Information Science in the web era: a term-based approach to domain mapping

Authors

Fidelia Ibekwe-SanJuan
ELICO, University of Lyon 3
4, cours Albert Thomas, 69008, Lyon Cedex
Email: fidelia.ibekwe-sanjuan@univ-lyon3.fr

We propose a methodology for mapping the research in Information Science (IS) field based on a combined use of symbolic (linguistic) and numeric information. Using the same list of 12 IS journals as in earlier studies on this same topic (White & McCain 1998 ; Zhao & Strotmann 2008a&b), we mapped the structure of research in IS for two consecutive periods: 1996-2005 and 2006-2008. We focused on mapping the content of scientific publications from the title and abstract fields of underlying publications. The labels of clusters were automatically derived from titles and abstracts of scientific publications based on linguistic criteria. The results showed that while Information Retrieval (IR) and Citation studies continued to be the two structuring poles of research in IS, other prominent poles have emerged: webometrics in the first period (1996-2005) evolved into general web studies in the second period, integrating more aspects of IR research. Hence web studies and IR are more interwoven. There is still persistence of user studies in IS but now dispersed among the web studies and the IR poles. The presence of some recent trends in IR research such as automatic summarization and the use of language models were also highlighted by our method. Theoretic research on "information science" continue to occupy a smaller but persistence place. Citation studies on the other hand remains a monolithic block, isolated from the two other poles (IR and web studies) save for a tenuous link through user studies. Citation studies have also recently evolved internally to accommodate newcomers like "h-index, Google scholar and the open access model". All these results were automatically generated by our method without resorting to manual labeling of specialties nor reading the publication titles. Our results show that mapping domain knowledge structures at the term level offers a more detailed and intuitive picture of the field as well as capturing emerging trends.

1. Introduction

The need to map scientific domains has long been established since the development of bibliometrics (Small 1999, Garfield 1979, White & McCain 1998). As more efficient and powerful clustering algorithms are developed and research in the information visualization progress, systems combine methodology and tools from these fields in order to visualize bigger networks of scientific data. Researchers in knowledge domain mapping have fruitfully combined tools from several domains to assist this process (Börner & Schiffrin 2004, Chen 2006, Boyack & Klavans 2007).

Author co-citation analysis (ACA) is an established method for mapping the intellectual base of a research field as reflected by the use of citations in published works. However, what is usually mapped by ACA are the co-cited authors, i.e; those whose works influenced the current published works. In the field of Information Science (henceforth IS), a series of ACA studies have been carried out on the most prominent co-cited authors. Amongst the most well known is the landmark study by White & McCain 1998. This study mapped the 120 most co-cited authors in IS field over a period of 24 years (1972-1995). Zhao & Strotmann (2008a) carried out a follow-up study by performing an ACA for the period 1996-2005. In both studies, the authors used the same 12 IS journals as in White & McCain (1998) and also limited their analysis to the first 120 most co-cited authors. Astrom (2007) performed a co-citation analysis of cited documents that significantly influenced the Library and Information Science (LIS) research. He started from a much larger selection of source journals (21 journals) than in White & McCain (1998) or Zhao & Stotmann (2008a,b) and he also covered a different period (1990-2004).

In the above three studies, what is mapped is the research carried out decades before the ones from which the co-cited data was retrieved. Hence, the results, while giving precious information on the intellectual base of the field, do not inform the reader on the current research topics or research authors. As a remedy to this gap, two studies focusing on current research but from different angles have been carried out. In a follow-up to their first paper, Zhao & Strotmann (2008b) carried out a co-citation analysis of the citing authors. The authors called this "author bibliographic-coupling analysis" (ABCA). The original method known as "bibliographic coupling" was introduced by Kessler in 1963. Zhao & Strotmann (2008b) adapted it to co-citing authors. Their aim was to map out current research dynamics as shown by current citing authors as opposed to mapping the older intellectual base by ACA. In this second study, the authors compared the two sides of the coin: active authors as obtained by ABCA and past authors of intellectual base as shown by ACA. They also covered the same period as in their previous study,

1996-2005 but split it here into two five-year periods: 1996-2000 and 2001-2005.

In ACA studies, the object of the mapping is the author names, not the publication contents. However, authors performing ACA or ABCA have found it necessary to go beyond the map of author names to understand what is going on in a field. Thus, it is customary in ACA studies to define cluster labels of research specialties and provide a narrative of what is going on in the field. These labels and narratives are a result of human interpretation. White & McCain (1998) had to label the 12 factors obtained from factor analysis as research specialties and provide an analysis of what each cited author represented in terms of research areas. Although this analysis was highly accurate and lent life to the maps of co-cited authors, it is reliant on the level of knowledge the authors possessed of the IS field. Likewise in Zhao and Strotmann (2008a&b), the authors had to examine the co-cited and co-citing authors' publications in order to manually label the factors. By this same token, some factors could not be labeled and were presented as "undefined" because they had too few authors or because their subject matter could not be determined.

Until recently, the IS field has mostly been studied via co-citation analysis. Research in knowledge domain mapping has made little or no use of recent advances in natural language processing (NLP) and of text mining to tackle the issue of mapping the field through the text contents of bibliographic records of publications or via the full texts. The majority of studies in this area have focused exclusively on factual bibliographic units (authors, documents, journals) or on keywords already furnished from a controlled vocabulary. Bibliometrics has had a history of mapping domains through keywords. The co-word analysis, developed by Callon, Courtial, Turner, and Bauin (1983) in the late eighties symbolizes this approach. However, the keywords have been criticized for their inertia and the co-word method for its apparent "inability" to represent domain knowledge structures (Leydesdorff 1997). The use of co-word analysis as it was presented by its authors therefore diminished in bibliometrics. Lately, some attempts have been made to add data from textual fields to ACA. Chen (2006) extracted n-grams from titles and abstracts fields using an existing phrase extractor. These noun phrases were mapped alongside co-cited authors to complement ACA results with phrases, thus rendering the results more intuitive.

