Framework for Conceptual Modeling on Natural Language Texts

Mikhail Bogatyrev, Kirill Samodurov

Tula State University, Tula, Russia

okkambo@mail.ru, zmeymc@gmail.com

Abstract. The paper presents the framework for conceptual modeling which has been used in on-going project of developing fact extraction technology on textual data. The modeling technique combines the usage of conceptual graphs and Formal Concept Analysis. Conceptual graphs serve as semantic models of text sentences and the data source for formal context of concept lattice. Several ways of creating formal contexts on a set of conceptual graphs have been investigated and resulting solution is proposed. It is based on the analysis of the use cases of semantic roles applied in conceptual graphs and their structural patterns. Concept lattice building on textual data is interpreted as storage of facts which can be extracted by using navigation in the lattice and interpretation its concepts and hierarchical links between them. Experimental investigation of the modeling technique was performed on the annotated textual corpus consisted of descriptions of biotopes of bacteria.

Keywords: conceptual modeling, conceptual graphs, concept lattice, biotopes of bacteria.

1 Introduction

Conceptual modeling in the Natural Language Processing (NLP) is a way of modeling semantics. Semantics of texts is transformed to semantics of conceptual models at a high level of abstraction, in terms of concepts. Conceptual graphs (CGs) [22] represent a well-known type of conceptual models and there are some applications of them in Text Mining problems solutions [13, 15].

Another paradigm of conceptual modeling is Formal Concept Analysis [10]. It is a mathematical theory of data analysis which studies how objects can be hierarchically grouped together according to their common attributes. Strong mathematical background of FCA (it is based on the lattice theory [2] and uses matrix model of so named “formal context”) provides its implementations as rigorous instrument for

Information Retrieval (IR). The number of FCA applications now is growing up including applications in Text Mining and linguistics [6, 19]. It is also applied in more general field of knowledge processing [17].

The idea of joining two paradigms of conceptual modeling - conceptual graphs and concept lattices - seems very attractive but not elaborated in the FCA community. There are several its realizations due the years, from early implementation in [23] up to recent investigations in [9].

This idea may get a second breath when FCA is utilized on textual data and conceptual graphs serve as conceptual model of text semantics. Acquiring conceptual graphs from natural language texts is non-trivial problem but it is quite solvable [5, 14]. The concepts of conceptual graphs may be treated as objects and attributes for formal context as far as the "attribute" conceptual relation really exists in conceptual graphs acquired from natural language texts. Actually, as it is followed from our investigations, the "attribute" relation is not always good and even enough for formal context. Except the "attribute" conceptual relation some other relations must be analyzed in conceptual graphs to find objects and attributes needed for formal context.

The main problem which arises in CGs – FCA applications is the problem of building formal concepts on conceptual graphs. Solution of this problem and the whole principle of applying FCA on textual data are closely depended on the real-life problems have been solved with FCA on textual data [12, 16]. In the sense of Information Retrieval these problems may be generalized to the fact extraction problem. Using FCA in its solution is based on that concept lattice built on textual data may be interpreted as storage of facts which can be extracted by using navigation in the lattice and interpretation its concepts and hierarchical links between them.

One of the fields where Text Mining applications are growing rapidly is Bioinformatics. New term of Biomedical Natural Language Processing (BioNLP) has been appeared there [1]. This is stipulated by huge amount of scientific publications in Bioinformatics and organizing them into corpora with access to the full texts of articles. FCA has great potential to take up a challenge from such areas as BioNLP.

In this paper we present the framework for conceptual modeling which has been used in on-going project of developing fact extraction technology on textual data.

The next section of the paper contains brief description of FCA basics and conceptual modeling technique which is used in the framework.

Section 3 is devoted to the framework; its structure and functionality are described there.

In the section 4 current experimental results of using framework on bacteria biotope textual corpus are presented and section 5 contains conclusion and planning future works.

2 CGs – FCA modeling on natural language texts

We are developing conceptual modeling technique which combines the usage of conceptual graphs and conceptual lattices from Formal Concept Analysis. Consider some FCA basics needed for understanding the modeling technique.

