Analysis of the DBLP Publication Classification Using Concept Lattices

Saleh Alwahaishi, Jan Martinovič, and Václav Snášel, and Miloš Kudělka

Department of Computer Science, FEECS, VŠB – Technical University of Ostrava,
17. listopadu 15, 708 33 Ostrava-Poruba, Czech Republic
{salehw, jan.martinovic, vaclav.snasel}@vsb.cz

Abstract. The definitive classification of scientific journals depends on their aim and scope details. In this paper, we present an approach to facilitate the journals classification of the DBLP datasets. For the analysis, the DBLP data sets were pre-processed by assigning each journal attributes defined by its topics. It is subsequently shown how theory of formal concept analysis can be applied to analyze the relations between journals and the extracted topics from their aims and scopes. It is shown how this approach can be used to facilitate the classifications of scientific journals.

1 Introduction

Formal Concept Analysis (FCA) was invented in the early 1980s by Rudolf Wille as a mathematical theory [1]. FCA is concerned with the formalization of concepts and conceptual thinking and has been applied in many disciplines such as software engineering, knowledge discovery and information retrieved during the last two decades. The mathematical foundation of FCA is described in [2]. In this paper, we describe how we used FCA to create a visual overview of the DBLP scientific journals classification based on their aims and scopes. As a case study, we zoom in on the top journals based on their impact factors.

FCA is a mathematical theory for concepts and concept hierarchies that reflects an understanding of "concept". It explicitly formalizes extension and intension of a concept, their mutual relationships, and the fact that increasing intent implies decreasing extent and vice versa. Based on lattice theory, it allows deriving a concept hierarchy from a given dataset. FCA is thus complementing other conceptual knowledge representations; and the combination of FCA with other representations has been the topic of many publications. For instance, several approaches combined FCA with description logics [3, 4] and with conceptual graphs [5, 6].

The remainder of this paper is composed as follows: In section 2 we introduce an overview of the Digital bibliography and Library project (DBLP). Section 3 visualizes, with an example, the literature using FCA lattice. In section 4 we explained the classification criterion of journals and applied the concept lattice on the selected journals. Section 5 concludes the paper.

V. Snášel, J. Pokorný, K. Richta (Eds.): Dateso 2011, pp. 132–139, ISBN 978-80-248-2391-1.

2 Digital Bibliography & Library Project (DBLP)

Digital libraries are collections of resources and services stored in digital formats and accessed by computers. Studying them offers an interesting case study for researches for the following reasons: Firstly, they grow quickly; secondly, they represent a multidisciplinary domain which has attracted researchers from a wide area of expertise. DBLP (Digital Bibliography & Library Project) is a computer science bibliography database hosted at University of Trier, in Germany.

It was started at the end of 1993 and listed more than one million articles on computer science in January 2010. These articles were published in Journals such as VLDB, the IEEE and the ACM Transactions and Conference proceedings [7, 8]. Besides DBLP has been a credible resource for finding publications, its dataset has been widely investigated in a number of studies related to data mining and social networks to solve different tasks such as recommender systems, experts finding, name ambiguity, etc. Even though, DBLP dataset provides abundant information about author relationships, conferences, and scientific communities it has a major limitation that is its records provide only the paper title without the abstract and index terms.

In addition to using the DBLP dataset for finding academic experts, it has been used extensively in academic recommender systems. A number of studies were conducted to recommend academic events and collaborators for researchers using different methods and techniques. For example, a recommender system for academic collaboration called DBconnect was presented in [9]. Authors of this paper used DBLP data to generate bipartite (author-conference) and tripartite (authorconference-topics) graph models, and designed a random walk algorithm for these models to calculate the relevance score between authors. And in another study [10] a recommender system for events and scientific communities for researchers was proposed based on social network analysis.

Querying large datasets produces large sets too, which makes the user unable to decide from where he has to start looking at the results. To solve this problem clustering and ranking were suggested in many papers. A system to visualize author information and relationships simultaneously was presented in [11]. The authors applied two types of clustering, keyword clustering and author clustering to visualize the relationships and groupings of authors. In [12] document clustering was applied to provide an overview of the recent trends in data mining activities. Clustering and ranking are often applied separately but in [13] a novel framework called RankClus was proposed to integrate them. To increase the accuracy of IR clustering, the authors in [14] proposed transferring knowledge available on the word side to the document side; they introduced a model based on nonnegative matrix factorization to achieve it.