Janssens, Leta, Glänzel, and De Moor (2006) carried out a quantitative linguistics approach to map the research specialties in the IS field over a three year period (2002-2004). Their data consisted of 938 full texts from five IS journals. This is a much smaller selection of journals than is usual in previous IS mappings but can be justified by the fact that the authors chose to work from full texts. This inevitably leads to a higher representation space than in ACA. However, in

their study, the labeling of clusters was done manually, based on inspection of the stemmed terms in each cluster. Although Janssens et al. (2006) worked from full texts , there are significant methodological differences between their approach and ours. Unlike in their study where most of the techniques used already existed (Porter's stemmer, LTChunker, Multi-Dimensional Scaling, Latent Semantic Indexing and the clustering algorithms have been around for some decades), we developed our own text mining platform in order to better adapt the natural language processing (NLP) component to text features. Also, we preferred to use titles and abstracts rather than from full texts as this reduces a considerable amount of noise found in full texts. Abstracts and titles are condensed forms representing the most salient contribution of a publication. We also do not stem words nor limit the noun phrases to any arbitrary length. Rather we extract them as they appear in the texts since this reflects the state of the field's terminology at that point in time. Owing to this and to the absence of stemming, our term frequencies are predictably lower and thus the frequency thresholds for selecting input terms for the analysis. This enables the method to eventually detect lesser known topics which can signal emerging trends.

In our study, we focus on mapping current research topics in IS as reflected by terms in titles and abstracts of published works. This is a different perspective from ACA which depicts past themes and the intellectual base of IS. While Zhao & Strotmann (2008b) in their ABCA study focused on "active authors", we focus on the "active topics" in IS, on the same period and after. Thus our study can be perceived as the thematic counterpart of theirs. The two studies would be complementary in providing a more complete view of the field.

2. Data collection

For ease of reference to previous studies on mapping the IS field, we used the same list of 12 journals as in White & McCain (1998), then Zhao & Strotmann (2008a,b). We split the time period such that the first period corresponded to the one studied in Zhao & Stromann (1996-2005). The second period (2006-2008) is an update of their study. The data collection was carried out on the 27th december 2008 from the ISI – Web of Science database.

Bibliographic references from these 12 journals were downloaded with the option "Full refs". No citation data was included as this was not the focus of our study. We recall the list of 12 journals in table 1. As we can see, this list does not include library journals per se and only four journals on library automation. In contrast, Astrom's (2007) chose a much broader selection of 21 library & information science (LIS) journals. Hence, the mappings obtained using data from these 12

journals will be constantly referred to as being of the IS field rather than of the broader LIS field.

TABLE 1. The 12 IS journals. Same list as in White & McCain (1998) and Zhao & Strotmann (2008a, b).

Information science	Library automation
1. Annual Review of Information Science and Technology 2. Information Processing & Management 3. Journal of the American Society for Information Science and Technology 4. Journal of Documentation 5. Journal of Information Science 6. Proceedings of the ASIST Annual Meeting 7. Library & Information Science Research 8. Scientometrics	1. Electronic Library 2. Information Technology and Libraries (and Journal of Library Automation) 3. Library Resources & Technical Services 4. Program—Automated Library and Information Systems

For the period 1996-2005, 6418 bibliographic records were obtained. Upon comparison of several fields in our dataset (titles, abstracts, authors and UT numbers), we found hundreds of duplicates which were removed, yielding finally 5535 records with titles and/or abstracts fields for this period which will be the input to our system. For the second period (2006-2008), we obtained 2438 records of which 1938 remained after duplicates removal. From these records, our system extracts the titles and abstracts which will be subjected to further processing. We also additionally extracted the keywords and authors fields as way to compare the maps obtained from the title and abstracts fields with other fields.

3. Methodology

Our methodology stems from a multi-disciplinary approach to text mining. It integrates state-of-the-art techniques from Natural Language Processing (NLP) and more specifically computational terminology, Clustering and Graph Theory. This methodology has been implemented in the TermWatch platform. Different stages of it have been described in earlier publications (SanJuan & Ibekwe-SanJuan, 2006). TermWatch relies on surface linguistic relations between multi-word terms (MWTs) to build semantically coherent clusters of terms. The process leading from the input of raw texts to the mapping of domain topics can be broken down into five major stages:

1. Multi-word term extraction and feature selection
2. Term variants identification
3. Term clustering by linguistic relations
4. Generating association graphs of clusters by co-occurrence information

5. Mapping and visualization of topics.

For visualization of the term networks produced by TermWatch, we use an external visualization package for which we adapt the system's output. In the current study, we used Pajek's implementation of the Kamada-kawai algorithm (Batagelj & Mryar, 2009).

3.1 Linguistic processing of texts

This consists of two components: term extraction and feature selection; term variant identification.

