International Journal of Competitive Intelligence, Strategic, Scientific and Technology Watch
SciWatch Journal 2 (2009), Issue 1, 72 - 73

REVIEW OF ARTICLES: “Towards mapping library and information science”

Information Processing and Management: an International Journal Special issue: Informetrics
Volume 42, Issue 6 Pages: 1614 - 1642
Frizo Janssens, Jacqueline Leta, Wolfgang Glänzel, Bart De Moor (2006)
(Review by M. Welling Flensborg)
e-mail: m.welling-flensborg@gmail.com

The article aims to explore the field of library and information science (LIS) and presents six different clusters of study that are a result of the process of data mining approximately 1000 articles and notes from five journals in the field of LIS. The study is concluded with the analysis of cluster representations by the selected sources of information.

The first idea presented in the text is the application of quantitative linguistics (text analysis) in informetrics and bibliometrics. At present, the most frequent techniques, co-word, co-heading and co-author clustering, are based on the analysis of co-occurring keywords, terms extracted from titles, abstracts and/or full text, subject headings or cited authors.

Different authors and their contributions to the field are cited, only to conclude that LIS is a very heterogeneous field that includes subdisciplines such as traditional library science, IR, scientometrics, informetrics, patent analysis and most recently the emerging specialty of webometrics. In this context, the authors’ objective is not providing a theoretical contribution to computational linguistics, but applying and extending our methodological approach to a broader, more heterogeneous set of documents. The challenge is not the growing number of articles, but the heterogeneity of this hybrid field and the variety of terms and concepts used by scientists in our field. The main techniques used by the authors in this paper are statistic-based and not language processing-based.

The study is based on the following questions as thesis:

1. Can the assumed heterogeneity be characterized by means of quantitative linguistics?
2. What are the main topics in current research in information science?
3. Have new, emerging topics, already developed their own ‘‘terminology’’?
4. Can the cognitive structure be visualized and represented using multivariate techniques?
5. How are topics and sub-disciplines represented by important journals of the field?

In order to be able to answer these questions, 938 articles and notes from 5 selected journals are processed and vocabularies for subdisciplines within LIS are presented. The article also compares different methods of clustering and mapping in order to reach the optimum presentation of the cognitive structure of our field (text representation, text preprocessing, text extraction, multidimensional scaling, clustering).

The results are presented in the following way: First of all, a general index of 11151 stemmed terms or phrases. Secondly a graphical representation of the articles, that meaning, for example, a multidimensional scaling (MDS) map of the 938 articles or notes in two and three dimensions (showing the relations, proximity, or similarities between the articles, for example). In third place, the presumable errors, ambiguities or mismatches resulting from the process are detailed and explained. Finally the authors analyze the relation that exists between the defined clusters and the journals from which the information was gotten in first place. For this matter, it is concluded that the relatively large distances among clusters and between each cluster and the journal, strongly indicates that a quite large spectrum of bibliometric, technometric and

informetric research using different vocabularies is covered by the journal “Scientometrics”. This observation is in line with the findings by other authors that affirm that scientometrics consists of several subdisciplines such as informetric theory, empirical studies, indicator engineering, methodological studies, sociological approach and science policy.

The article concludes that, after the process of data mining (the analysis of the conceptual structure of five journals representing a broad spectrum of topics in the field of LIS, focusing on the analysis of the ‘‘pure’’ text corpus, excluding any bibliographic or bibliometric components which might influence or even distort the quantitative linguistic analysis of the scientific text.) six different clusters are clearly defined: two clusters in bibliometrics, of which a big one is about bibliometrics/research evaluation, and a smaller one about methodological/theoretical issues; also two large clusters in information retrieval and general and miscellaneous issues were found and, finally, two small emerging clusters in webometrics and patent and technology studies were defined. Within the IR cluster, we have found a small sub cluster on music retrieval, which might be a temporary phenomenon since the journal JASIST has published a special issue on this topic. Aside from clustering, interesting patterns have been found (e. g. in a 3-D model, a diffused tripod is defined; the papers published in Scientometrics were arranged in two of the three legs forming the tripod. The ‘‘two legs’’ were formed by Bibliometrics1 and Patent on the one hand, and Bibliometrics1 and Bibliometrics2 on the other hand. The border between the two bibliometrics clusters is fuzzy. Moreover, the cluster dendrogram has shown that Bibliometrics1 is combined with Patent first, before being combined with Bibliometrics2.