# Author Profile Identification Using Formal Concept Analysis*

Martin Radvanský, Zdeněk Horák, Miloš Kudělka, and Václav Snášel

VSB Technical University Ostrava, Ostrava, Czech Republic
{martin.radvansky.st,zdenek.horak.st,
milos.kudelka,vaclav.snasel}@vsb.cz

**Abstract.** This paper presents results of the finding of the author's profiles using formal concepts generated from DBLP database. Our main aim was to evaluate the use of formal concept analysis as a method for extracting the author's profiles. There are several commonly used methods for clustering and for finding experts in a large database. These methods are mainly based on different kinds of clustering and metrics which are sometimes difficult to understand. Finding experts for particular and mainly special areas of research is not an easy task. Formal concept analysis (FCA) is a method with a very strong mathematical background, which makes it easy to understand. Properties of FCA can give us a very strong tool for finding author's profiles.

**Keywords:** DBLP, author profile, formal concept analysis, concept stability.

## 1 Introduction

Digital Bibliography & Library Project (DBLP) is one of the most known collections of electronic resources which can be accessed over the Internet. This project was founded in 1993 and contains, among other things, more than 1,800,000 papers. These papers come from computer science and were published in different journals and conference proceedings. Although DBLP is primarily used for finding publication in the library, this fast increasing database is often used by researchers as a good dataset for data mining tasks, such as finding experts, recommendation systems, social networks algorithms, etc. However, the DBLP contains only a limited amount of information about particular papers - there are no abstracts or index terms stored in it. On the other hand, DBLP provides a lot of information about the publication activity of authors, conferences and author relationships. There was a lot of research done to find experts, extract

* This paper has been elaborated in the framework of the IT4Innovations Centre of Excellence project, reg. no. CZ.1.05/1.1.00/02.0070 supported by Operational Programme 'Research and Development for Innovations' funded by Structural Funds of the European Union and state budget of the Czech Republic, this work was also supported by the Bio-Inspired Methods: research, development and knowledge transfer project, reg. no. CZ.1.07/2.3.00/20.0073 funded by Operational Programme Education for Competitiveness, co-financed by ESF and state budget of the Czech Republic, and partially supported by SGS, VSB-Technical University of Ostrava, Czech Republic, under the grants No. SP2012/58.

V. Snasel et al. (Eds.): SOCO Models in Industrial & Environmental Appl., AISC 188, pp. 485–494.
springerlink.com

their working areas, analyse communities in the social network based on DBLP and much more.

In this paper we have processed the DBLP in order to extract the author's profiles based on keywords used in their papers. The author's profiles can help us to find groups of keywords that were often and repeatedly used by authors in their titles of papers. The main expected result of our work is to find author's profiles. These profiles can be used for identification of experts for particular area of research.

This paper is organized as follows: Section 2 contains an overview of related work. Section 3 explains the methods used for data evaluation. Section 4 is focused on the finding of author's profiles. Last section 5 concludes the paper.

## 2 Related Work

Growing databases of documents, research papers and other document-oriented databases, during the last thirty years bring new challenges to the researchers. Many methods have been introduced that were focused on the fast searching, grouping and finding similar documents. In the following paragraphs we will review some of the most related approaches.

In [13] we can find an efficient algorithm for topic ranking. The authors show a method for the extraction of keyword sets and cluster of research papers using these keyword sets. The evolution of topics over time and their ranking is studied. Paper [4] covers the bibliometrics perspective. It investigates the frequency and impact of conference publications in computer science and compares it to journal papers. The author uses statistical methods for analysing DBLP. Paper [15] introduces alternative measures for ranking venues. They create new bibliometrics that can be used in ranking publication venues. These bibliometrics are easy to implement and bring more accuracy to the evaluation of venues. An application based on stability (a measure from formal concept analysis which is discussed later) can be found in [9]. In this paper the stability is used for pruning conceptual lattice which was constructed from the ECSC dataset. Analysis of the DBLP publication and their classification by using Concept lattices can be found in [1]. This paper shows how concept lattice can cover relational and contextual information of analysed papers.