3.1.1 Multiword term extraction and feature selection

The corpus of titles and abstracts was tagged using TreeTagger (Schmid, 1999) in order to obtain parts-of-speech (POS) information for every word. We wrote a few contextual rules to extract multi-word terms based on their morphological and syntactic properties. The extracted terms can be simplex noun phrases (NPs) like "information science" or complex ones like "library information science abstracts " which embed simpler NPs. No limit is imposed on the length of the extracted terms thus ensuring that new terms coined by authors of papers are extracted 'as is' and that existing domain concepts with multi-words are not altered or split. Extracted terms are subjected to a term selection function which eliminates the most unlikely candidates. This function computes the geometric mean G(t) of the inertia induced by two tf.idf like functions. One function is based on the whole term occurrence, the other is based on the occurrence of component words using MySql's match function and document length normalization1. Candidate terms are thus ranked according to the harmonic mean of the score from these two functions. From an initial list of 51 000 candidates, this function retained 8029 terms.

3.1.2 Term variant identification

After term selection, a semantic variant identifier searches for relations amongst the set of terms. These relations are lexico-syntactic variations which affect the form and the structure of terms. By lexico-syntactic variations, we refer mainly to two linguistic operations: lexical inclusion and lexical substitution. By lexical inclusion, we refer to the case where a shorter term is

embedded in a longer one through insertions or additions of modifier or head words as in "ISI impact factor / ISI journal impact factor". Lexical inclusion reflects hierarchical relations between a generic term (hypernym) and its more specific variant (hyponym). Spelling variants (rank-frequency distribution / rank frequency distribution) and synonyms are acquired by consulting WordNet (Fellbaum 1998). However this process is not straightforward as WordNet only has synonymous relations between lone words. We implemented rules to expand one word synset relations to multiword terms by stipulating that this relation obtained only if the synonymous words occupied the same grammatical function in the two terms considered.

The linguistic theory behind the grouping of terms either by shared modifiers or by shared head is known as distributional analysis. It was originally introduced by Harris (1968) and later taken up by various authors working on automatic thesaurus construction (Grefenstette 1994) or on terminology engineering (Jacquemin 2001). We extended the definition of the types of relations identified and added additional constraints like the position of added words and restricted the number of substituted words to 1 to avoid generating spurious variants. The details of the linguistic rules by which these semantic variants are identified can be found in (Ibekwe-SanJuan 1998). We simply point out at this stage that the entire process leading from corpus tagging to term variant identification is fully automated.

3.2 Term Clustering

TermWatch offers several possibilities for choosing relations used for clustering. The variation relations identified amongst terms in the preceding stage are used to form a first level of semantically motivated clusters. In a second stage, an association graph (or co-occurrence matrix) is built from cluster labels appearing in the same documents. Two terms (representing cluster labels) are said to co-occur if they or one of their semantic variants appeared in the same document. In this way, at the first level, we generate semantically coherent clusters and then aggregate them using co-occurrence information. We describe this two-tier process in more details below.

3.2.1 Clustering by linguistic relations

We designed an agglomerative hierarchical algorithm called CPCL which

considers any type of relation between two items. The clustering module starts by building connected components using a subset of the variation relations called COMP which correspond to synonyms and spelling variants. A component is thus a group of terms that typically depict the same concept but in various forms (variants).

In a second stage, components are clustered based on the second subset of relations called CLAS that modify the concept (the term's focus). These relations typically depict associations between related concepts (head expansion, head substitution).

Clustering relies on a measure of the strength of the links between two components, which is a proportion of the type and number of variations between them compared to the total number of that type of variation in the whole graph. We represent this measure by a dissimilarity index d(i,j) computed as follows:

$$d_{ij} = \sum_{R \in CLAS} \frac{n_R(i, j)}{|R|}$$

Equation 1. Edge coefficient

where N_R (i,j) is the number of variations of type $_R$ in CLAS relations between components i and i. |R| is the total number of variations of that type in the whole graph. Thus the clustering is not done at the item level (single terms) but at the level of groups of terms (components) formed based on their semantic similarity. Clusters are labeled automatically by the algorithm as the term with the highest number of semantic variants within the cluster.

3.2.2 Generating association graphs through co-occurrence of cluster labels

The linguistic clusters obtained in the preceding stage represent homogeneous topics. We need to study the way in which these clusters are associated to documents. The idea is to utilize co-occurrence information to measure their associations in order to map out the research specialties in the field. The labels

of the clusters obtained in the preceding stage form the input to this stage. Two clusters are said to be associated if their labels or one of their semantic variants co-occurred in the same document. In this way, we not only look at the co-occurrence of the labels but also that of the cluster contents. In order to measure the association between two clusters, we need to compute an association index. We experimented with four commonly used association measures: mutual Information index (MI), log llikelihood test, chi-2 and the equivalence index (Eij). This last one was the same principle used in co-word analysis (Callon et al., 1983). The equivalence index produced a better ranking than the other three measures. MI and log likelihood produced similar rankings when looking at the first 50 pairs of clusters. The chi-2 test produced the worst rankings by putting at the topmost positions pairs of clusters that did not reflect the best related topics in the field.

Once an association index has been computed for each pair of clusters, thresholds can be fixed such that the corresponding map shows only clusters whose associations are above the given threshold. Considering the fact that text units have notoriously low frequencies, these thresholds cannot be as high as the ones usually fixed in ACA. We experimented with different thresholds and found that best results were obtained with a raw co-occurrence of >2 and an Eij >=0.01. In the next section, we analyze the results obtained in the two periods of our corpus on the IS field.

To summarize the approach to clustering implemented in TermWatch: the rationale is to first use linguistic relations to obtain good quality groupings (the linguistic clusters), then connect these clusters through co-occurrence information and measure their association index to yield the final topics. The end results are clusters of high semantic homogeneity which also capture the most salient association links. This way of building clusters by first grouping semantic variants of the same terms, then by gradually connecting them based on their co-occurrences in the same documents is unique to the best of our knowledge.