2.1 Formal Concept Analysis basics

There are two basic notions FCA deals with: *formal context* and *concept lattice*. Formal context is a triple $\mathbf{K} = (G, M, I)$, where G is a set of objects, M – set of their attributes, $I \subseteq G \times M$ – binary relation which represents facts of belonging attributes to objects. The sets G and M are partially ordered by relations ϕ and $-$, correspondingly: $G = (G, \phi)$, $M = (M, -)$. Formal context may be represented by [0, 1] - matrix $\mathbf{K} = \{k_{i,j}\}$ in which units mark correspondence between objects $g_i \in G$ and attributes $m_j \in M$. The concepts in the formal context have been determined by the following way. If for subsets of objects $A \subseteq G$ and attributes $B \subseteq M$ there are exist mappings (which may be functions also) $A' : A \to B$ and $B' : B \to A$ [1] with properties of $A' := \{\exists m \in M \mid < g, m > \in I \ \forall g \in A\}$ and $B' := \{\exists g \in G \mid < g, m > \in I \ \forall m \in B\}$ then the pair (A, B) that $A' = B$, $B' = A$ is named as formal concept. The sets A and B are closed by composition of mappings: $A'' = A$, $B'' = B$; A and B is called the *extent* and the *intent* of a formal context $\mathbf{K} = (G, M, I)$ respectively.

A formal concept is a pair (A, B) of subsets of objects and attributes which are connected so that every object in A has every attribute in B, for every object in G that is not in A, there is an attribute in B that the object does not have and for every attribute in M that is not in B, there is an object in A that does not have that attribute.

The partial orders established by relations ϕ and $-$ on the set G and M induce a partial order $\leq$ on the set of formal concepts. If for formal concepts (A_1, B_1) and (A_2, B_2), $A_1 \phi A_2$ and $B_2 - B_1$ then $(A_1, B_1) \leq (A_2, B_2)$ and formal concept (A_1, B_1) is less general than (A_2, B_2). This order is represented by *concept lattice*. A lattice consists of a partially ordered set in which every two elements have a unique *supremum* (also called a least upper bound or *join*) and a unique *infimum* (also called a greatest lower bound or *meet*).

According to the central theorem of FCA [10], a collection of all formal concepts in the context $\mathbf{K} = (G, M, I)$ with subconcept-superconcept ordering $\leq$ constitutes the *concept lattice* of $\mathbf{K}$. Its concepts are subsets of objects and attributes connected each other by mappings A', B' and ordered by a subconcept-superconcept relation.
Although that level of abstraction makes FCA suitable for use with data of any nature, its application to specific data often requires special investigation. It is fully relevant for using FCA with textual data.

---

[1] More rigorous definition assumes that these mappings are different: $\varphi : A \to B, \psi : B \to A$ but it is not a matter of principle here.

2.2 FCA on textual data

The main problem in applying FCA on textual data is the problem of building formal context. If textual data is represented as natural language texts then this problem becomes especially important.

There are several approaches to the construction of formal contexts on the textual data, presented as separate documents, as data corpora. One, mostly applied variant of context is that its objects are text documents and the attributes are the terms in these documents [6, 7]. The main problem which can be solved with that formal context and concept lattice is the problem of retrieving textual documents.

Another variant of formal context is building directly on the texts. In the general case, various word combinations constitute its concepts and the number of such concepts may be very large. An advantage of such variant is that this context contains potentially more information about texts than previous one and more general problems such as fact extraction problem can be solved on that formal context. The disadvantage of it is its great dimension and possible many pointless concepts.

Restricting the dimension of formal context and giving it more semantics is doing by representing in it the various features of its source texts: semantic relations (synonymy, hyponymy, hypernymy) in a set of words for semantic matching [12], verb-object dependencies from texts [7], words and their lexico-syntactic contexts [16].

For building formal context, one needs to distinguish some of these lexical elements in texts as objects and attributes. There are following approaches to solve this problem:

- adding special descriptions to texts which mark objects and attributes and partial order – this is usually done manually;
- using semantic models of texts and corpus tagging [7].

We apply the second approach and use conceptual graphs for representing semantics of individual sentences of a text.

2.3 CGs – FCA modeling process

The whole process of CGs – FCA modeling has the following steps.

1. *Acquiring a set of conceptual graphs from input texts*. Conceptual graph [22] is bipartite directed graph having two types of vertices: concepts and conceptual relations. These vertices are connected by arrows representing binary relations. Conceptual graphs can be created by our tool CGs Maker [2]. Some details about it can be found in [13, 14].