Our approach is inspired by the previous research, but we have tried to address several issues in a different way. In this paper we try to search keywords that are used by authors periodically and frequently. This set of keywords is included in the intent of concepts. Because the set of concepts is greater than number of keywords, we have used the approach based on concept stability. This is the main difference of our approach and previously mentioned related work. In the next section we summarize used tools and techniques.

## 3 Applied Methods and Data Collection

This section provides some basic notions and techniques applied in our experiments.

### 3.1 Formal Concept Analysis

In our paper we use Formal concept analysis as a technique for unsupervised clustering. This method helped us to find non-trivial clusters of authors and their keywords. In the next paragraph we briefly describe Formal concept analysis.

Formal concept analysis (FCA) is a general data analysis method based on the lattice theory. FCA was introduced in 1982 by Wille [14]. The basic algorithms for concept lattice computation were published by Ganter in 1984 [5]. More recent publications of these founders can be found in ([6], [7], [8]). Carpineto and Romano summarized in ([2], [3]), both the mathematical and computer scientist's (with a focus on information retrieval) perspective of the FCA. A good overview of the recent state was written also by Priss in [11].

The input data for FCA is called formal context $C$, which can be described as $C = (G,M,I)$ - a triplet consisting of a set of objects $G$ and set of attributes $M$, with $I$ as relation of $G$ and $M$. The elements of $G$ are defined as objects and the elements of $M$ as attributes of the context.

As an example of using FCA we have selected five authors $\mathrm{a}_1$,..., $\mathrm{a}_5$ and five keywords that were often used by these authors in the title of their papers. These keywords are "database - $\mathrm{k}_1$", "algorithm - $\mathrm{k}_2$", "distributed - $\mathrm{k}_3$", "data mining - $\mathrm{k}_4$" and "analysis - $\mathrm{k}_5$". The relation between author and keyword is shown as a cross in the Table 1.

**Table 1.** Formal context

| | database $\mathrm{k}_1$ | algorithm $\mathrm{k}_2$ | distributed $\mathrm{k}_3$ | data mining $\mathrm{k}_4$ | analysis $\mathrm{k}_5$ |
|---|---|---|---|---|---|
| author $\mathrm{a}_1$ | × | × | × | | × |
| author $\mathrm{a}_2$ | × | × | | × | |
| author $\mathrm{a}_3$ | × | | × | | |
| author $\mathrm{a}_4$ | × | | | × | × |
| author $\mathrm{a}_5$ | | × | | × | |

Density of the formal context $(G,M,I)$ is defined as proportion of elements of $I$ with respect to the size of $GM$. The density calculated for the context depicted in the Table 1 is $56\%$.

For a set $A \subseteq G$ of objects we define $A^{\uparrow}$ as the set of attributes, common to the objects in $A$. Correspondingly, for a set $B \subseteq M$ of attributes we define $B^{\downarrow}$ as the set of objects which have all attributes in $B$. A formal concept of the context $(G,M,I)$ is a pair $(A,B)$ with $A \subseteq G$, $B \subseteq M$, $A^{\uparrow} = B \; and \; B^{\downarrow} = A$. The set $A$ is called extent of a concept, while the set $B$ is called intent of a concept. $\mathscr{B}(G,M,I)$ denotes the set of all concepts of context $(G,M,I)$ and forms a complete lattice (so-called Galois lattice). For more details see ([7], [8]). All concepts from our example are shown in the Table 2. The figure 1 depicts concept lattice of our example.

For selection of interesting concepts we have used a method based on concept stability that is described in the next section.