4. Structure of research IS between 1996-2008

As our study covers the same period and the same source journals as in Zhao & Strotmann's (2008b), it is with respect to their ABCA mappings (citing authors) that we can make some loose

comparisons, pointing out similarities and differences when applicable.

4.1 Period I: 1996-2005

The resulting map for this period is shown in Figure 1. The size of a node reflects the importance of the topic in terms of its size (number of terms) and its degree (number of links to other nodes). The colour of a node has no particular significance. 109 clusters were obtained for this period. With a co-occurrence threshold of >2 and an Eij >=0.01, we selected the most significant links between clusters. The field of IS is structured around three big research poles: automated information retrieval (IR); web studies and co-citation studies. Apart from the prominence of "web studies", the two other poles correspond to the already observed "two-camp structure" of IS (White & McCain 1998). We analyze in more details each pole and how they relate to one another.

4.1.1 Automated Information Retrieval (IR)

The automated IR cluster is at the core of a research dynamic with several sub-specialties: user studies (user profile), document collection, knowledge creation & management and online search process. Smaller clusters surrounding the IR cluster reflect well known IR themes such as "IR system, structured query, term frequency, vector space, text retrieval conference (TREC), average precision figure". The "user profile" cluster refers to relevance feedback studies as it pertains to the use of automated IR systems (document ranking) and not to cognitive user-oriented studies or interactive user studies. Zhao & Strotmann (2008b) made a similar observation about user studies from their ABCA maps, which according to them appeared "to be more about users' interaction with information retrieval systems than about user information behavior in general".

The IR specialty is connected to the web studies pole via an intermediary zone with middle-sized labeled "search intermediary postsearch questionnaire, pre postsearch interview, search result". These clusters reflect research on user evaluation of online search systems.

4.1.2 Web-based studies

This pole is organized around two prominent clusters labeled "web-based" and

"world wide web" located in the north west part of the map. Web-based studies occupy as much place as IR on the map, suggesting that this specialty has become as prominent as research on IR systems. Surrounding the two web-based clusters are clusters labeled "UK academic web, different topological web graph structure, relevant web page, institutional research, web search session, social network theory, major internet search service, electronic book". The web studies area portrays current research on webometrics. This group of clusters corresponds roughly to "scholarly communication and web'' and later to "webometrics" on the ABCA maps of Zhao & Strotmann (2008b).

Above the web-based cluster is a relatively prominent cluster labeled "originality value". Upon consulting documents where this term occurred, we found out that this cluster is really about recovery or preservation of lost or damaged archives during the two wars (WWI and WW2). Hence the presence of clusters labeled "online public access catalogue (OPAC), library and information science" around it. This area also reflects the concerns about the integration of e-books into digital libraries. Just as web studies and IR share common themes, so do research on digital libraries and web studies.

Figure 1. Map of research topics in IS for the period 1996-2005. 'Term x term' association graph.
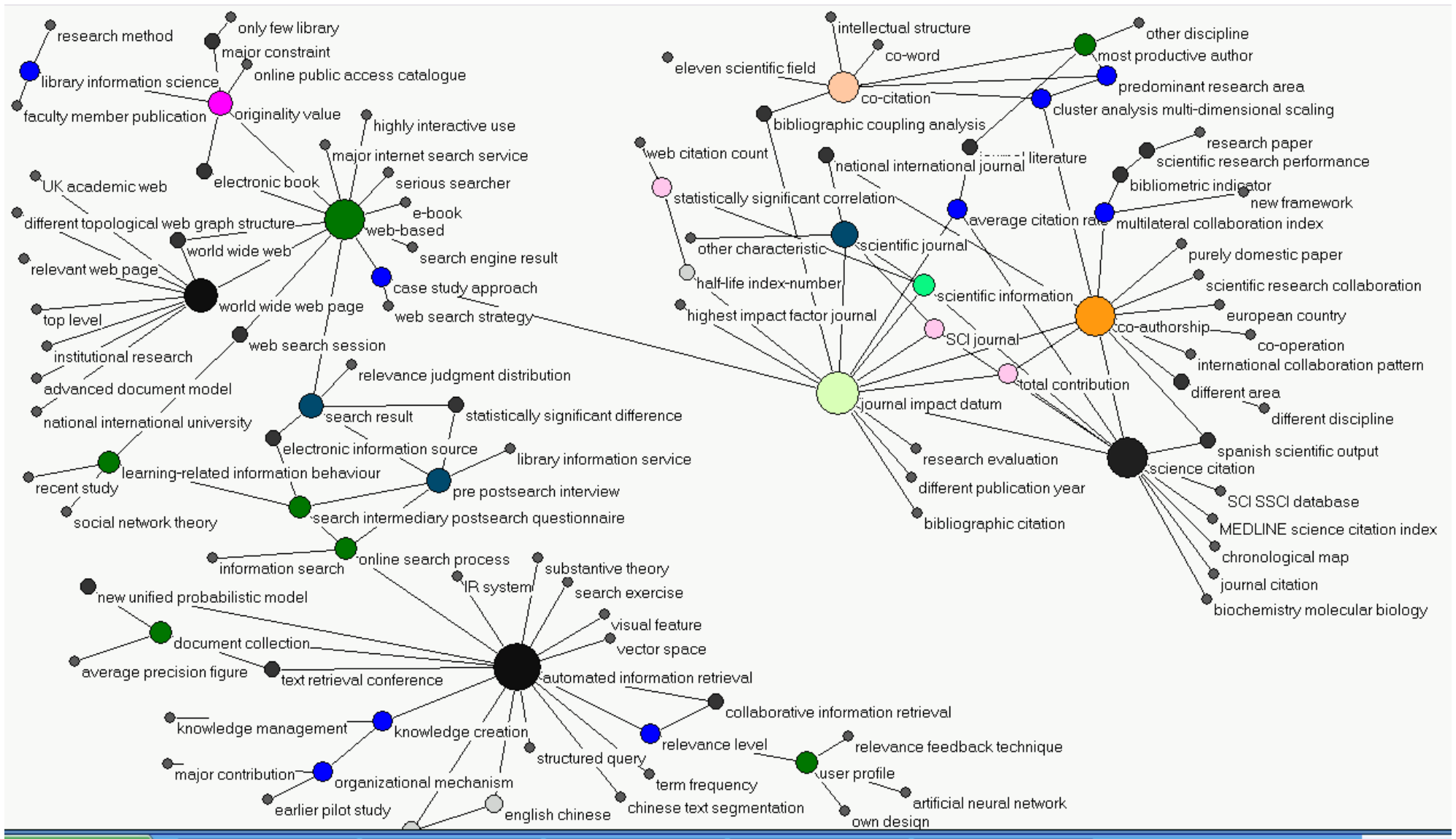
4.1.3 Citation studies