2. *Aggregating the set of conceptual graphs*. Aggregation is needed to exclude excessive dimension of conceptual models, not related to useful information. We have tested two ways of conceptual graphs aggregation: conceptual graphs clustering and restricting the number of conceptual graphs by identifying and excluding sentences which are not corresponded to the problem solving with the current technique.

[2] The lightweight online version of CGs Maker for simple English and Russian texts can be found at http://85.142.138.156:8888 .

3. *Creating formal contexts*. One or several formal contexts are built on the aggregated conceptual graphs. The number of formal concepts and the method of building them have been determined in the solving problem.

4. *Building concept lattice*. Having a concept lattice, it is possible to identify connections between the concepts according to the principle of "common – particular". Each concept, the node in the lattice is interpreted as the set of potential facts of certain level, which is associated with other facts.

5. *Fact extraction from concept lattice*. Concept lattice is the data storage for fact extraction system. This system has domain oriented user interface for query processing and generating output.

This paper reflects results of investigations corresponded to steps 1-3 of the process. On the step 4 we used standard open source tool for building and visualizing concept lattices [8] which we integrated into the whole modeling system. Creating the fact extraction system (step 5) is separate problem currently being under development.

2.4 Usage of conceptual graphs

The crucial step in the described process of CGs – FCA modeling is creating formal contexts on the set of conceptual graphs. At first glance, this problem has simple solution: those concepts which are connected by "attribute" relation have been put into formal context as its objects and attributes. Actually the solution is much more complex. To illustrate it consider conceptual graph for the sentence "*Xylella fastidiosa is a gram-negative fastidious, xylem-limited bacterium*" shown on Fig. 1. This sentence is from bacteria biotopes textual corpus [4] which we use for our method evaluation.

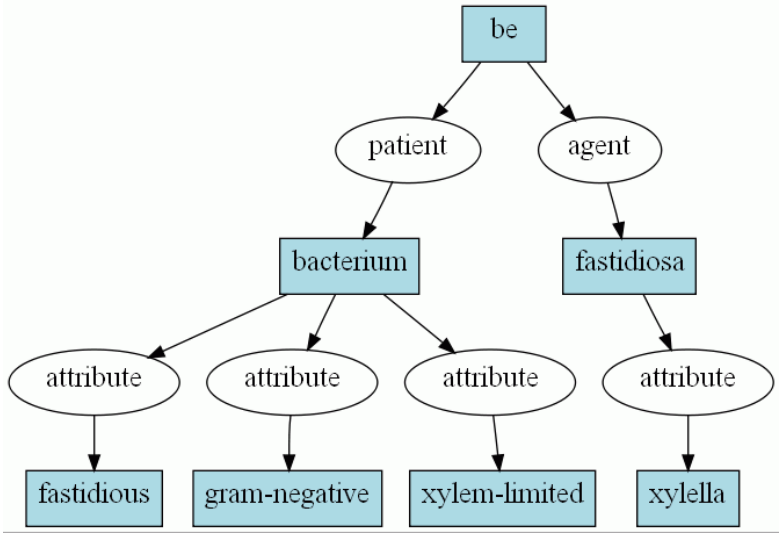

Fig. 1. Conceptual graph for the sentence "Xylella fastidiosa is a gram-negative fastidious, xylem-limited bacterium."

Conceptual graph on the Fig.1 has four conceptual relations “attribute” but only three of them indicate real objects and attributes for formal context. Using “fastidiosa” as object and “Xylella” as attribute in the formal context is wrong way because “Xylella fastidiosa” is known full name of this bacterium. Full names of bacteria have to be objects in the formal context devoted to bacteria. Word combinations denoting the names of bacteria must be recognized before conceptual graphs building. There is no other way of doing this than to use an external source of information, for example, the corpus tagging.

We also realize the following rules for creating formal contexts on conceptual graphs.