**Table 2.** Formal concepts extracted from context in Table 1

| concept | extent | intent |
|---|---|---|
| c(0) | $\{a_1, a_2, a_3, a_4, a_5\}$ | $\{\}$ |
| c(1) | $\{a_2, a_4, a_5\}$ | $\{k_4\}$ |
| c(2) | $\{a_1, a_2, a_5\}$ | $\{k_2\}$ |
| c(3) | $\{a_2, a_5\}$, | $\{k_2, k_4\}$ |
| c(4) | $\{a_1, a_2, a_3, a_4\}$ | $\{k_1\}$ |
| c(5) | $\{a_1, a_4\}$ | $\{k_1, k_5\}$ |
| c(6) | $\{a_2, a_4\}$ | $\{k_1, k_4\}$ |
| c(7) | $\{a_4\}$ | $\{k_1, k_4, k_5\}$ |
| c(8) | $\{a_1, a_3\}$ | $\{k_1, k_3\}$ |
| c(9) | $\{a_1, a_2\ \}$ | $\{k_1, k_2\}$ |
| c(10) | $\{a_2\}$ | $\{k_1, k_2, k_4\}$ |
| c(11) | $\{a_1\}$ | $\{k_1, k_2, k_3, k_5\}$ |
| c(12) | $\{\}$ | $\{k_1, k_2, k_3, k_4, k_5\}$ |

### 3.2 Concept Stability

The main problem of using FCA as a clustering method is that we often obtain very large and complicated structure, which is hard to understand and interpret. Technically speaking, we can get a large number of concepts even for a relatively small context. There are several methods which can be used to select only some part of concepts. We have used the so-called concept stability to filter only the interesting ones. As an interesting concept we considered a concept which is, up to a certain degree, resistant to the change of a particular object (removing particular object does not cause the change of the intent).

Stability of a concept (introduced by Kuznetsov in [9]) expresses the dependency between the intent and extent of the concept. Following the notions from [10], for a particular concept $(A,B)$ of a concept lattice $\mathscr{B}(G,M,I)$, the stability is defined as:

$$\sigma(A,B) = \frac{|\{C \subseteq A | C^{\uparrow} = B\}|}{2^{|A|}} \tag{1}$$

Higher stability causes higher immunity of concept to changes in particular objects. An efficient way to compute the stability of all concepts (using a bottom-up lattice traversal) is described in [12].

As a continuation of our example we can compute stability of concept C(8) by using Equation (1) as:

$$\sigma(\{a_1,a_3\},\{k_1,k_3\}) = \frac{2}{2^2} = \frac{1}{2} \tag{2}$$

### 3.3 Pre-processing of Data Collection

On December 12, 2011, we downloaded the DBLP dataset in XML[1] and pre-processed it for further usage. First of all, we selected journal volumes and conferences held by

[1] Available from `http://dblp.uni-trier.de/xml/`

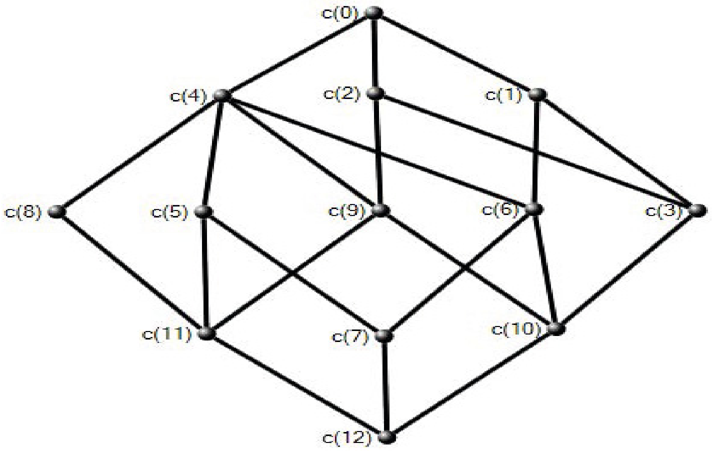

**Fig. 1.** Concept lattice created from concepts in the Table 2

IEEE, ACM and Springer. For every record we identified the month and year of the publication. In the next step, we extracted all authors having at least one published paper (11,355 authors) in a selected period. Then, we extracted keywords and phrases from paper titles. The approach was based on Faceted DBLP set[2]; 1,134 keywords and phrases we used in total.

For our paper we have selected a time period up to the year 2010 to get the most complete dataset. Then, we divided the entire recorded publication period of conferences into one-month time periods. If during one month an author has published a paper then we set keyword records, corresponding to the paper title. For each author we obtained a list of months with occurred keywords. To reduce the number of authors we have used secondary filtering based on occurrences of keywords during the selected period. From a set of 11,355 authors, we have selected only authors with more than 2 keywords in theirs papers (1,735). Figure 2 display histogram of number of keywords in author's papers.