This the second biggest pole of research in IS. In this area of the map (east side), we find four big clusters labeled "journal impact datum, science citation, co-authorship, co-citation". Journal impact datum reflects research on Journal Impact Factor as evidenced by the surrounding clusters "highest impact factor journal, half-life index-number, scientific journal, SCI journal, scientific information, statistically significant correlation, web citation count, bibliographic citation, research evaluation". The big "science citation" cluster is surrounded by smaller clusters on the different citation databases "SCI SSCI database, Medline science citation index".

Both "journal impact datum" and "science citation" clusters share links with the "co-authorship" cluster. Surrounding clusters deal with different research issues on evaluating collaboration networks: "international collaboration pattern, co-operation, european country, scientific research collaboration, purely domestic paper, multilateral collaboration index, bibliometric indicator, scientific research performance". The co-authorship cluster is clearly on using bibliometric indicators to evaluate research and individuals both at national and international levels. Here again, we find a correlation with an observation made in Zhao & Strotmann (2008b) from their ABCA map and we quote "In the scientometrics area, influential papers [ACA] were more about citation behavior (e.g., motivations), while the actual research appeared to focus on citation-analysis studies for research evaluation."

The fourth prominent cluster in this area "co-citation" deals with methods for co-citation analysis as shown by clusters labeled "bibliographic coupling analysis, cluster analysis multi-dimensional scaling, intellectual structure". The citation group is linked to the web studies group by a cluster labeled "case study approach". There is no direct link between the citation studies clusters and the IR clusters. This confirms the observation made in previous studies (White & McCain 1998, Zhao & Strotmann 2008a) that both poles – IR and citation, are still distinct communities with few or no interactions. It is interesting to have this finding confirmed many years later, at the term level.

Globally, the map obtained by terminological analysis (figure 1) corroborates a certain number findings in the ABCA by Zhao & Strotmann's (2008b) on the

same period. These authors labeled three specialities "IR systems", OPAC, "IR-interaction" in the period 1996-2000. We can see the same links in our maps: research on automated IR is linked to that of user studies and to digital libraries. The two-camp structure of IS with sparse connections observed in earlier studies (White & McCain 1998, Zhao & Strotmann, 2008a&b) is still visible in this period. There is evidence that the two poles – IR and citation studies - continue to be the two structuring poles of research in IS, although web studies (or webometrics) is catching up as a third pole. The IR specialty continues to aggregate research on IR interaction and IR systems but is more and more interwoven with web studies and to a lesser degree, is connected to research institutional digital libraries, to scholarly communication, and to studies on theoretical foundations of information science. The citation pole on the other hand is still solidly entrenched on research on scientific collaboration, co-citation methods, journal impact factor studies, bibliometric indicators and research evaluation.

4.2 Period II: 2006-2008

No mapping of the IS field exists on this period as it is the most recent one studied. Therefore, a comparison with other studies is not possible. However, we can refer to the perspectives on the evolution of the IS field outlined in Zhao & Strotmann (2008b) and see whether they are verified. An optimal visualization of topics for this period was reached with 124 clusters at the same co-occurrence and association thresholds (co-occurrence>2, Eij >=0.01). Figure 2 shows the map of topics for this period. The two-camp picture "automated IR systems" vs "co-citation studies" although still discernible starts to show some noticeable changes.

