Original scientific paper

Telling the Future of Information Sciences: Co-Word Analysis of Keywords in Scientific Literature Produced at the Department of Information Sciences in Zagreb

Siniša Bosanac, student
Department of Information Sciences
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10 000 Zagreb, Croatia
sbosanac@ffzg.hr

Marija Matešić, student
Department of Information Sciences
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10 000 Zagreb, Croatia
mmatesi1@ffzg.hr

Nino Tolić, student
Department of Information Sciences
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10 000 Zagreb, Croatia
nitolic@ffzg.hr

Summary

Research fields dealing with information from various perspectives have been rapidly developing throughout the last few decades. Information science is one of the most prominent among them. The aim of this article was to investigate how concepts related to information sciences in Croatia change over time, and by doing this to show the development of the field. For this purpose, co-word analysis has been used. Using this method, the most important concepts of information sciences that appeared in the 14-year period 1995-2009 were indentified, and the results have been visualized. The concepts are visualized in the form of a network along with their respective clusters for the whole time frame, and also separately for two seven-year periods, 1995-2002 and 2002-2009. The analysis has shown concepts centred on education and community to be the most prominent and stabile. New concepts in the network appear independently, as a replacement for similar concepts, or as a result of braking down of general concepts into more specific ones. Results presented in this paper are of purely quantitative nature and, if combined with observations of relevant external *factors, can serve as a basis for a study of trends in scientific production, and ultimately, their prediction.*

Key words: co-word analysis, clustering, data visualization, information science, key words, scientific literature

Introduction

Predictions are a key factor in every decision-making process. In every aspect of life we operate using predictions and in most cases we are not even aware of it. Short-term predictions are more accurate, but long-term ones enable us to make decisions that lead to much greater gains. All professional long-term predictions are based on previous trends and regularities. In order to recognize trends and regularities, we need detailed and precise data on previous development in the area on which we focus. The time-span and the quality of data determine the quality of predictions.

Human action is arguably the most difficult phenomenon to predict. In science, things are made somewhat easier by the fact that science is a very structured activity which records even the smallest steps it makes in the form of scientific literature. The analysis of these records is done using bibliometric methods. It is only natural that those methods are used to describe the field in which they originated – information sciences.

One of the key characteristics of the Department of Information Sciences at the Faculty of Humanities and Social Sciences in Zagreb is its interdisciplinary area of activities and increased dynamic of its development, which is also typical for other academic organizations in this field. Being a scientific and educational organization, the Department has two kinds of output. The first kind, the *scientific output*, is principally measured by the number of completed research projects and produced scientific literature. The second kind*, educational output*, can be measured by the type and number of available courses and the number of its graduated students and their specializations.

In this paper, we have established a framework for representing the development of its *scientific output*. For this purpose we have used co-word analysis, an established technique for mapping the structure and dynamics of science[1], to analyze keywords in scientific papers produced by members of the Department in the 1995 – May 2009 period catalogued in the Croatian Scientific Bibliography. The aim of this article was to describe the methods used and to provide quantitative results. In order to interpret them and to draw qualitative conclusions about the development of information sciences, it would be necessary to take into account various other factors. One of those factors is the question whether keywords provided by authors accurately represent the actual topic of a scien-

[1] Qin He. Knowledge Discovery Through Co-Word Analysis. // *Library Trends*. Vol. 48 (1999), No. 1; pp. 135-159.

tific paper's content.[2] We have not dealt extensively with this issue in our article, and for its purposes, description using author-provided keywords is taken to be accurate.

Method

Co-word analysis is a bibliometric technique that examines co-occurrence of keywords. (Glänzel, 2004) Words are the most important in this analysis, and can be extracted from various types of scientific publications. They can be mined from titles, abstracts, full texts or keyword lists of various types of scientific publications. The main purpose of this technique is to show the dynamics of scientific field's development by visually representing the co-occurrence matrix of words chosen according to their frequency in the corpus. Higher co-occurrence frequency of two keywords indicates closer and stronger links between them. The closer links between two keywords represent closer relationships between the concepts they refer to.

Data harvesting

A total of 376 articles authored by 35 members of the Department of Information Sciences at the Faculty of Humanities and Social Sciences in Zagreb were harvested[3] from the Croatian Scientific Bibliography (CROSBI)[4] from 1995 to the present. This period was selected because information on publications dating earlier than 1995 were unavailable at the time the data harvesting was conducted. Publications were selected according to type of publication, field of science, and author. Types of publications that were included in the harvesting process were primarily scientific articles. Books, book abstracts, book chapters, conference reports, unpublished papers, or graduation thesis were not included.
The publications' title, keywords and year of publishing were harvested according to the chosen 35 authors that are members of the Department of Information Sciences at the Faculty of Humanities and Social Sciences in Zagreb. That was chosen harvesting filter to include information science related publications. Publications that did not have English keywords listed were excluded from the analysis. Keywords were not standardized because a thesaurus was not available, so there is a possibility of inconsistencies with standard terminology of information sciences.

