On Different Aspects of Usage of One-Sided Concept Lattices in Data Analysis

Peter BUTKA

Technical University of Košice
Faculty of Electrical Engineering and Informatics
Department of Cybernetics and Artificial Intelligence
Letná 9, 040 01 Košice, Slovakia
peter.butka@tuke.sk

Abstract. This paper summarizes the work related to the usage of one-sided concept lattices in different data analysis tasks. During the work on the project we have analysed several aspects of this problem, from the theoretical contemplations to the practical problems of their implementations. Concept lattice is hierarchical structure which organizes studied objects into similar groups using the theory known as Formal Concept Analysis, which is used as conceptual data analysis technique applicable in different areas like text-mining, information retrieval, decision support systems, knowledge management, etc. In this paper the summary of papers on several studied aspects is presented, especially on the theory of generalized version of one-sided concept lattices, its standard and sparse-based implementations, or its practical aspects like usage in portal-based data mining software or hybrid information retrieval system.

1. Introduction

The growing needs for the analysis of data on web brings up the new challenges to the area of data mining. One of the conceptual data mining method is called Formal Concept Analysis (FCA, [13]), which is a theory of data analysis for identification of conceptual structures among data sets and has been found useful in data/text mining, knowledge discovery, information retrieval, or business intelligence.

FCA produces concept lattice among the data that can be understand as knowledge-based model or concept hierarchy. Standard FCA works in binary (crisp) case, which is based on binary data tables (object has/has not attribute). Due to fact that many data tables usually contain different types of attributes, some fuzzification of classic crisp method is needed. The generalization of such approaches based on the Galois connections was presented in [14] (another extension in [15]). This approach was used as a basic for the development of our model known as Generalized One-Sided Concept Lattices (GOSCL), which provide the cornerstone for all works summarized in this paper within the next section.

2. Studies on different aspects of generalized one-sided concept lattices

2.1 Theoretical papers on GOCSL

The notion of Generalized One-Sided Concept Lattices was introduced in [1] and is based on the special (one-sided) case of generalized fuzzy concept lattices previously described in [14]. In this basic paper we have provided a generalization of one-sided concept lattices based on Galois connections. Our approach allows analysis of object-attribute models with different structures for truth values of attributes. Moreover, we proved that this method of creating one-sided concept lattices is the most general one, i.e., with respect to the set of admissible formal contexts, it produces all Galois connections between power sets and the products of complete lattices. Also, the incremental algorithm for the creation of such models (called also GOSCL) was firstly presented there. More details on its implementation in Java, together with some example from the real dataset were presented in [4].

In [3] we described the problem of the limitations of crisp approaches and how the usage of GOSCL can be helpful for analysis of different (and also complex user-defined) attributes within one data table, while the interpretation of concept hierarchy is same as in classical FCA (without the need for specific unified preprocessing to binary case). Then, in [9] we have described also how the GOSCL approach can be helpful in joining of existing simple object-attribute models created by different non-generalized approaches.

Another interesting result related to GOSCL is presented in [11], where we described the closure operator, which corresponds to the intersection of particular closure systems obtained from various object-attribute models with different types of attributes. Each particular closure system is defined via one-sided concept lattices approaches applied for particular data tables, which are then combined as intersection of particular concepts in result set of concepts (concept lattice).

The last work mentioned here is [8], where the relationship between one-sided concept lattices and their restricted version was described, i.e., we presented the approach (and algorithm) for generation of one-sided concept lattices from restricted formal context (or we can call them also projections), which are obtained from given one by omitting some subset of the set of all attributes.

2.2 Time complexity of GOSCL on different data tables

One of the problems of the methods for construction of one-sided concept lattices is computational complexity, i.e., we have decided to study several aspects of GOSCL complexity. In [5] we have shown that for fixed input data table and attributes time complexity of GOSCL is asymptotically linear with the increasing number of objects, i.e., after some time (which is specific for the input context) new concepts are added linear to the increment of objects. Moreover, in [6] we have analysed the reduction of computation times of the algorithm for sparse input data tables. The main aim of this paper was to show experimental study that analyzes the influence of the sparseness of the data tables on the computation times. Sparse data tables are input object-attribute models for which many of their values are equal to zero or bottom element of attribute lattice. In our analysis we showed the significant reduction of computation times with the higher sparseness of input data.

In [7] we did another step for reduction of computation times - implementation of algorithm specifically for sparse inputs, i.e., we have implemented sparse-based representation of data input vectors and also sparse-based versions of some critical functions in GOSCL algorithm. In this paper we shown the comparison of computation times between standard and sparse-based implementation, which proved another possible reduction of computation times for sparse input data tables. Another way for reduction of computation times is presented in [2], where we have designed distributed version of GOSCL algorithm based on the decomposition of input context to several subtables with the separated (and disjoint) subsets of rows, i.e., data table is decomposed to smaller tables (using recursive bisection-based method), for which small lattices are created and these are iteratively combined (by defined merging procedure) to one large concept lattice for the whole data table.

2.3 Practical aspects of GOSCL usage in different tasks Subchapter

From the practical aspects we recall papers related to design of information retrieval (IR) system and the usage of GOSCL as data analysis module in portal-based solution.