Figure 2. Map of research in IS for the period 2006-2008. 'Term x term' association graph.
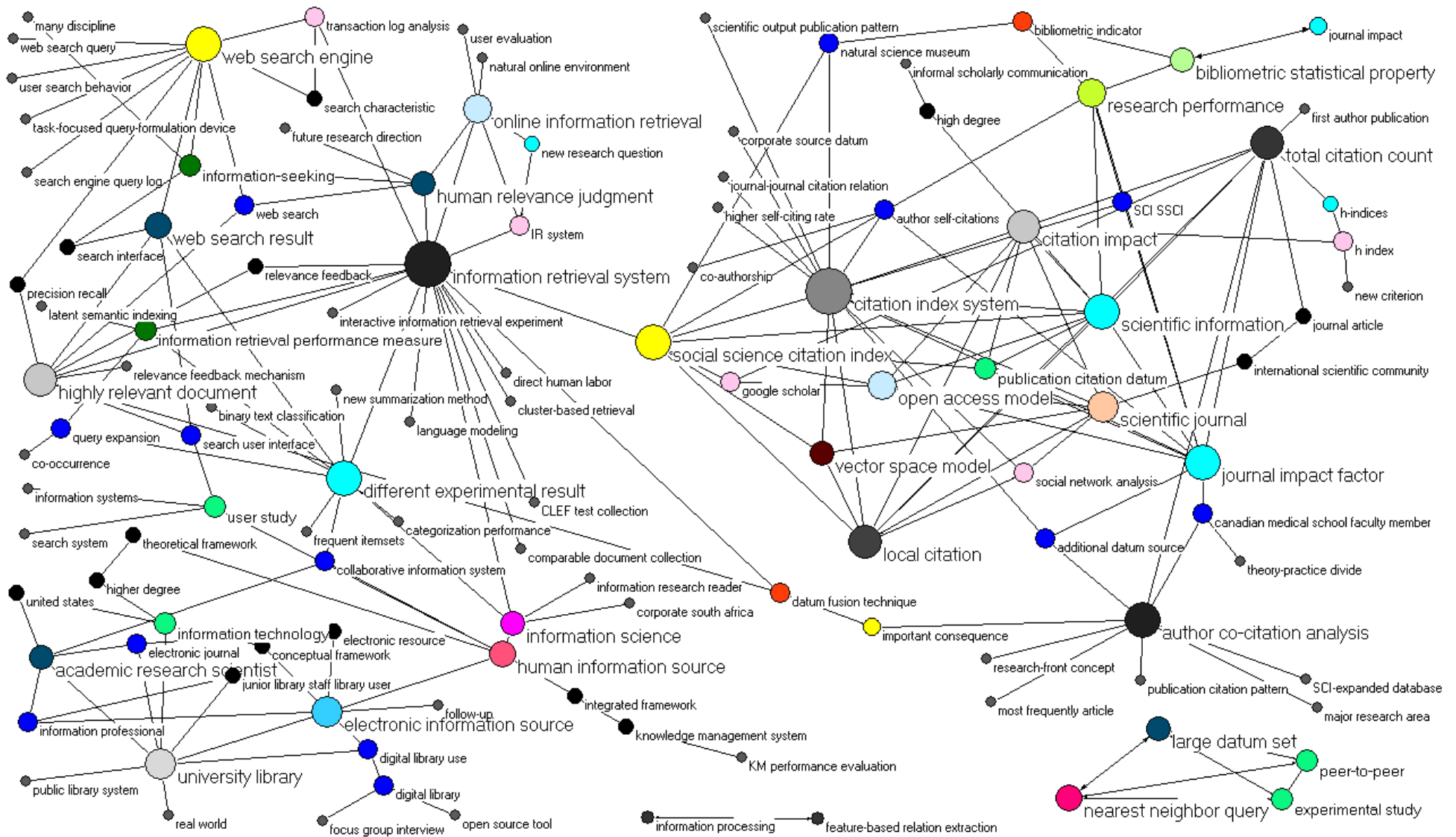
4.2.1 Evolutions in IR research: text mining tasks, machine learning techniques

The IR specialty, although still a major one has become more diverse in this second period (2006-2008). Several related sub-specialties are branching out from it and becoming more prominent. This is evidenced by the proximity of moderately sized-clusters such as "human relevance judgement, online information retrieval, different experimental result, highly relevant document". We also observe the presence of smaller clusters which were already present in the first period such as "IR systems, relevance feedback, CLEF test collection". More importantly, we observe the appearance of newer research topics like "language modeling, new summarization method, binary text classification". They reflect recent trends in IR research which draw upon machine learning techniques for specific text mining tasks.

4.2.2 Evolutions in web studies: web user studies

This pole whose prominence was already observed in the first period (1996-2005) continued to grow in the second period. The focus is now on more on user-oriented search behavior. Surrounding the big "web search engine" cluster, we find two moderately sized clusters labeled "transaction log analysis, web search result". Other smaller clusters are "web search, search characteristic, task-focused query-formulation device, user search behaviour search engine query log, web search query, many discipline". These clusters all point to the prominence of research on transaction log analysis in order to analyze user search behaviour and subsequently measure the performance of web search engines. This focus was not so clear in the first period which was more on webometrics (analysis of web links and web page topology). Thus this second period points to a resurgence in user-oriented studies of web and IR engines. This observation is also in agreement with Zhao & Strotmann's findings (2008b) that webometrics which seemed active in the period 1996-2000 appeared to be on the decline in the later period 2001-2005. This point is further buttressed by the appearance in the lower part of the map of the cluster "user study" linked to clusters on "collaborative information system, information systems, human information source". The user-oriented pole is however not a homogeneous nor quite distinct one. It is split between the IR and the web studies poles thus underlining the transcendental nature of user studies which can concern almost

every aspect of research in IS.

4.2.3 Satellite specialties: information science, digital libraries and knowledge management

The presence of moderately sized clusters labeled "information science, human information source" in the southern part of the map (below the IR cluster) reflect the persistence of research on more theoretical aspects of IS. This group of clusters interestingly lead to smaller clusters on knowledge management (integrated framework, knowledge management system, KM performance evaluation) which were already present in the first period, although the focus in KM research has now shifted to systems design and evaluation.

Leading from this group of clusters is a moderate sized cluster on "electronic information source" linked to another moderately sized cluster on "university library". These two reflect research focus by the digital libraries communities on how these resources can be accessed by the academic public at large. "Electronic information source" leads to a smaller group of clusters labeled "digital library use, digital library, open source tool, focus group interview" and is linked to the bigger cluster "university library". This last cluster "University library" aggregates other clusters on the use of digital library resources by academic faculty members and their interaction with information professionals as evidenced by clusters labeled "junior library staff library use, academic research scientist, information professional". In the first period (1996-2005), digital library research focused mostly on archives preservation, recovery of lost or damaged libraries during the two wars, the integration of e-books and on OPACS. In the second period (2006-2008), the focus is now on access of digital resources. Zhao & Stromann"s (2008b) had observed a similar trend in an earlier period (2001-2005). Quoting the authors "A small but distinct research area has emerged (i.e., e-resources in scientific communication), which studies how scientists interact with electronic resources." The two specialties – digital libraries and information science (theoretical aspects) continue to share connections via "human information source".