[2] For a discussion on this problem see Whittaker et al. Creativity and Conformity in Science: Titles, Keywords and Co-word Analysis. 1989.

[3] Harvesting process was automated by using CURL module in manually written PHP script.

[4] The website of Croatian Scientific Bibliography (CROSBI) can be found at http://bib.irb.hr/

Data processing

A total of 689 unique keywords from 1136 tokens (keyword forms) were harvested out of 367 articles covering the 1995-2009 period. Keywords unmistakably denoting the same concept, or occurring in different forms were standardized through the process of normalization and lemmatization. As a part of lemmatization process all inflected forms were reverted to their base form, except when changing the form would change the meaning of the whole keyword. During normalization, abbreviations were converted into their full form, e.g., CAL + Computer assisted learning = Computer assisted learning; CALL + Computer Assisted Language Learning = CALL; EU + European Union = European Union; HMM + Hidden Markov Model = Hidden Markov Model; IT + Information Technology = Information Technology; LIS + Library and information science = Library and information science; LMS + Learning Management System = Learning Management System.

After lemmatization and normalization data were converted to a data format supported by Bibexcel[5] – freeware software for bibliometric analysis, and co-word analysis in particular. Within Bibexcel as a tool, word frequency is calculated. Finally, words with frequency more than two were selected for the next step - co-word analysis. A co-occurrence matrix was formed that shows relationships between phrases or words.

To provide a very clear view what is happening in co-occurrence matrix a visualization tool called Pajek[6] was used to map co-occurrence data. Pajek is an open source program for large network decomposition, visualization and clustering. Co-occurrence data were visualized using Kamada & Kawai algorithm as it available in Pajek. This draws general graphs with minimal energy[7]. In order to keep visualization readable, the analysis was limited to words that co-occurred with frequency more than two.

Results

In order to show the development of the observed scientific field it is necessary to show how the results changed over time. To facilitate this, the results were divided into three parts. The first part shows the results for the whole period 1995-2009. The other two show results for two seven-year periods, 1995-2002 and 2002-2009. The analysis was done independently for each period.

[5] Bibexcel is publicly available at: http://www.umu.se/inforsk

[6] The homepage of Pajek can be found at http://pajek.imfm.si/doku.php

[7] Wikipedia Contributors. Force-based algorithms. Wikipedia, The Free Encyclopedia. http://en.wikipedia.org/wiki/Force-based_algorithms (Jul 24, 2009), Aug 16, 2009.

1995 - 2009

A total number of 376 scientific articles by 35 authors were collected for the whole period. These articles contain 689 unique keywords. Finally, a data matrix of 64 co-occurring words was created and visualized using Pajek.

Visualization of co-occurring words for whole period shows the existence of five main clusters gathered around strongest connected nodes which are *museology, Croatian language, scientific communication, information literacy and education* as it is presented in Figure1. Cluster *museology* includes topics relating to users, museums and repositories. Cluster *Croatian language* is related to natural language processing concepts and topics as the field of artificial intelligence. *Information literacy* as a core cluster contains knowledge and library, which are public administration oriented concepts. Cluster *scientific communication* includes topics on scientific activities. Cluster *education* includes concepts related to networked society, which is a result of the impact of the Internet and information technology. An isolated cluster which contains keywords *computer-assisted language learning*, *web application* and *Croatian old dictionary portal* can also be observed. This cluster is not related to any of the given clusters.