3 Concept Analysis Of Journals Classification

This section describes how formal concept analysis is employed to analyze the DBLP's journals classification. Formal Concept Analysis can be used as an unsupervised clustering technique. The starting point of the analysis is a database table consisting of rows G (i.e. objects), columns M (i.e. attributes) and crosses I

$\subseteq G \times M$ (i.e. relationships between objects and attributes). The mathematical structure used to reference such a cross table is called a formal context (G, M, I).

A group of interested similar journals, which covered the scope of computer science, were selected. The list of selected journals (objects) was obtained from well-known DBLP database that contains information about the published articles and their authors as well. The selected list of links to journals has the size of 115 items. The next step was to identify main topics (attributes), which each of the journals covers. From the journal web sites we have found the aim and scope of each journal, and have manually extracted the main topics, such as Pattern Recognition, Image Processing, etc. Each journal has been identified by an existing classifier by company due to the problem with using their own names or similar names of topics. The used classifier that contains about 1224 sub disciplines classified to disciplines and those classified to discipline field, e.g. sub discipline Pattern Recognition is in disciplines Artificial Intelligence and Image Processing and that is in Information and computing sciences [15]. We selected only sub disciplines in the field Technology and Information and computing sciences. Our manually extracted topic from journals in many cases correspond the classified disciplines, but in some cases it was necessary to assign the extracted topic to sub discipline, which was almost similar. Therefore, journals were classified into a list of topics based in their relation to the topic. The classification process ends up with ten main topics that have twenty nine subfields or disciplines. Table 2 shows the main topics and their subfields.

A journal is represented as a list of topics. The topics are the disciplines that being covered by all journals, based on the extracted data from their aims and scopes. Each topic is assigned a weight of 0 or 1. A topic's weight for a journal expresses the coverage possibility of the topic by the related journal. A value of 1 denotes that the journal covers the column's topic and 0 denotes the lack of coverage. Formally, these data can be represented as a matrix of journals by topics whose m rows and n columns correspond to m journals and n topics, respectively. The elements of the journal-topic matrix are the weights of each term for a particular document, that is:

$$\boldsymbol{Y} = \begin{bmatrix} y_{11} & \cdots & y_{1n} \\ \vdots & & \vdots \\ y_{m1} & \cdots & y_{mn} \end{bmatrix}$$

Where y_{ij} denotes the weight assigned to topic T_j for journal J_i.

The formal concept analysis of the data starts with the creation of a formal context. The formal objects of the formal context are the journals J_i that were retrieved from DBLP database. The set of these journals is denoted by J. Using the information that was extracted from the aim and scope of the journals in J. The coverage possibility T_j that shows the topic coverage by the journals in J, constitute the formal attributes of the formal context. The set containing these attributes is denoted by T.

The cross table of the resulting formal context has a row for each journals in J, a column for each topic in T and a cross in the row of J_i and the column of T_j if the corresponding weight y_{ij} is 1. To minimize the cross table size, journals impact factors will be considered to decrease the number of tested journals. The journals with an

impact factor of 3.0 and above will be enlisted in the matrix, dropping the number of selected journals to be 18 as shown in Table 3. After the formal context is constructed, formal concept analysis is applied to produce the concept lattice.

Table 4 represents the formal context. A cross in the row of J_i and the column of T_j indicates that T_j is believed to be a covered topic by the journal of J_i.

Table 1. Journals' impact factors and abbreviations

Abbreviation	Journal	Impact Factor
A	Nucleic Acids Research	6.878
B	IEEE Transactions on Pattern Analysis and Machine Intelligence	5.96
C	International Journal of Computer Vision	5.358
D	Computer Applications in the Biosciences	4.328
E	Journal of Selected Areas in Communications	4.249
F	Transactions on Medical Imaging	4.004
G	Transactions on Information Theory	3.793
H	BMC Bioinformatics	3.78
I	Transactions on Neural Networks	3.726
J	Journal of Chemical Information and Computer Sciences	3.643
K	Transactions on Fuzzy Systems	3.624
L	Journal of Computational Chemistry	3.39
M	Transactions on Graphics	3.383
N	Transactions on Mobile Computing	3.352
O	Transactions on Image Processing	3.315
P	Pattern Recognition	3.279
Q	Automatica	3.178
R	Information Sciences	3.095

The intent of each formal concept contains precisely those topics covered by all journals in the extent. Conversely, the extent contains precisely those journals sharing all topics in the intent.