For the following evaluation we constructed binary formal context. The rows of context represent authors of papers and columns correspond to the keywords used in particular paper in DBLP. The value of an intersection between row and column contain value "1" when author used keyword or "0" otherwise.

## 4 Searching for Author's Profiles

This section describes our method for searching author's profiles.

*Author's profile by meaning is used in this paper as a set of characteristic keywords. These keywords were used often and repeatedly by an author in his papers during an observed period. Groups of authors with the same profile, can be seen as a group of experts in the particular research area, covered by profile keywords.*

[2] `http://dblp.l3s.de/browse.php?browse=mostPopularKeywords`

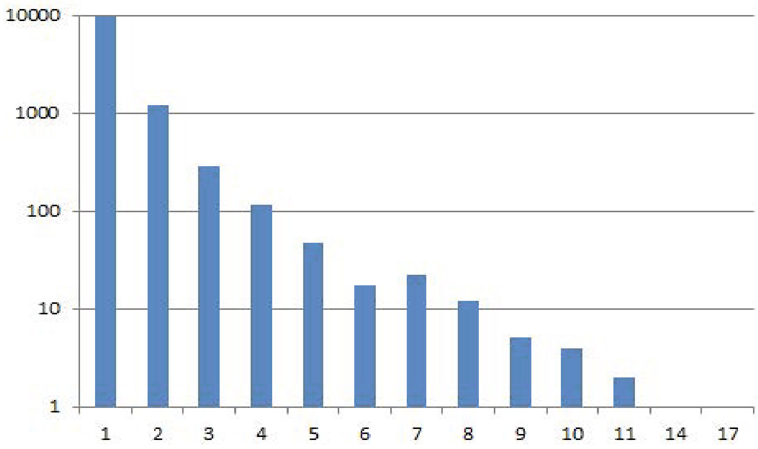

**Fig. 2.** Histogram of keywords used by authors in their papers

### 4.1 Basic Properties Data Collection

In order to create formal contexts from the described data collection, we have got a context containing 1,735 rows (authors) and 525 columns (keywords). The density of the context was about 5%. A small example of selected keywords are a set of "analysis, algorithms, applications, coding, testing, modeling, attacks, aggregation, logic, rdf, energy".

In the next section FCA is used as a main method for clustering and finding authors' profiles.

### 4.2 Finding Profiles of Authors by FCA

FCA gave us a tool for finding profiles of authors, based on the keywords they use in papers. For the context created in the previous step we have computed a concept lattice. We are interested in nontrivial concepts where author's profiles are the intents of these interesting concepts. In order to decide which concepts are interesting for us, we used the concept stability method, described in the section 3.2. This method helped us reduce the size of concept lattice and intent of concepts are more confident as the author's profile. For finding the right level of stability threshold, we have computed a number of concepts that satisfied the level of stability see figure 3.

Choosing the value of stability threshold has been done according to the basic meaning of stability (see section 3.2). Stability 0.5 gives us information that there exists one half subsets of all possible subsets of authors in the concept that has special property. Removing subset authors from concept extent (set of authors) does not cause a change in the intent (set of keywords) of concept. Higher value of concept stability makes concept more confident. We have used concept stability for pruning concept lattices. After applying pruning lattice by concept stability we obtained a relative small number of concepts which can be easily explored.