1. Not only individual concepts and relations, but also patterns of connections between concepts in conceptual graphs represented as subgraphs have been analyzed and processed. The pattern “agent - patient” is mostly frequent in biotope texts.
2. The hierarchy of conceptual relations in conceptual graphs is fixed and taken into account when creating formal context. Such hierarchy exists on the Fig.1: relations “agent” and “patient” are on the top level and relations "attribute" belong to underlying level. Using this hierarchy of conceptual relations we can select for formal contexts more or less details from conceptual graphs. This makes conceptual graphs more power and flexible semantic model for FCA than n-grams or collocations.
3. FCA – model for fact extraction is domain specific. Domain information is also taken into account in conceptual graphs building. This information is from external resources – thesauruses or tagging of textual corpuses.

   Concrete implementations of these rules are in the section 4.

3 Architecture and Functionality of the Framework

Architecture of the CGs – FCA modeling framework is shown on the Fig. 2. Consider its main elements.

Database. Database is very important part of the framework. We use relational database on the SAP-Sybase platform. It was built with CASE technology PowerDesigner™ [18] and may be scaled and expanded. Database stores texts, conceptual graphs, formal contexts and concept lattices. Special indexing is applied to textual data.

Conceptual graphs building module. This module and several other modules constitute the NLP block of modules of the framework. They realize our algorithm of acquiring conceptual graphs from texts, visualization of conceptual graphs and their clusters, interaction with external resources including WordNet.

English and Russian languages have been supported in the framework. The framework has internal dictionaries and may communicate with external ones.

Representing of modeling results. Modeling results have been presented as visualization of conceptual graphs and concept lattices as in table and textual forms. Storing all objects in database allows analyzing its data and computing conceptual graphs and concept lattice characteristics.

Programming environment. Java is the main programming platform which is used in the framework. Some modules of NLP block have been written on PowerScript language of SAP-Sybase platform.

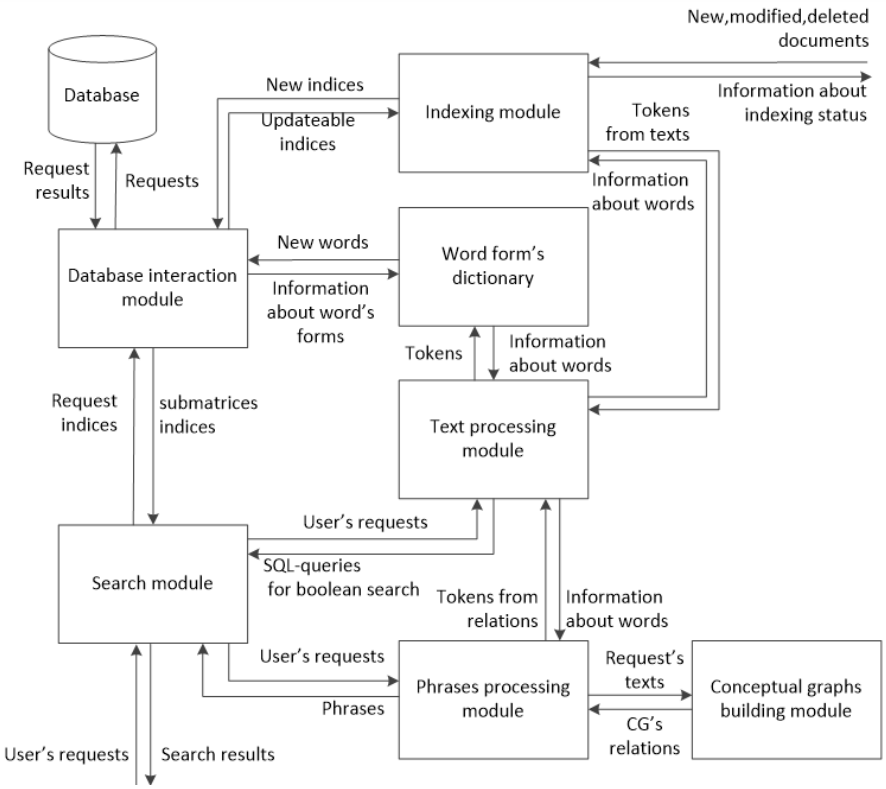

Fig. 2. Architecture of the framework

4 Experiments and Results

Experimental evaluation of CGs – FCA modeling technique has been carried out on the textual corpus of bacteria biotopes which is used in the innovation named as BioNLP Shared Task [4]. This innovation includes three IR tasks: the Bacteria Gene Renaming, the Bacteria Gene Interaction and the Bacteria Biotopes. The Bacteria Biotope task is formulated as consisting of two standard Text Mining tasks of Named Entity Recognition (NER) and Relations Extraction (RE) [20].