4.2.4 Emerging topics in citations studies: vector space, open access, Google scholar and h-index

Like in period 1996-2006, citation studies remain the second biggest research pole in IS. However, we observe some shifts in focus. The most prominent cluster in this pole is labeled "citation index system", followed by "social science citation index, journal impact factor, scientific information, author co-citation analysis". Clearly there are research issues related to the different citation databases (SCI, SSCI) and the evaluation of authors's impact in their fields via author co-citation analysis. Mapping the IS from terms also bring to the forefront specific methodological issues. Research performance evaluation continues to be a driving motivation in bibliometrics research.

We also observe some new focus of co-citation studies which were absent in the first period (1996-2006): the appearance of the cluster "vector space model, open source model, Google scholar" in this second period.

Vector space model is usually a term associated to the IR pole. This cluster is linked to the "citation index cluster" and to the "scientific journal cluster". Upon searching the associated MySql database to understand this association, we found only one paper by Loet Leydersdorff published in a JASIST volume of january 2007, entitled "Visualization of the citation impact environments of scientific journals: An online mapping exercise". In this paper, the author used the vector space model for normalization of the journal-journal co-citation counts. It is remarkable that this co-appearance in a single paper of the association between "vector space model" and several co-citation terms should be picked up by our system, from the thousands of associations of term-pairs from this corpus. In this paper, the term vector space model co-occurred with "citation impact, citation index system, journal citation, local citation, social network analysis, social science citation index, total citation count". It is indeed a confirmation that mapping from the publication content and from linguistic relations can uncover emerging trends in a timely manner.

Recent developments in the open source community and the expansion of Google's technology also become visible in this last period. Seemingly, the bibliometrics community has now to deal with the implications of open source models of dissemination, hence the emergence of the moderately-sized cluster cluster "open access model". This topic is well connected to several of the prominent research issues in co-citation studies such as "social science citation

index, scientific information, citation impact, journal impact factor". Not surprisingly, the two emerging concerns – the impact of "Google scholar" and "open access model" also share a link. The Google scholar cluster contains two components labeled "google scholar" and "google scholar citation". Although the beta version of Google scholar was released in 2004, its more enhanced features have only been added from 2006. Since then, Google scholar, for some bibliometric tasks, may appear as a possible rival of the more established ISI-Thomson's citation databases for elaborating research performance indicators.

Our method was also able to capture the emerging nature of the "h-index" in citation studies. This index, proposed in the landmark paper by Hirsch (2005) measures a researcher's contribution to his field. This topic was correctly identified as a small and marginal cluster suspended to the citation studies clique via the clusters "citation" and "total citation count". Our variation identification component was also able to determine that the two variants "h-index" and "hirsch index" were synonyms and also recognized the spelling variant "h index" as the same concept. All three terms were put in the same component and ended up in the same cluster. On the other hand, the clustering algorithm correctly formed two clusters on this topic based on the fact that there were was another set of variants related to the h-index but not its semantic equivalence. Indeed, a series of variants were formed around the term "h-indices" which became the label of the cluster of the same name. The "h-indices" cluster reflects the scientific impact of Hirsch's publication and the subsequent debate surrounding the h-index. This has led to counter-proposals and modifications of the original h-index (g-index, modified h-index). Hence, forming a separate cluster on "h-indices" which is linked only to the original "h-index" cluster is a way of attracting the readers attention to the fact that there are "h-index like" things. Upon querying the MySql database associated to this period of our corpus, we found that the "h-indices" cluster contains two components labeled "h-indices" and "successive h-indices". On the other hand, the cluster "h index" contained three components labeled "hirsch index, h index, successive H index", thus with explicit reference to the original h-index. Other terms in this cluster refer to other indices that are being compared to the h-index: jaccard index, citation index, price index, coupling index, full-content index, ACIF index. Our method was thus able to distinguish between the original "h-index" and all the modified versions proposed since. More important for

timely novelty detection is the fact the h-index cluster is associated to a cluster labeled by a seemingly trivial term "new criterion". In the light of what we know of the interest generated by the h-index among the bibliometric and the scientific communities at large, this term is not so trivial. It alerts researchers to the novelty of this topic in unambiguous terms.

We generated other maps of the IS field based on keywords and on authors but owing to space limitations, we cannot show them in the current paper. Interested readers can find all the maps at http://pub.termwatch.es/imagesLIS/ .

5. Conclusion

We proposed a methodology combining symbolic and numeric information for domain mapping. The linguistic components first effect meaningful groupings of terms into tight semantic components leading to semantic clusters. The labels of clusters are automatically derived from titles and abstracts based on linguistic criteria and they reflect the actual terms used by the authors. Ultimately, what our method contributes is to lend life to the map by automatically labeling the specialities and providing sufficient topical elements such that manual reading of author publications is rendered if not totally unnecessary, greatly alleviated. This equally removes the need to resort to extensive human background knowledge in order to provide the narrative on the maps. Also, mapping from the text level has brought to light smaller but emerging research trends. Our methodology was for instance able to capture the budding nature of the "h-index, google scholar and open access models" in the citation studies specialty and capture recent trends in IR research such as automatic summarization and the use of language models in some IR tasks. Mapping from text fields gives a more detailed picture of the domain and yields more intuitive results.