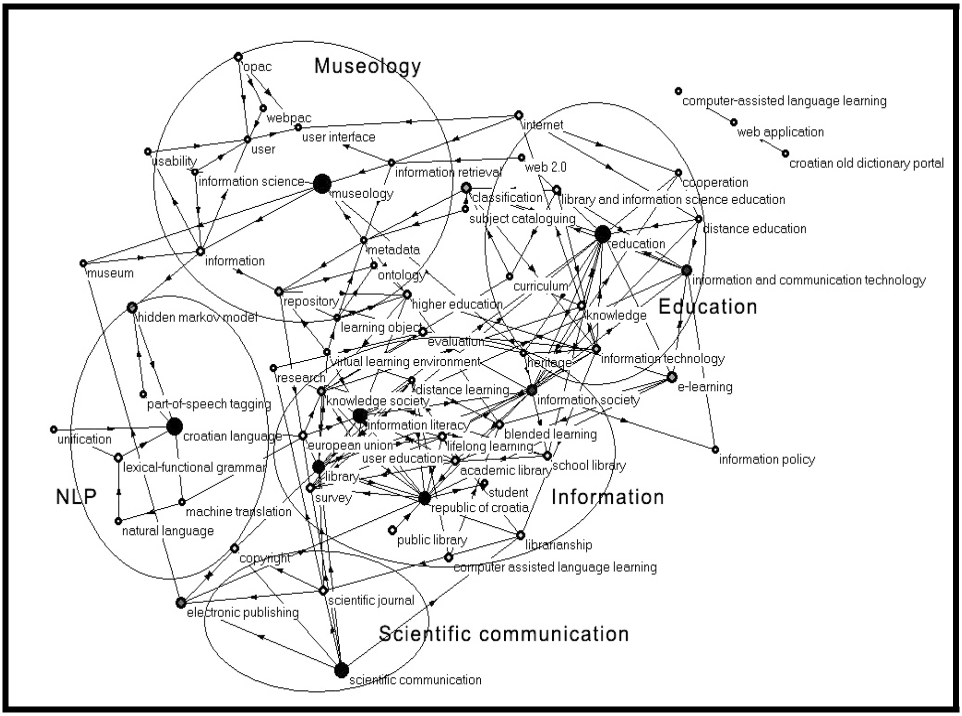

Figure 1: Visualization of 64 words with frequency greater than 2 for 376 information science related articles for the whole period (1995-2009)

1995 – 2002

In the first time period from 1995 to 2002 there is a total of 35 items shown in the diagram (Figure 2). The three largest nodes are as follows: *education, information communication technology* and *library*. As shown on Figure 2, five larger clusters are visible: *information technology, education, NLP, community* and *Interfaces*. Cluster *information technology* contains nodes which are linked to technology itself, as well as possible applications of technology, such as *classification* and *computer assisted language learning.*

Cluster *community* contains nodes related to general public such as *library, knowledge society, library users* and the like. This cluster is worth examining in greater detail, due to the possible relevancy to the time frame in question.

Cluster *education* contains nodes related to education itself and its evaluation such as *comparison* and *characteristics.* Close relationship this cluster shares with cluster *information technology* could be worth examining further.

Cluster *interfaces* contain nodes related to information search and retrieval such as *OPAC* and *WEBpac*. Isolated cluster *NLP* contains nodes related to natural language processing of the Croatian language.

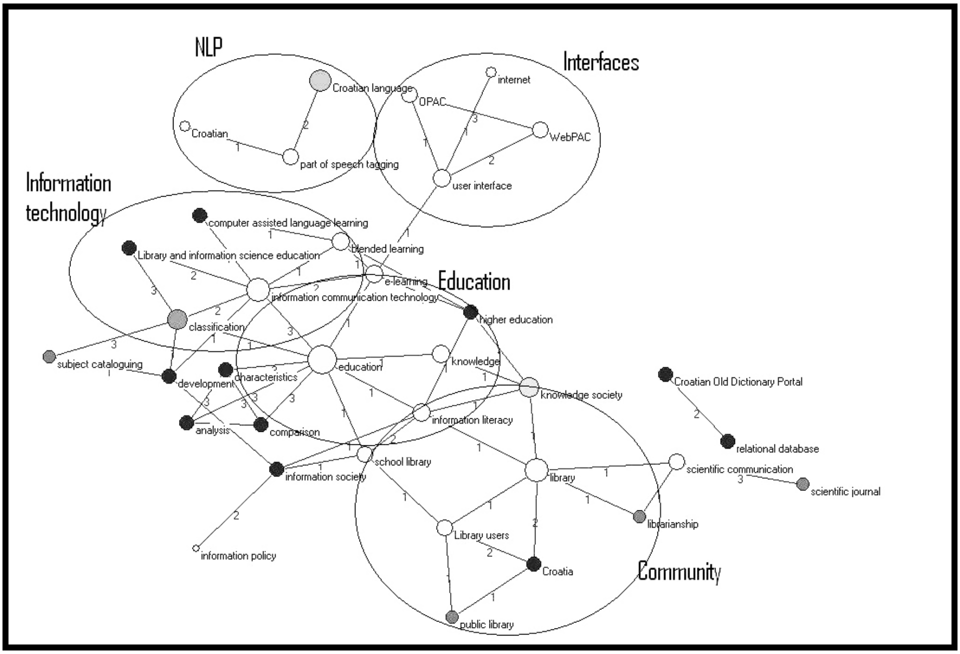

Figure 2: Visualization of 35 words with frequency greater than 2 for 187 information science related articles for the time period between 1995 and 2002

2002 – 2009

In the second time period from 2002 to May 2009 there is a total of 37 items shown in the diagram (See Figure 2). The five most interconnected nodes are the following: *education*, *repository*, *information literacy, higher education*, and *information and communication technology*.