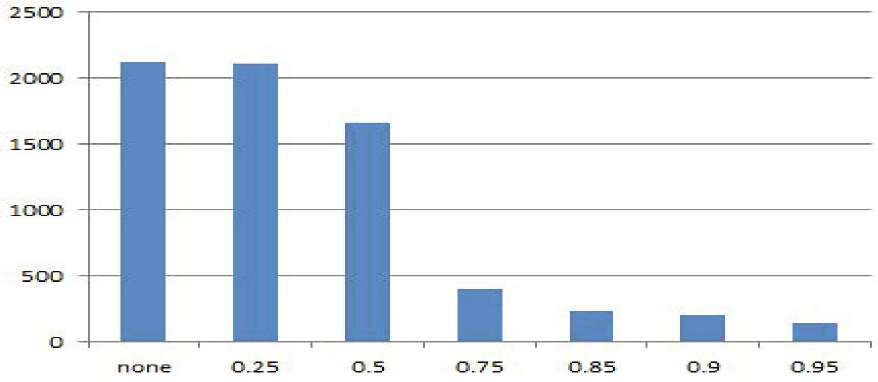

**Fig. 3.** Number of formal concepts reduced by different level of stability

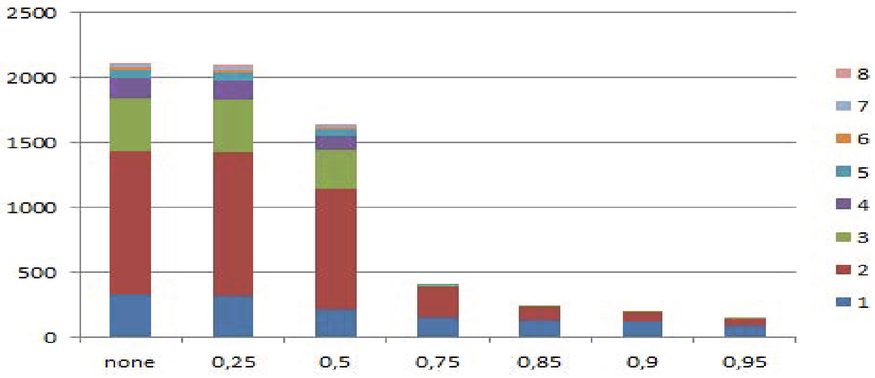

**Fig. 4.** Proportion of number of keywords for different level of stability

Figure 4 illustrates the proportion of number of keywords in the intent of computed concepts in dependence on different level of stability. According to the selected level of stability 0.5, the most interesting concepts for creating author's profiles were concepts with more than one attribute. Author's profiles are concepts that contain particular keywords in their intents.

As an example we have selected concepts having the keyword "factorization" in its intent, see Table 3.

This table depicts five concepts and theirs intents create the author's profiles. We can identify five sets of authors in the extent of these concepts who are experts in this particular area of research. In our example we have identified four authors and each of them have in his author's profile keyword "factorization" which is more confident than

**Table 3.** Formal concepts with keyword "factorization" in their intents

| id | stability | extent | intent |
|---|---|---|---|
| c(1) | 0.625 | Tao Li, Amir Shpilka, Jan Platos, Ilya Volkovich | factorization |
| c(2) | 0.5 | Amir Shpilka, Ilya Volkovich | factorization, testing |
| c(3) | 0.5 | Jan Platos | factorization, data minning, genetic algorithm |
| c(4) | 0.5 | Amir Shpilka | factorization, boolean functions, coding, finite fields, testing |
| c(5) | 0.5 | Tao Li | factorization, adaptation, algorithms, analysis, applications, clustering, detection, neural networks |

**Table 4.** Author's profiles for particular area of research

| author | keywords |
|---|---|
| Amir Sphilka | factorization (0.625), boolean functions (0.5), coding (0.5), finite fields (0.5), testing (0.5) |
| Ilya Volkovich | factorization (0.625), testing (0.5) |
| Jan Platos | factorization (0.625), data minning (0.5), genetic algorithm (0.5) |
| Tao Li | factorization (0.625), adaptation (0.5), algorithms (0.5), analysis (0.5), applications (0.5), clustering (0.5), detection (0.5), neural networks (0.5) |

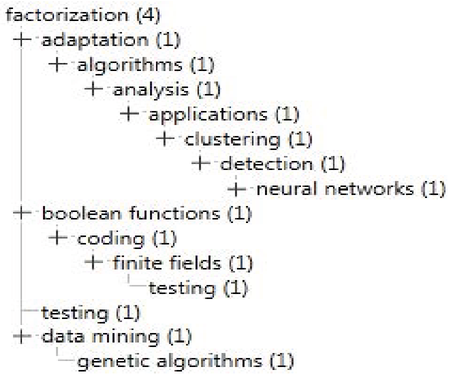

**Fig. 5.** Hierarchy of keywords connected with keyword "factorization"

other keywords. This has to be done by meaning of concept stability (0.625 > 0.5) see Table 4. On similar bases we can extend author's profile of each author by the other keywords from the set of concepts.