The line diagram of the concept lattice, showing the partially ordered set of concepts is shown in Fig 1, has the minimal set of edges necessary; all other edges can be derived by using reflexivity and transitivity. Journals and topics label the node that represents the formal concept they generate. All concept nodes above a node labeled by a journal have the journal in their extent. All concept nodes below a node labeled by a topic have the topic in their intent. The extent of the concept node labeled by the topic "STVV" for example is easily found by collecting the journal H labeling this concept node on a path going downward.

Table 2. Formal context

	AIIP	CTM	CS	ISLIS	DF	DC	CT	CA	DIP	STVV
A		x	x							
B	x							x		
C	x			x						
D		x	x							
E							x			
F	x	x	x					x		
G	x	x		x	x		x			
H		x	x							x
I			x							
J	x	x		x						
K	x									
L		x	x							
M	x									
N			x			x	x			
O	x				x					
P	x									
Q	x	x	x	x				x	x	
R	x	x	x		x		x	x		

The intent of this concept is found by first collecting the topic “STVV” and by going upward to collect the topic “CTM”, and “CS” labeling the two concepts found on paths going upward. The resulting extent-intent pair of this concept is ({H}, {CTM,CS,STVV}).

The concept generated by the topic “DF” is a sub concept of the concept generated by the topic “AIIP”, for the extent of the former concept is contained in the extent of the latter concept. All journals classified by the topic “DF” were also classified by the topic “AIIP”, suggesting that within the given formal context “DF” is a more specific topic than “AIIP”.

Another multi constructed example is found in the extent of the concept node labeled by the topic “DIP”, which is found by collecting the journal Q labeling this concept node on a path going downward. The intent of this concept is found by collecting the topics “CTM”, “CA”, and “ISLIS” labeling the three concepts found on paths going upward. The latter two topics, however, are sub concepts of the concept generated by the topic “AIIP”. The resulting extent-intent pair of this concept is ({Q}, {AIIP,CTM,CS,ISLIS,CA,DIP}).

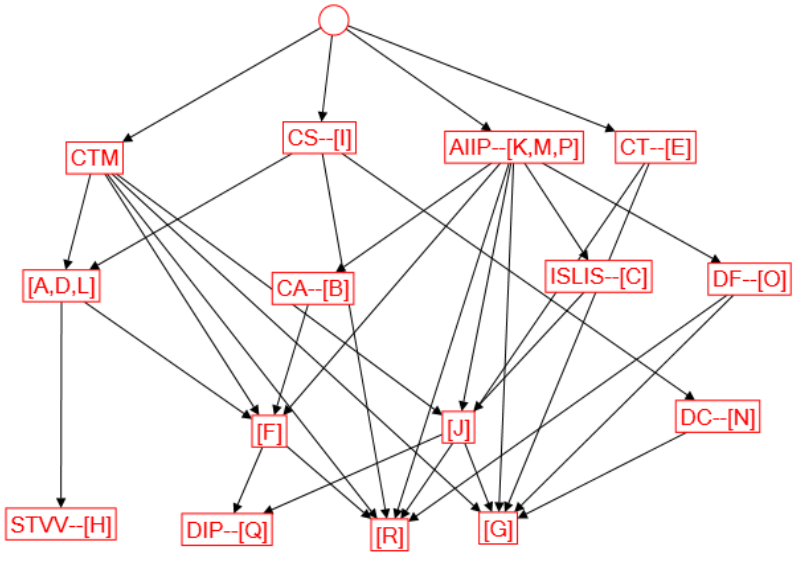

Fig. 1. Concept lattice for journals classification

4 Conclusion

The concept lattice uncovers relational and contextual information. Journals' topic categorizations are put into relational context depending on how they are associated by the journals' aims and scopes. The topics "Computer Theory and Mathematics – CTM", and "Data and Information Processing –DIP" for example are shown as related because these topics share a similar classification context. The implicit structures revealed help researchers to classify journals more efficiently. This approach has the potential to support the emergence of new knowledge by identifying concept relations, making these explicit and enabling researchers to inspect these concept relations.

Concept lattices are not intended to build or substitute traditional static ontologies, rather they aim to support specifications of less rigorous relations, or associations [16], which might be more intuitive to knowledge workers and lead to more interesting links via associations.