We have already pointed out some major correlations between the structure of research in IS revealed by term maps and the one obtained through author-bibliographic-coupling-analysis (ABCA) in Zhao & Strotmann (2008b). Here we point out a few differences in the results from both studies. Concerning the structure of IS research as revealed by term maps, we did not observe a decline in the dominant position of research on IR systems contrary to the observations made in Zhao and Strotmann (2008b) for the same period. The term maps (Figures 1 & 2) consistently showed IR as occupying a central place in IS although it has become more diverse in the later period of 2006-2008. The IR pole has expanded to include various issues relating to IR

systems design and evaluation (TREC, CLEF, document collections) but also user-oriented IR studies as they pertain to user interaction with online IR systems.

Zhao & Stotmann (2008b) did however observe that "The IR camp, by contrast, displays evidence of major internal restructuring during this decade". Our findings are in agreement with this statement that the IR pole is expanding and perhaps re-structuring. Zhao & Stotmann (2008b) observed on other hand that "the literatures camp" (citation studies) showed remarkable stability throughout the period of their study (1996-2005) whereas one would expect that its connections with webometrics should have induced some changes. The maps we obtained at the term level corroborate this observation that the citation studies pole up till now, continued unperturbed by going-ons around it.

Another difference between our findings and those of Zhao and Strotmann (2008b) is the persistence on the term maps, of a small group of clusters on "knowledge management" in the vicinity of the IR pole. In Zhao and Strotmann (2008b), "knowledge management" appeared on the author co-citation analysis (ACA) as an influence on current IS research but not on the ABCA maps as a current active topic in the period 2001-2005.

While the methodology we have designed shows a lot of promise, there is certainly room for improvement on some aspects. Determining the optimal number of clusters for any clustering algorithm is still an open research question. A lot of research has been carried out on this topic, but no one solution fits all. In Janssens et al. (2006), Ben-Hur's (2002) cluster stability method was used to select six clusters as the optimal number for partitioning words from five IS journals. The authors observed however that the stability diagram did not show a clear cut solution for their data and that overall mean silhouette values were low. This led to some mis-classifications. Some clustering algorithms require that this number be fixed a priori (k-means) based on the analyst's perception of what the optimal partition should be. This is still a matter for further research.

Acknowledgements

References

1.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, 2007, 58(7): 947–957.

2.Batagelj V., Mryar A. (2009) Pajek. Program for Large Network Analysis. Retrieved online on 15 january 2009 [http://vlado.fmf.uni-lj.si/pub/networks/pajek/]

3.Boyack K., Börner K., Klavans R. (2007). Mapping the structure and evolution of chemistry research. In Torres-Salinas, D., & Moed, H.F. (eds.)., Proceedings of the 11th International Conference of the International Society for Scientometrics and Informetrics. Center for Scientific Information and Documentation of the Spanish Research Council, Madrid, Spain, 112–123.

4.Ben-Hur A., Elisseeff A., Guyon, I. (2002). A stability based method for discovering structure in clustered data, Pacific symposium on biocomputing, 7, 6–17.

5.Callon M., Courtial J. P., Turner, W., & Brain, S. (1983). From translations to problematic networks. An introduction to co-word analysis. Social Science Information, 22(2), 191–235.

6.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), pp. 359-377.

7.Dunning T. (1993). Accurate methods for the statistics of surprise and coincidence. Computational Linguistics, 19(1), 61–74.

8.Garfield E. (1979). Citation indexing – its theory and application in science, technology and humanities, John Wiley & Sons, NY, 274p.

9.Harris Z. S., (1968) Mathematical Structures of Language, New York: Wiley, 1968.

10.Hirsch J. E. (2005)., An index to quantify an individual's scientific research output, PNAS 102 (46): 16569–16572.

11.Ibekwe-SanJuan F. (1998), Terminological variation, a means of identifying research topics from texts, Joint International Conference on Computational Linguistics (COLING-ACL'98), Montréal, Québec, 10-14 August 1998, 564-570.

12.Jansens F., Leta J., Glanzel W., De Moor B. (2006), Towards mapping, library and information science, Information Processing and Management, 42, 1614–1642.

13. Kessler M.M., Bibliographic coupling between scientific papers. American Documentation, 1963, 14, 10–25.

14. Leydesdorff L. (1997). Why words and co-words cannot map the development of the sciences, Journal of the American Society for Information Science, 48(5), 418–427.

15. Salton G., McGill, M. J. (1986). Introduction to modern information retrieval. New York: McGraw-Hill, Inc.

16. SanJuan E., Ibekwe-SanJuan F., (2006) Textmining without document context, Information Processing & Management, Special issue on Informetrics II, 42(6), 1532-1552.

17. Schiffrin R., Börner K. (2004), Mapping knowledge domains, Publication of the National Academy of Science (PNAS), 2004, 101(1), 5183-5185.

18. Small H. (1999), Visualizing science by citation mapping, Journal of the American society for Information Science, 50(1999), n° 9, 799-813.

19. 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 and Technology, 49(4):327–355, 1998.

20. Zhao D., Strotmann A., (2008a), Information Science during the First Decade of the Web: An Enriched Author Cocitation Analysis, Journal of the American Society for Information Science and Technology, 59(6):916–937, 2008.

21. Zhao D., Strotmann A., (2008b), Evolution of Research Activities and Intellectual Influences in Information Science 1996–2005: Introducing Author Bibliographic-Coupling Analysis, Journal of the American Society for Information Science and Technology, 59(13):2070–2086, 2008.