In [10] we proposed the usage of conceptual models based on GOSCL, which are locally created for subsets of documents represented by object-attribute table (document-term table in case of vector representation of text documents). Consequently, these local concept lattices are combined to one merged model using agglomerative clustering algorithm based on the descriptive (keyword-based) representation of particular lattices. Finally, we defined basic details and methods of IR system that combines standard full-text search and conceptual search based on the extracted conceptual GOSCL-like model.

In [12] we described the use of cloud computing platform for support of distributed creation of conceptual models based on the GOSCL model. This extension also proposed a solution for creation of GOSCL models, which can be searched in analytical tool (within specific module incorporated in portal-based data mining tool) by data analyst. Cloud infrastructure is then used for increase of computational effectiveness, because particular models are built in parallel/distributed way.

3. Conclusions

In this paper the summary of works related to the theory and applications of one-sided concept lattices is presented. In the future we would like to follow the research in both theoretical and practical way with the main aim for introduction of new methods for the fields of data mining, text mining, information retrieval and business intelligence.

Acknowledgement: This work was supported by the Slovak Research and Development Agency under contract APVV-0208-10 (Tradice project), by the Slovak VEGA grant No. 1/1147/12, by the Slovak KEGA grant No. 065TUKE-4/2011, and also by the Development of the Center of Information and Communication Technologies for Knowledge Systems (ITMS project code: 26220120030). I also would like to thank to all my co-authors of the mentioned works and to their funding projects, which are mentioned in original papers.

References

to other papers publishing the results that are summarized here

[1] Butka, P., Pócs, J.: Generalization of one-sided concept lattices. *Computing and Informatics*, (2013), vol. 32, no. 2, pp. 355–370.

[2] Butka, P., Pócs, J., Pócsová, J.: Distributed version of algorithm for generalized one-sided concept lattices. *Studies in Computational Intelligence*, vol. 511, (2014), pp. 119–129.

[3] Butka, P., Pócs, J., Pócsová, J.: Use of concept lattices for data tables with different types of attributes. *Journal of Information and Organizational Sciences*, vol. 36, no. 2, (2012), pp. 1–12.

[4] Butka, P., Pócsová, Pócs, J.: Design and Implementation of Incremental Algorithm for Creation of Generalized One-Sided Concept Lattices. In: *Proc. of the 12th IEEE International Symposium on Computational Intelligence and Informatics*, IEEE, (2011), pp. 373–378.

[5] Butka, P., Pócsová, Pócs, J.: On Some Complexity Aspects of Generalized One-Sided Concept Lattices Algorithm. In: *Proc. of the 10th IEEE Jubilee International Symposium on Applied Machine Intelligence and Informatics*, IEEE, (2012), pp. 231–236.

[6] Butka, P., Pócsová, Pócs, J.: Experimental Study on Time Complexity of GOSCL Algorithm for Sparse Data Tables. In: *Proc. of the 7th IEEE International Symposium on Applied Computational Intelligence and Informatics*, IEEE, (2012), pp. 101–106.

[7] Butka, P., Pócsová, Pócs, J.: Comparison of Standard and Sparse-based Implementation of GOSCL Algorithm. In: *Proc. of the 13th IEEE International Symposium on Computational Intelligence and Informatics*, IEEE, (2012), pp. 67–71.

[8] Butka, P., Pócsová, Pócs, J.: On generation of one-sided concept lattices from restricted context. In: Proc. of the IEEE 10th Jubilee International Symposium on Intelligent Systems and Informatics, IEEE, (2012), pp. 111–115.

[9] Butka, P., Pócsová, Pócs, J.: Use of Generalized One-sided FCA Approach for Joining of Simple Object-attribute Models. In: *Proc. of the 8th IEEE International Symposium on Applied Computational Intelligence and Informatics*, IEEE, (2013), pp. 331–336.

[10] Butka, P., Pócsová, Pócs, J.: A Proposal of the Information Retrieval System Based on the Generalized One-Sided Concept Lattices. In *Applied Computational Intelligence in Engineering and Information Technology*, Springer, Berlin, (2012), pp. 59–70.

[11] Butka, P., Pócs, J., Pócsová, J., Sarnovský, M.: Multiple data tables processing via one-sided concept lattices. *Advances in Intelligent Systems and Computing*, (2013), vol. 183, pp. 89–98.

[12] Sarnovský, M., Butka, P., Pócsová, J.: Cloud computing as a platform for distributed fuzzy FCA approach in data analysis. In: *Proc. of the IEEE 16th International Conference on Intelligent Engineering Systems*, IEEE, (2012), pp. 291–296.

Other references

[13] Ganter, B., Wille, R.: *Formal Concept Analysiss*. Springer, Berlin, (1999).

[14] Pócs, J.: Note on Generating Fuzzy Concept Lattices via Galois Connections. *Information Sciences*, (2012), vol. 185, no. 1, pp. 128–136.

[15] Pócs, J.: On possible generalization of fuzzy concept lattices using dually isomorphic retracts. *Information Sciences*, (2012), vol. 210, pp. 89–98.