Journal Publishing in Developing, Transition and Emerging Countries

Proceedings of the 5th Belgrade International Open Access Conference 2012

Belgrade, Serbia, May 18-19, 2012

Edited by Pero Šipka
Journal Publishing in Developing, Transition, and Emerging Countries

Proceedings of the 5th Belgrade International Open Access Conference 2012

Belgrade, Serbia
May 18-19, 2012

Pero Šipka (ed.)

Publisher
Centre for Evaluation in Education and Science
Kneza Miloša 17
11.000 Belgrade
Serbia
tel. +381 11 32 38 506
www.ceon.rs
Copyright information
© 2013 Centre for Evaluation in Education and Science

ISSN (print): 2334-7880
ISSN (online): 2334-7880

Cover by PeteF

Ordering information
Print-on-demand by Amazon.

Web access and citation information
This entire proceedings can also be viewed on the web at http://boac.ceon.rs/public/full/5th-bioac.pdf. Each paper has a unique identifier (DOI) which can be added to citations to facilitate access. The DOI # should not replace the full citation. Each e-version contains instruction on how to cite a paper.
Maps of Science Based on Keywords of Articles’ Antecedences, Presences, and Consequences: Application of the CEON/CEES Model of Multi-Perspective Description of Articles

Tanja Jevremov¹, Dejan Pajić², Miša Sotirović³ and Pero Šipka⁴

¹ jevremov@neobee.net
University of Novi Sad, Faculty of Philosophy, Dpt. of Psychology, Dr Zorana Đinđića, 21000 Novi Sad (Serbia)

² dpajicns@gmail.com
University of Novi Sad, Faculty of Philosophy, Dpt. of Psychology, Dr Zorana Đinđića, 21000 Novi Sad (Serbia)

³ sotirovic@ceon.rs
Centre for Evaluation in Education and Science, Kneza Miloša 17, 11000 Belgrade (Serbia)

⁴ sipka@ceon.rs
Centre for Evaluation in Education and Science, Kneza Miloša 17, 11000 Belgrade (Serbia)

Abstract

A model of multi-perspective article description (MPAD) was explored and preliminarily tested. The model assumes that journal articles should be described for bibliographic purposes from three different perspectives, i.e. by using keywords extracted from: (1) article titles and abstracts, (2) titles of their cited references, and (3) titles and abstracts of articles citing them in the future. In order to explore the relationships among keyword types and to test the model preliminarily, a method labeled as Multistage Indexing of Subject Headings (MISH) based on the Keyphrase Extraction Algorithm (KEA) was employed to provide all three types of keywords for all articles from the sample. The articles were sampled from SCIndeks: The Serbian citation index. Three separate maps of (local, peripheral) science were constructed, each based on a different type of keywords. The Partitioning Around Medoid method (PAM) for cluster analyses, followed by multidimensional scaling for visual representations of extracted clusters, was employed.
Results suggest that the three types of keywords generate relatively similar maps, encouraging keywords aggregation for practical purposes. Some differences among the maps are not fully consistent with the predictions derived from the model. They reveal some methodological deficiencies of the study and indicate the most promising directions for further research.

Keywords: citations; titles; models; science; maps; indexing; articles; abstracts; references; clusters

Introduction
Description of research papers by keywords extracted from their titles and abstracts is a rather standard procedure to ensure information retrieval (Medelyan & Witten, 2008). Although most widespread, this procedure is heavily criticized for numerous imperfections. Keywords generated by authors are repeatedly found to be too diverging, those generated by experts too expensive, while the ones extracted by non-human (intelligent) agents too simple-minded to reveal the implicit contents of papers.