Biotope is an area of uniform environmental conditions providing a living place for plants, animals or any living organism. According to [4] there are two types of entities to be extracted: the names of bacteria and their locations. We added third entity of pathogenicity of bacteria.

It is preliminarily clear that the task of extracting the names of bacteria and the task of extracting locations and pathogenicity have different complexities. For extracting the names of bacteria some words or collocations (*Xylella fastidiosa*) have to be analyzed in the text. Locations and pathogenicity may be represented by more complex and long word combinations. As for bacterium *Xylella fastidiosa* on the Fig. 1, the example of its location is the following fragment from the text about it: "*the bacteria … receive a safe environment and metabolites from the insect*". To extract "*insect"* as location of bacterium we need to analyze some relations between words in the sentence. This is done also through the use of conceptual graphs.

Biotope texts tagging includes full names of bacteria, its abbreviated names and unified key codes in the database. We add additional tags if special words (*extreme, obligately*, etc.) recognized in the texts.

A BioNLP data is always domain-specific. All the texts in the corpus [4] are about bacteria themselves, their areal and pathogenicity. Not every text contains these three topics but if some of them are in the text then they are presented as separate text fragments. This simplifies text processing. According to these three topics of interest we construct three different formal contexts of "Entity", "Areal" and "Pathogenicity". They engender three different concept lattices which are connected each other. To join lattices we use facet technology [19].

Our solution of the task of Named Entity Recognition is supported by conceptual graphs. As it is illustrated above (Fig. 1) conceptual graphs can represent names of bacteria as named entities. Named Entity Recognition also includes anaphora resolution.

4.1 Anaphora resolution and noise reduction

Anaphora resolution is the problem of resolving references to earlier or later items in the text. These items are usually noun phrases representing objects called referents but can also be verb phrases, whole sentences or paragraphs. Anaphora resolution is the standard problem in NLP.

Biotope texts we work with contain several types of anaphora:

- hyperonym definite expressions ("bacterium" - "organism", "cell" - "bacterium"),
- higher level taxa often preceded by a demonstrative determinant ("this bacteria", "this organism"),
- sortal anaphors ("genus", "species", "strain").

For anaphora detection and resolution we use a pattern-based approach. It is based on fixing anaphora items in texts and establishing relations between these items and the objects in conceptual models we use. These objects are bacteria names for "Entity" context, mentions of water, soil and other environment parameters for "Areal" context and names and characteristics of diseases for "Pathogenicity" context.

Corpus tagging is also used for anaphora detection. In particular encoding bacteria (for instance bacterium *Burkholderia phytofirmans* is encoded as PsJN) is found from tagging and further used as its name in text processing.

Noise is constituted by the text elements that contain no facts or cannot be interpreted as facts. Also noise consider the data that are deliberately excluded from

consideration, for example, information about when and by whom a bacterium was first identified.

4.2 Data Processing

We have selected 130 mostly known bacteria and processed corresponding corpus texts about them. Three formal contexts of "Entity", "Areal" and "Pathogenicity" had built on the texts. They have the names of bacteria as objects and corresponding concepts from conceptual graphs as attributes.
Table 1 shows numerical characteristics of created contexts.

Context name	Number of objects	Number of attributes	Number of formal concepts
Entity	130	26	426
Areal	130	18	127
Pathogenicity	130	28	692

Table 1. Numerical characteristics of created contexts

As it is followed from the table there is relatively small number of formal concepts in the contexts. This is due to the sparse form of all contexts generated by conceptual graphs and noise reduction.

4.3 Fact extraction

Extracting facts from concept lattices is realized by forming special views constructed on the lattice and corresponded to certain property (intent in the lattice) or entity (extent in the lattice) on the set of bacteria. Every view is a sub lattice. It shows the links between concrete bacterium and its properties.

An example of such view as the fragment of lattice is shown on Fig. 3. The lattice on the Fig. 3 contains formal concepts related to the following bacteria: *Borrelia turicatae, Frankia, Legionella, Clamydophila, Thermoanaerobacter tengcongensis, Xanthomonas oryzae*. Highlighted view on the figure corresponds to gram-negative property of bacteria. Such bacteria are resistant to conventional antibiotics.