Visualization of co-occurring words for time period between 2002 and 2009 contains four clusters: *Education*, *Repository*, *Web 2.0*, and *Community*. The largest is the one centered on *education*. Its strongest connections are with nodes that could be roughly described as analytical in meaning: *analysis*, *comparison,* and *characteristic.* The rest of nodes in the cluster could be described as web-related: *content management system*, *web 2.0*, *semantic web*. Links with strength 1 are those with *information and communication technology*, *e-learning*, *information literacy*, *school library*, and *knowledge*.

Second largest is the cluster with *repository* as its central node. Its members could be described as mostly maintenance-related; *technical support*, *valorization*, *evaluation*, *standardization*, and there is also *digital educational material*, which is slightly different by its meaning. The weaker links with its central member are with *information*, *metadata*, *ontology*, *higher education*, and *survey*.

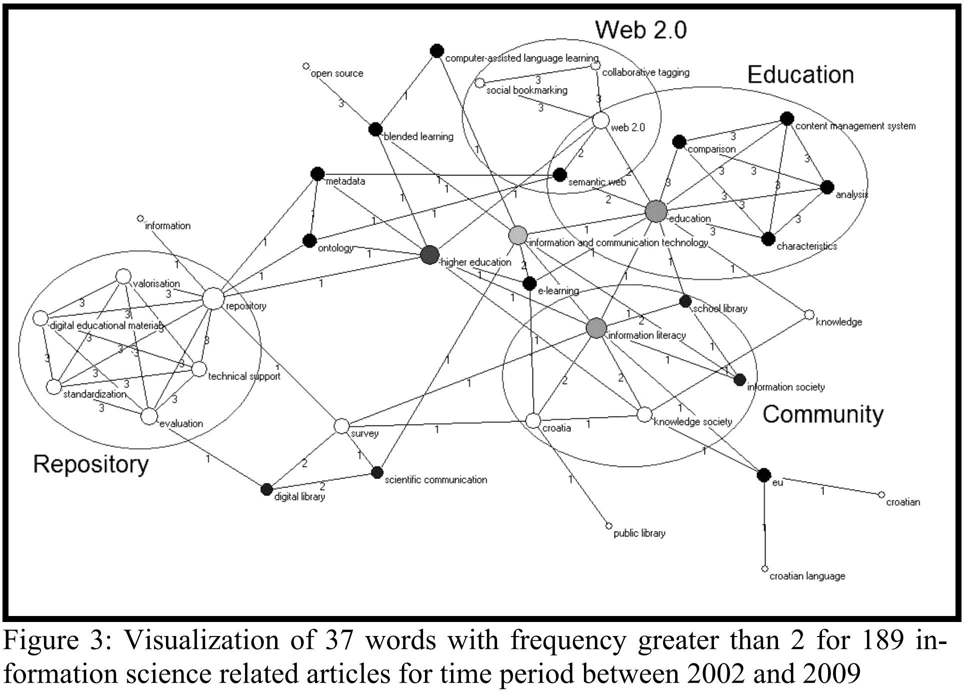

Figure 3: Visualization of 37 words with frequency greater than 2 for 189 information science related articles for time period between 2002 and 2009

The third cluster is that centered on *web 2.0*. Its member-nodes are typical subordinate concepts of web 2.0: *social bookmarking*, *collaborative tagging* and *semantic web*. There are also links with *education* and *higher education*.

The fourth primary cluster, *Community* is the one centered on *information literacy*. Its members are *school library*, *knowledge society*, and *Croatia*, with links to *EU*, *information society*, *knowledge society*, *survey*, *higher education*, *education* and *information and communication technology*.

There are also several secondary clusters: *digital library - scientific communication - survey, blended learning – open source - computer-assisted language learning.* Links between *education* - *knowledge* and *knowledge society* could also be very interesting to examine, same as those between *EU – Croatian – Croatian language – knowledge society,* and *information literacy*.