The criticism in question stimulated research looking for alternative solutions. This resulted in the emergence of some innovative approaches, the most important being citations-based (related records), but efforts to improve text-based codification of research papers proceeded. Research along this line, that has continued to these days, was focused in its early phases on the automatic extraction of keywords from articles’ references. The method was implemented in ISI (now Thomson) citation databases under the trademark KeyWords Plus® (Garfield & Sher, 1993). The interest was recently extended to include the utilization of metadata from citing papers. The primary interest in experiments using this approach varied from identifying papers relevant to a research topic of interest (Qazvinian & Radev, 2008) to the summarization of the results of cluster analysis (Chen et al., 2010). Studies on “reference direct indexing” (Bradshaw, 2003; Ritchie, Teufel & Robertson, 2008), “citation context” (Elkiss et al., 2008), and “citation-based automatic indexing” (Mahdi & Joorabchi, 2010) showed that using information from citing documents can substantially improve the full-text indexing and, consequently, the searching of scientific literature.
By capitalizing on the above ideas and findings, we developed a model of multi-perspective articles description (MPAD; Šipka et al., 2011). MPAD is based on the assumption that a single piece of research has to be observed and described “in motion”, starting with the previous research from which it arose and ending with the subsequent research it triggered. Accordingly, keywords describing an article as a report of a single piece of research should be generated not only from this article’s metadata (title and abstract) or full text, but also (a) from the titles of all its references and (b) metadata belonging to all papers citing this particular article. MPAD can be illustrated by a logical framework matrix reflecting various phases of the research cycle (as different perspectives of viewing the targeted research) within various paradigms existing in the philosophy of science and research methodology, including their genuine parallel from the economy of industry (Table 1).

<table>
<thead>
<tr>
<th>cycle phases</th>
<th>paradigms</th>
<th>economic concepts</th>
<th>bibliometric metaphors</th>
<th>substrate location</th>
<th>keywords extraction sources</th>
<th>keywords types</th>
</tr>
</thead>
<tbody>
<tr>
<td>raw material</td>
<td>intellectual base</td>
<td>previous studies</td>
<td>titles of article references</td>
<td>KW-Rs</td>
<td></td>
<td></td>
</tr>
<tr>
<td>value added</td>
<td>research front</td>
<td>present study</td>
<td>article titles and abstracts</td>
<td>KWs</td>
<td></td>
<td></td>
</tr>
<tr>
<td>product</td>
<td>research heritage</td>
<td>subsequent studies</td>
<td>citing articles’ titles and abstracts</td>
<td>KW-Cs</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

MPAD was implemented in a CEON/CEES web application supporting journal publishing (ASEESTANT: The SouthEast European Journal Production Assistant, http://aseestant.ceon.rs). ASEESTANT is a journal management system created primarily to serve the idea of “quality enforcement”, in order to bring journal publishing in the region to the level of world reputed periodicals. Enforcement of the quality is to be ensured, among else, by KwASS (Keywords Assignment Support System), a module designed to produce three kinds of keywords posited by the MPAD model. All three types of keywords are generated by the same algorithm.
Some restrictions make MPAD difficult to apply in practical solutions. A necessary condition for its implementation is the availability of titles of both references and citations. This practically limits the use of MPAD to journal papers indexed in the same citation index or citation-enhanced database. The additional problems come from the fact that the number of citations to an article is virtually infinite. As a consequence, an upper limit for the number of keywords isolated from citing articles had to be stated, leaving at the same time the possibility for the permanent accumulation of new KW-Cs in an “article profile”. This was to be managed through an automated software agent.

In building MPAD into ASEESTANT, we surmounted these and similar difficulties. Yet, many methodological problems remained. Within this study we focused on the relationship among the three types of MPAD keywords. Two interrelated questions were sought to be answered: (1) are different types of keywords summative, i.e. is it appropriate to aggregate them into a common profile for individual articles; and (2) are differences among the keyword types, if existent, interpretable, i.e. consistent with predictions drawn from MPAD.

Answers to the above questions were sought by comparing the bibliographic maps of science produced by the three different types of MPAD keywords. The observed basic similarity of maps was supposed to be an argument in favor of their summative nature, while their observed differences were expected to be in accordance with the MPAD description of the three stages of the research cycle: intellectual base, research front, and research heritage.

**Method**

**Indexing method**

Automatic generation of keywords in ASEESTANT is based on KEA: Keyphrase Extraction Algorithm (Medelyan & Witten, 2008). KEA initially identifies candidates for keyphrases by extracting n-grams from text. In the second phase, the algorithm uses machine learning techniques in order to select the most important candidates, by considering their attributes, such as frequency, inverse document frequency, position in the
text and number of words in a phrase. KEA offers possibilities for free keyphrase extraction and indexing using controlled vocabularies. However, several successive trials during the development process revealed the shortcomings of both methods. Free indexing was found to produce too large a number of divergent terms, many of which were grammatically incorrect phrases. On the other hand, assignment based on controlled vocabularies tended to isolate too small a number of terms, which in most cases turned out to be too general and imprecise.