Abbreviation	Main Topic	Subfields
AIIP	Artificial Intelligence and Image Processing	Adaptive Agents and Intelligent Robotics, Neural, Evolutionary and Fuzzy Computation, Simulation and Modeling, Computer Vision Pattern Recognition and Data Mining, Signal processing, Image Processing
CTM	Computation Theory and Mathematics	Computer Graphics
		Other Computation Theory and Mathematics
		Numerical Computation
		Applied Discrete Mathematics
		Computational Logic and Formal Languages
		Analysis of Algorithms and Complexity
CS	Computer Software	Software Engineering
		Operating Systems
		Computer System Security
		Bioinformatics Software
ISLIS	Information Systems and Library and Information Studies	Database and Database Management
		Information Retrieval and Web Search
		Inter-organizational Information Systems and Web Services
		Information Systems Management
		Information Systems Development Methodologies
DF	Data Format	Data Encryption
		Data Structures
DC	Distributed Computing	Mobile Technologies
		Distributed Computing
CT	Communications Technologies	Computer Communications Networks (computer network)
		Wireless Communications
		Other Communications Technologies (telecommunications)
CA	Computer Architecture	
DIP	Data and Information Processing	
STVV	Software Testing and Verification & Validation	

Table 3. Journals' main topics and subfields

5 References

1. Wille. R. (1982). Restructuring lattice theory: an approach based on hierarchies of concepts. In I. Rival (Ed.). Ordered sets. Reidel. Dordrecht-Boston. 445-470.
2. Ganter, B., Wille, R. (1999) Formal Concept Analysis: Mathematical foundations. Springer
3. Beydoun, G. (2009) Using Formal Concept Analysis towards Cooperative E-Learning. D. Richards and B.H. Kang (Eds.): PKAW, LNAI 5465, 109-117. Springer
4. Priss, U. (2006), Formal Concept Analysis in Information Science. Cronin. Blaise (ed.). Annual Review of Information Science and Technology, ASIST, Vol. 40.
5. Ganter, B., Kuznetsov, S.O. (2008) Scale Coarsening as Feature Selection. Medina and S. Obiedkov (Eds.) : ICFCA, LNAI 4933, 217-228. Springer.
6. Stumme, G., Wille, R., Wille, U. (1998) Conceptual knowledge discovery in databases using Formal Concept Analysis Methods. PKDD, 450-458.
7. Ley, M. (2002) The dblp computer science bibliography: Evolution, research issues, perspectives. SPIRE 2002: Proceedings of the 9th International Symposium on String Processing and Information Retrieval. London, UK: Springer-Verlag, pp. 1–10.
8. URL, http://en.wikipedia.org/wiki/DBLP.
9. Zaiane ,O. R. Chen, J. and Goebel, R. (2007) Dbconnect: mining research community on dblp data. WebKDD/SNA-KDD '07: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis. New York, NY, USA: ACM, pp. 74–81.
10. R. Klamma, P. M. Cuong, and Y. Cao (2009) You never walk alone: Recommending academic events based on social network analysis. Complex (1), pp. 657–670.
11. Chan, S. Pon, R. and Cardenas, A. (2006) Visualization and clustering of author social networks. Distributed Multimedia Systems Conference, pp. 174–180.
12. Peng, Y. Kou, G. and Shi, Y. (2006) Recent trends in data mining: Document clustering of dm publications. International Conference on Service Systems and Service Management, vol. 2, pp. 1653–1659.
13. Y. Sun, J. Han, P. Zhao, Z. Yin, H. Cheng, and T. Wu, "Rankclus: integrating clustering with ranking for heterogeneous information network analysis," in EDBT '09: Proceedings of the 12th International Conference on Extending Database Technology. New York, NY, USA: ACM, 2009, pp. 565–576.
14. T. Li, C. Ding, Y. Zhang, and B. Shao, "Knowledge transformation from word space to document space," in SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval. New York, NY, USA: ACM, 2008, pp. 187–194.
15. Obadi G., Drazdilova P., Hlavacek L., Martinovic J., and Snasel V. (2010) A Tolerance Rough Set Based Overlapping Clustering for the DBLP Data, Web Intelligence and Intelligent Agent Technology, IEEE/WIC/ACM International Conference on, pp. 57-60, 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology.
16. Krohn U, Davies NJ, Weeks, R. (1999) Concept lattices for knowledge management. BT Technology Journal, 17(4):108-116.