Using this view, some facts about bacteria can be extracted:

- only three bacteria from the set, *Thermoanaerobacter tengcongensis, Clamydophila* and *Xanthomonas oryzae*, are gram-negative;
- two gram-negative bacteria, *Thermoanaerobacter tengcongensis* and *Xanthomonas oryzae*, have the shape as rod;
- one of gram-negative bacteria, *Clamydophila*, is obligately pathogenic.

Note that attribute *obligately pathogenic* was formed directly from the same two words in the text according to the rule of marking words denoting extreme situation.

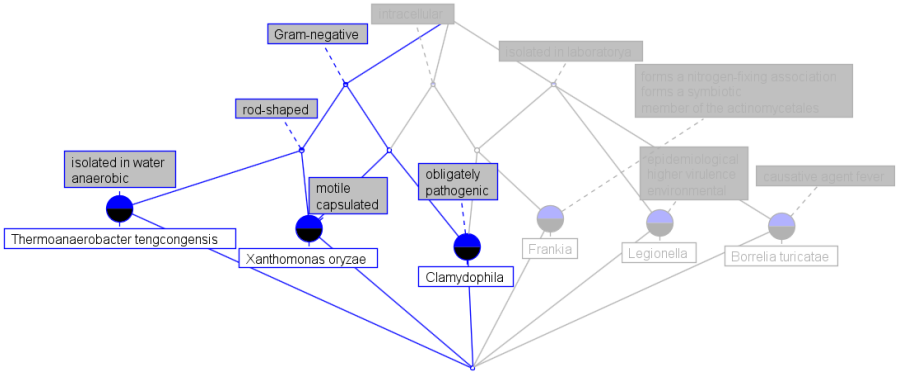

Fig. 3. Example of view concerned gram-negative property of bacteria.

We can compare our results with the known ones, most completely presented in the work [20]. Although we use the same corpus and some resembling methods (they use pattern-based approach and domain lexical resources) our results are different in fact. Our main result is not certain words extracted from texts as solution of NER and RE tasks but the whole information resource of concept lattice which is similar to ontology. So we resume that CGs – FCA modeling provides solving wider set of tasks than Named Entity Recognition and Relations Extraction, the set which corresponds to fact extraction problem.

5 Conclusion and Future Work

This paper describes the first but very important stage of creating environment for performing experiments of CGs – FCA modeling in the project of creating fact extraction technology on natural language texts. Some parts of this project are under construction but current results demonstrate effectiveness of CGs – FCA modeling.

Conceptual graphs were recognized as valid low level conceptual model for creating high level such model of concept lattice. Using conceptual graphs, it is possible to control semantic depth of representing sentences in formal concepts by selecting certain levels (sub graphs) of graph structure.

Among the topics of our future work there are the following.

Now the verb-centric approach which we use in acquiring conceptual graphs is not fully applied for creating formal contexts. When conceptual graph has the pattern <concept> - (agent) – <verb> – (patient) - <concept> the verb serves as condition which links two concepts. In other patterns with other conceptual relations including attribute verbs play the same role. This opens the need to construct tricontexts on conceptual graphs. We plan to construct multidimensional data model on our database under SAP PowerDesigner™ CASE technology and apply OLAP for modeling tricontexts and triclusters.

We also plan to use SAP HANA Environment [21] for work with big textual data.

Acknowledgments. The paper concerns the work which is partially supported by Russian Foundation of Basic Research, grant № 15-07-05507.