In order to improve KEA performance, in the final stage of ASEESTANT development a new method was developed. The method, dubbed Multi-stage Indexing of Subject Headings (MISH), employs KEA’s free indexing algorithm as a first step to extract candidates for descriptors. The extracted terms are then matched against terms from some of the pre-selected international domain-specific thesauri and from two home-made CEON/CEES vocabularies. Matched multi-word terms are finally stemmed in order to collapse those with the same meaning into a single entry to avoid duplication.

The above procedure was applied to produce all three types of MPAD keywords. Only papers having titles, abstracts, and at least one reference in English could be included in the analysis, since the thesauri and vocabularies employed are all monolingual.

Data sample

The sample of papers included 13,032 papers indexed with all three types of automatically generated descriptors. All papers were taken from SCIndeks: The Serbian citation Index (Šipka, 2005; http://ceon.rs). Due to the above stated language restrictions, a majority of them came from journals covering chemistry (19%), agriculture (18%), and medicine. The sample of terms included 96,766 KWs (7.4 per paper on average), 96,537 KW-Rs (7.4 per paper), and 154,305 KW-Cs (11.8 per paper). The analysis included only KWs having a frequency higher than 17 (1,006 terms), KW-Rs with a frequency higher than 14 (1,279 terms) and KW-Cs with a frequency higher than 27 (1,030 terms).
Data analysis
Maps were generated using a combination of techniques which have proved to be appropriate for the representation of a large amount of entities (Moya-Anegón, Herrero-Solana & Jiménez-Contreras, 2006). Distance matrices calculated using the Jaccard index were created for each sample of words separately. Cluster analyses were applied on these distance matrices, with the Partitioning Around Medoids method (PAM). Silhouette and connectivity measures were employed to estimate the number of clusters (Kaufman & Rousseeuw, 1990). Clusters of terms were visually represented in two-dimensional space using multidimensional scaling (MDS). Distances among clusters on the map reflect the average distances of terms within clusters which are their best representatives. The size of the circles representing clusters is proportional to the number of terms in each of them.

Results
Number of clusters
The validity of cluster solutions estimated by the silhouette and connectivity measures is shown in Figures 1 and 2. Better solutions are indicated by higher silhouette and lower connectivity values. The results suggest that the clusters are of low quality and unclear structure. The solution giving 9 clusters in mapping KWs and KW-Cs seems to be the most appropriate. For KW-Rs, solutions resulting with 9 to 11 clusters appeared to be more interpretable than others. Among these, the ten-clusters solution was chosen for the main analysis.

Maps of science
The bibliographic maps are presented in chronological order, from the map of cited papers’ keywords (KW-Rs; Figure 3) to the map based on keywords extracted from citing papers (KW-Cs; Figure 5). The maps are generally similar both in terms of content and the structure of clusters. Natural sciences are basically reduced to two areas most heavily represented in the sample: agriculture and chemistry. Judged by the number of clusters, agriculture seems to be the most differentiated research area. Similar configurations with the predominance of agriculture and chemis-
try were already observed in the profiles of other developing countries (Schultz & Manganote, 2012). Another similarity comes from the close positions of social sciences and medicine in all three maps. Finally, all maps have failed to show some disciplines, such as mathematics, computer sciences, and earth sciences, which can normally be found on large

**Figure 1. Silhouette for diff. number of clusters**

**Figure 2. Connectivity for diff. number of clusters**
scale world-science maps (Boyack, Klavans & Börner, 2005). It should be noted, however, that the KW-Rs map corresponds slightly more closely to global maps of science, thanks to the better differentiation of basic natural sciences.

![Figure 3. Map of science described by KW-Rs](image)

Regarding the content of clusters, maps based on keywords describing the papers themselves (KWs; Figure 4) and citing papers’ keywords (KW-Cs; Figure 5) look more similar. Both these types of keywords were isolated from SCIndeks papers, published mainly by domestic, Serbian authors who share the same orientation towards findings application. In such papers more uniform and less technical terminology, typical of peripheral science, is predominant. On the other hand, the titles of cited papers condensed into KW-Rs come mainly from international publications offering a greater diversity of topics and richer vocabulary. Apart from better outlining some applied disciplines like metallurgy and earthquake estimation, the cited references (KW-Rs) map is more convincing in featuring basic disciplines, especially biology and microbiology. As for social sciences, clusters in this map have a more general and fundamental profile instead of an applied one.