Time period comparison

It is important to note that these represent only the most frequent concepts in a relatively broad time period.

The most obvious difference between the two time periods is the number of clusters (first period 5, second 4), and the fact that the later time period does not have isolated clusters and nodes.

The *Education* cluster is the most prominent in both time periods. Its main members and links have remained mostly unchanged – particularly members belonging to group of analytical concepts, and links with *information and communication technology*, *knowledge, knowledge society, school library,* and *e-learning*. The difference in this cluster is that it lost its links with *classification*-related nodes, and formed several strong ones with *Web 2.0* cluster.

Although the nodes from the larger part of *Information technology* cluster from the first period are also present in the second, it was not designated as an independent cluster in the second period because the links between its members were weaker. A significant difference in this group is the disappearance of three prominent nodes with strong connections - *classification*, *subject cataloguing*, and *library and information science education*.

Cluster *Community* has kept a significant number of its nodes, namely, *knowledge society*, *school library*, *Croatia*, and *information literacy*. The cluster in the second period is centered on *information literacy*, which is much more pronounced than in the first period. Other changes were the reduction of library-related concepts, and the emergence of *European Union* as a concept interconnected with cluster *Community*.

NLP cluster, which was isolated in the first period, dissolved in the second period with the disappearance of its central member – *part of speech tagging*, while its remaining members formed links with *European Union* node.

Interfaces cluster also dissolved with all of its members disappearing. *Internet* node, which had strong connections, also disappeared, but it was replaced by an entire cluster named *Web 2.0*, whose members can actually be considered as subordinate concepts of the *Internet*. This would mean that the node *Internet*, actually multiplied and became more specific.

Cluster *Repository*, which appeared in the second time period, consists of newly formed nodes, but also has connections with some of the "old" ones, such as *scientific communication* and *higher education*.

Some independent nodes denoting more specific concepts, such as *information policy*, *scientific journal,* and *Croatian Old Dictionary Portal*, are not present in the second time period, while new ones, such as *open source*, *metadata,* and *digital library,* have appeared. More general concepts, e.g. *knowledge*, and those denoting scientific methodology, such as *survey*, *analysis*, *comparison*, and *characteristics*, also did not change. From this we can conclude that specific concepts are more dynamic than the general ones, and those describing methodology.

Conclusion

The goal of this article was to investigate the main topics and trends within the field of information science during the time period from 1995 to 2009. Using co-word analysis and by comparing the two time periods, first from 1995 to 2002, second from 2002 to 2009, along with the overarching period, we have endeavored to present our findings through quantitative analysis. Several interesting topic shifts were uncovered, all meriting further qualitative and quantitative research.

In order to improve upon our research we propose several modifications. For a more exact analysis, it would be necessary to include more publications, to display the results in more segmented and narrower time spans as to increase the resolution of graphical representations. A guideline for author-added keywords that would prescribe the classification of keywords according to a hierarchy of concepts, and whether they describe the method, or the actual topic of the article, would greatly improve not only the quality of their retrieval, but also the precision of similar analyses. Also, it should be noted that improvements in the storage, classification and general usability policies on CROSBI servers are in order.

With these modifications the research would be more precise and perhaps, would uncover more. Hopefully, the topic in question will be revisited on a broader scale than was possible in this article.

References

Ding Ying, Matthew; Gobinda G. Chowdchury, Foo, Schubert. Bibliometric cartography of information retrieval research by using co-word analysis. 2001. http://www3.ntu.edu.sg/home/assfoo/publications/2000/00ipm_fmt.pdf (Jun 24, 2009)

Glänzel, W. Bibliometrics as research field. 2004. http://www.norslis.net/2004/Bib_Module_KUL.pdf (May 10, 2009)

Jokić, Maja. Bibliometrijski aspekti vrednovanja znanstvenoga rada. Zagreb: Sveučilišna knjižara, 2005

Leydesdorff, Loet. The university-industry knowledge relationship: Analyzing patents and the (May 10, 2009)

Qin He. Knowledge Discovery Through Co-Word Analysis. // *Library Trends*. Vol. 48 (1999), No. 1; pp. 135-159

Whittaker, John; Courtial, Jean-Pierre; Law, John. Creativity and Conformity in Science: Titles, Keywords and Co-Word Analysis. 1989. http://www.jstor.org/stable/285083 (May 13, 2009)

Wikipedia Contributors. Force-based algorithms. Wikipedia, The Free Encyclopedia. http://en.wikipedia.org/wiki/Force-based_algorithms (Jul 24, 2009), Aug 16, 2009.