References

1. BioNLP 2014. Workshop on Biomedical Natural Language Processing. Proceedings of the Workshop. The Association for Computational Linguistics. Baltimore, 2014. 155 p.
2. Birkhoff, G.: Lattice Theory. Providence, RI: Amer. Math. Soc. (1967)
3. Bogatyrev, M. Y., Vakurin V. S. Conceptual Modeling in Biomedical Data Research. Mathematical Biology and Bioinformatics. 2013. Vol. 8. № 1, pp. 340–349. (in Russian).
4. Bossy R, Jourde J, Manine A-P, Veber P, Alphonse E, Van De Guchte M, Bessières P, Nédellec C: BioNLP 2011 Shared Task - The Bacteria Track. BMC Bioinformatics. 2012, 13: S8, pp. 1-15.
5. Boytcheva, S. Dobrev, P. Angelova, G.: CGExtract: Towards Extraction of Conceptual Graphs from Controlled English. Lecture Notes in Computer Science № 2120, Springer Verlag (2001)
6. Carpineto, C., Romano, G. Using Concept Lattices for Text Retrieval and Mining. In B. Ganter, G. Stumme, and R. Wille (Eds.), Formal Concept Analysis: Foundations and Applications. Lecture Notes in Computer Science 3626, pp. 161-179. Springer-Verlag, Berlin, 2005.
7. Cimiano, P. Hotho, A. Staab, S. Learning Concept Hierarchies from Text Corpora using Formal Concept Analysis. Journal of Artificial Intelligence Research, Volume 24, pp. 305-339. 2005.
8. ConExp-NG. https://github.com/fcatools/conexp-ng
9. Galitsky, B., Dobrocsi, G., de la Rosa, J.L. Kuznetsov, S.O. From Generalization of Syntactic Parse Trees to Conceptual Graphs. In: M. Croitoru, S. Ferre, D. Lukose, Eds., Conceptual Structures: From Information to Intelligence, Proc. 18th International Conference on Conceptual Structures (ICCS 2010), Lecture Notes in Artificial Intelligence (Springer), vol. 6208, pp. 185-190, 2010.
10. Ganter, B., Stumme, G., Wille, R., eds., Formal Concept Analysis: Foundations and Applications, Lecture Notes in Artificial Intelligence, No. 3626, Springer-Verlag. 2005
11. Gildea D., Jurafsky D.: Automatic labeling of semantic roles. Computational Linguistics, 2002, v. 28, 245-288. (2002)
12. Meštrović, A. Semantic Matching Using Concept Lattice. Concept Discovery in Unstructured Data, CDUD 2012, pp. 49-58.
13. Michael Bogatyrev and Alexey Kolosoff. Using Conceptual Graphs for Text Mining in Technical Support Services. Pattern Recognition and Machine Intelligence. - Lecture Notes in Computer Science, 2011, Volume 6744/2011, p.p. 466-471. Springer-Verlag, Heidelberg, 2011.
14. Michael Bogatyrev, Vadim Nuriahmetov. Application of Conceptual Structures in Requirements Modeling. – Proc. of the International Workshop on Concept Discovery in Unstructured Data (CDUD 2011) at the Thirteenth International Conference on Rough Sets, Fuzzy Sets, Data Mining and Granular Computing - RSFDGrC 2011. Moscow, Russia, 2011, pp. 11-19.
15. Montes-y-Gomez, M., Gelbukh, A., Lopez-Lopez, M.: Text Mining at Detail Level Using Conceptual Graphs. Lecture Notes In Computer Science; vol. 2393, pp. 122 – 136 (2002)

16. Otero P. G., Lopes G. P., Agustini, A., Automatic Acquisition of Formal Concepts from Text. Journal for Language Technology and Computational Linguistics. Vol. 23(1), pp. 59-74. 2008.
17. Poelmans J., Kuznetsov S. O., Ignatov D. I., Dedene G. Formal Concept Analysis in knowledge processing: A survey on models and techniques // Expert Systems with Applications. 2013. Vol. 40. No. 16. P. 6601-6623.
18. PowerDesigner 16.2. Sybase documentation, DC 00121-01-1520-01. February 2015.
19. Priss, U., Linguistic Applications of Formal Concept Analysis. In: Ganter; Stumme; Wille (eds.), Formal Concept Analysis, Foundations and Applications. Springer Verlag. LNAI 3626, p. 149-160. 2005.
20. Ratkovic, Z., Golik, W., Warnier, P. Event extraction of bacteria biotopes: a knowledge-intensive NLP-based approach. - BMC Bioinformatics 2012, 13, (Suppl 11): S8, pp. 1-11.
21. SAP HANA Environment. https://hana.sap.com/abouthana.html
22. Sowa, J.F., Knowledge Representation: Logical, Philosophical, and Computational Foundations, Brooks Cole Publishing Co., Pacific Grove, CA, (2000).
23. Wille, R.: Conceptual Graphs and Formal Concept Analysis. Proceedings of the Fifth International Conference on Conceptual Structures: Fulfilling Peirce's Dream. 290 - 303. Springer-Verlag, London. (1997)