Science (PNAS), 101*(1), 5183–5185.

[19] Sparck Jones, K. (1971). *Automatic keyword classification for information retrieval*. London: Butterworths.

[20] White, H. D., & Mccain, K. W. (1989). *Bibliometrics*. In M. E. Williams (Ed.), *Annual review of information science and technology* (pp. 119–186). New York: Elsevier Science.

[21] Yeung, K. Y., Haynor, H., & Ruzzo, W. L. (2001). Validating clustering for gene expression data. *Bioinformatics, 17,* 309–318.

[22] Zamir, O., & Etzioni, O. (1998). Web document clustering, a feasibility demonstration. *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 46–54).