In the figure 5 we can see hierarchical view on the set of concepts. The numbers next to the keywords give us information of occurrences these keywords have in concepts. This hierarchical structure is connected to the underlying conceptual lattice what was pruned by concept stability.

This example of author's profiles is just a small part of the whole process for creating author's profiles. For creating a full author profile we intersect all keywords in all particular author's profiles. As a result we will get all keywords usually and often used by an author in his papers.

## 5 Conclusion and Future Work

In this paper, we have introduced an approach for finding interesting author's profiles, based on keywords used in titles of papers in the DBLP database. Using of FCA together with stability of concepts helped us to find these profiles. The used method gave us a very interesting hierarchical view of the keywords as the author's profile. In our future work we plan to take a closer look at the evolution of author's profiles during the time and the cooperation of authors based on their profiles.

## References

1. Alwahaishi, S., Martinovič, J., Snášel, V., Kudělka, M.: Analysis of the DBLP Publication Classification Using Concept Lattices. In: DATESO 2011, pp. 132–139 (2011)
2. Carpineto, C., Romano, G.: Concept Data Analysis. John Wiley and Sons, New York (2004)
3. Carpineto, C., Romano, G.: Using Concept Lattices for Text Retrieval and Mining. In: Ganter, B., Stumme, G., Wille, R. (eds.) Formal Concept Analysis. LNCS (LNAI), vol. 3626, pp. 161–179. Springer, Heidelberg (2005)
4. Franceschet, M.: The Role of Conference publications in CS. Communications of the ACM 53(12), 129–132 (2010)
5. Ganter, B.: Two Basic Algorithms in Concept Analysis. In: Kwuida, L., Sertkaya, B. (eds.) ICFCA 2010. LNCS, vol. 5986, pp. 312–340. Springer, Heidelberg (2010)
6. Ganter, B., Stumme, G., Wille, R.: Formal Concept Analysis. LNCS (LNAI), vol. 3626. Springer, Heidelberg (2005)
7. Ganter, B., Wille, R.: Applied Lattice Theory: Formal Concept Analysis. In: Grätzer, G.A. (ed.) General Lattice Theory, pp. 592–606, Birkhäuser (1997)
8. Ganter, B., Wille, R.: Formal Concept Analysis: Mathematical Foundations. Springer, Berlin (1999)
9. Kuznetsov, S.O.: On stability of a formal concept. In: Annals of Mathematics and Artificial Intelligence, vol. 49(1), pp. 101–115 (2007)
10. Kuznetsov, S.O., Obiedkov, S., Roth, C.: Reducing the Representation Complexity of Lattice-Based Taxonomies. In: Priss, U., Polovina, S., Hill, R. (eds.) ICCS 2007. LNCS (LNAI), vol. 4604, pp. 241–254. Springer, Heidelberg (2007)
11. Priss, U.: Formal concept analysis in information science. Annual Review of Information Science and Technology 40 (2006)

12. Roth, C., Obiedkov, S., Kourie, D.G.: Towards Concise Representation for Taxonomies of Epistemic Communities. In: Yahia, S.B., Nguifo, E.M., Belohlavek, R. (eds.) CLA 2006. LNCS (LNAI), vol. 4923, pp. 240–255. Springer, Heidelberg (2008)
13. Shubhankar, K., Singh, A.P., Pudi, V.: An Efficient Algorithm for Topic Ranking and Modeling Topic Evolution. In: Proceeding DEXA 2011 of the 22nd International Conference on Database and Expert Systems Applications, pp. 320–330. Springer, Berlin (2011)
14. Wille, R.: Restructuring lattice theory: an approach based on hierarchies of concepts. In: Rival, I. (ed.) Ordered Sets, pp. 445–470. Reidel, Dordrecht (1982)
15. Yan, S., Lee, D.: Toward Alternative Measures for Ranking Venues: A Case of Database Research Community. In: Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries, Vancouver, Canada, pp. 235–244 (2007)