The expectation based on previous studies showing that cited references reflect intellectual base seems to be at least partially confirmed. However,
the results failed to fully support the hypothesis about keywords (KW) being good descriptors of research presence (“state-of-the art”), and KW-Cs being descriptors of research results’ consequences (“research heritage”). The two cluster analyses resulted in structures too similar to reveal interpretable differences between these two stages of the research cycle. This can be attributed to the rather imitative character of research presented in SCIndeks, as well as to the lack of interdisciplinarity of research in science periphery. Apparently, the findings are too rarely applied outside their narrow disciplinary area to be recognized, when transformed into KW-Cs, as something forming a new, discrete segment of research space.

Differences among the maps in terms of the interpretability of MDS dimensions are more or less in line with the results of the cluster analyses. The KWs map somehow provided more interpretable dimensions than the other two maps. One of the dimensions, plotted on the x-axis, apparently differentiates social sciences from sciences, while the second one contrasts life sciences with the group consisting of chemistry, physics and, interestingly enough, research on society dominated by legal issues.

The basic nature of the dimensions in the other two maps seems to be quite similar. At the same time, some congruent deviations from the KWs map can be observed. The main dimension from this map, reflecting the
science-social sciences dichotomy, appeared as the y-axis on the other two maps, accounting for less variance than previously stated. Also, the second dimension is obviously shaped not only by the genuine differences among the research areas, but also by their bibliometric specificities. Polarization on this dimension seems to have occurred as a result of the activity of two dominant research groups, one of which is extremely large, open, and highly productive (agricultural researchers), while the other is rather closed and self-sufficient (mathematical chemists). The latter group is small and characterized by an extremely high self- and inter-citation rate. In the analysis based on KWs, it was not recognized as a cluster at all. More research is needed to see if its isolation, position on the map, and influence on the whole constellation of research areas is an artifact or the structural characteristic of the research space under study.

Figure 5. Map of science described by KW-Cs

Conclusions
The three bibliographic maps obtained seem to be sufficiently congruent to allow for the routine aggregation of different types of MPAD keywords, i.e. combining them into unique article profiles. Some observed differences among the maps are not clearly attributable to differences among the three stages of the research cycle, as hypothesized by MPAD and expected on the basis of previous studies about the influence of “his-
historical variation” on the research front (e.g. Lucio-Arias & Leydesdorff, 2009).

The reasons for the lack of support for MPAD might be quite different. Some hints suggest that they are methodological in nature and most likely related to sampling issues. Future studies should be concentrated on overcoming some obvious deficiencies of the present study:

1. Papers published in reputed international journals were not covered for practical reasons. This may have affected the resulting picture of science, which is reduced to its “import-oriented” part, hiding an important share capable of “findings re-export”.

2. Papers not having English titles and/or abstracts, most of which lacking English references as well, were not covered for technical reasons. This may have affected results on clustering, and consequently mapping.

3. The differences in the number of papers among various disciplines were not controlled. This probably affected the nature of isolated MDA dimensions. In a study aimed at testing a model such as MPAD, the area sampling of papers would have probably been more appropriate.

4. The differences in information pregnancy of sources used for keyword extraction was not possible to control. KW-Rs were extracted only from reference titles, while in producing KW-s and KW-Cs article abstracts were used as well.

These and other deficiencies suggest caution in accepting the findings of this study, calling for further research. They, however, do not make the results invalid, and certainly do not make MPAD unfeasible.

References


Maps of Science Based on Keywords of Articles’ Antecedences, Presences, and Consequences:
Application of the CEON/CEES Model of Multi-Perspective Description of Articles


These Proceedings are a record of the 5th Belgrade International Open Access Conference, organized by the Centre for Evaluation in Education and Science.

Centre for Evaluation in Education and Science (CEON/CEES) is a virtual institute gathering experts from various R&D organizations to work together on issues of evaluation in science and higher education. Our primary activities are focused around adjusting, developing, and implementing scientometric indicators suitable for the use in small, developing, transition, and emerging countries. An important part of CEON/CEES's program is reserved for the development of information systems to support decision making in STI sector, as well as quality research publishing based on quality control, monitoring, and evaluation.