

# Fact Extraction from Natural Language Texts with Conceptual Modeling

Mikhail Bogatyrev<sup>(✉)</sup>

Tula State University, Tula, Russia  
okkambo@mail.ru

**Abstract.** The paper presents the application of Formal Concept Analysis paradigm to the fact extraction problem on natural language texts. Proposed technique combines the usage of two conceptual models: conceptual graphs and concept lattice. Conceptual graphs serve as semantic models of text sentences and the data source for concept lattice – the basic conceptual model in the Formal Concept Analysis. With the use of concept lattice it is possible to model relationships between words from different sentences from different texts. These relationships have been collected in formal concepts of concept lattice and provide interpreting formal concepts as possible facts. Facts can be extracted by using navigation in the lattice and interpretation its concepts and hierarchical links between them. Experimental investigation of the proposed technique is performed on the annotated textual corpus consisted of descriptions of biotopes of bacteria.

**Keywords:** Conceptual graphs · Formal Concept Analysis · Concept lattice

## 1 Introduction

The problem of fact extraction from text is the part of more general problem of knowledge extraction from text [1]. Methods for solving this problem are strongly depended on whether the text is structured or not. We will use the term “text” for natural language text and the term “textual data” when text is structured by means of database or corpus. Facts and events form a kind of knowledge which represents semantics of a certain portion of text. In this area of research the term “event” is applied in the literature more often than “fact” [2] and sometimes these terms have similar meaning. However we distinguish facts and events in the corresponding problems of knowledge extraction.

Both facts and events extracted from texts can be represented by words and relationships on the sets of words. An example of fact is phrase “SAP has purchased SYBASE” and this phrase also denotes an event of purchasing. The model of this event may be in the form of pattern  $\langle agent \rangle$ -purchase- $\langle patient \rangle$  where concrete words may be substituted as semantic roles of agent and patient. In the survey [2] facts are defined as “statistical relations”, so the evidence of facts is detected statistically and discovered relations “are not necessarily semantically valid, as semantics (meanings) are not explicitly considered, but are assumed to be implicit in the data” [2]. Now this definition may be replenished so that relations in the fact model can be found semantically valid and the evidence of facts is detected not statistically but also semantically. Certain

technologies, including one presented in this paper, devoted to extract facts using semantics explicitly presented in corresponding semantic models of text. Many of these models are the same as in the fact extraction problems as in the event extraction problems: for example lexico-syntactic and lexico-semantic patterns are applied there. These models are also applied for solving Named Entity Recognition (NER) problem. Solutions of this problem often come as the base for solutions of fact extraction problem [2].

We consider fact as realized or occurred event. So the modeling of events and facts may be implemented in a same way. We apply conceptual modeling [5] in the fact extraction problem. This method is based on the usage of two conceptual models, conceptual graphs and concept lattice, to discover facts as formal concepts and their relationships in concept lattice.

Conceptual modeling is one of the ways of modeling semantics in the Natural Language Processing (NLP) [6, 7]. Every conceptual model has its own semantics which represents the meanings of concepts and relationships on them.

Formal Concept Analysis (FCA) [17] is the paradigm of conceptual modeling which studies how objects can be hierarchically grouped together according to their common attributes. In the FCA, its conceptual model is the lattice of formal concepts (concept lattice) which is built on the abstract sets treated as objects and their attributes. Concept lattices have been applied as an instrument for information retrieval and knowledge extraction in many applications. The number of FCA applications now is growing up including applications in social science, civil engineering, planning, biology, psychology and linguistics [22, 23]. Several successful implementations of FCA methods on fact extraction on textual data [12, 13] and Web data are known [19]. Although the high 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 on textual data.

Another paradigm of conceptual modeling is Conceptual Graphs [25]. Conceptual graph is bipartite directed graph having two types of vertices: concepts and conceptual relations. Conceptual terms of entities and relationships are represented in conceptual graphs as its concepts and conceptual relations.

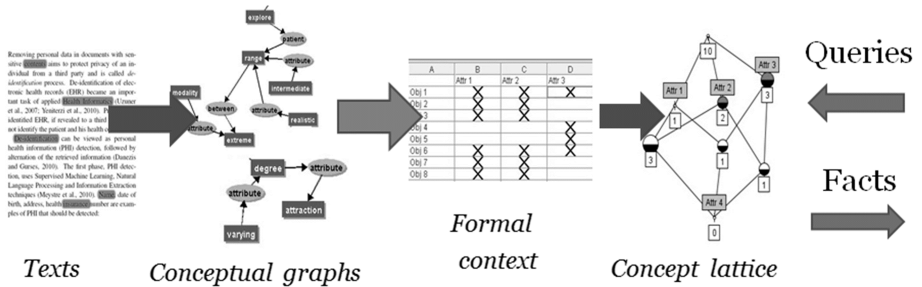
Conceptual graphs have been applied for modeling many real life objects including texts. Acquiring conceptual graphs from natural language texts is non-trivial problem but it is quite solvable [5].

There is great number of various methods of solving fact and event extraction problems which can be distinguished according to data-driven and knowledge-driven approaches [2]. Data-driven approach is based on the idea that knowledge (facts or events) presented explicitly in data whereas knowledge-driven approach requires external resources or expert knowledge for solving the problem.

Fact extraction technology proposed in this paper is hybrid. Using conceptual graphs as semantic model of text we follow the data-driven approach. Expert knowledge-driven methods are applied in the output of the technology when facts have to be detected and presented in the output interface. The principles of creating this technology are described in [6] and its implementation in biomedical data research is described in [9]. In this paper we present some generalizations of these principles and new experimental results of investigation of biotopes of bacteria.

## 2 Fact Extraction Technology

The work of fact extraction technology is illustrated on the Fig. 1.



**Fig. 1.** Elements of the fact extraction technology.

The elements of this technology have the following content.

1. Input data in the form of plain text is transformed to the set of conceptual graphs. The maximal number of conceptual graphs is equal to the number of processed sentences of texts.
2. According to FCA paradigm, so called formal context is building on the set of conceptual graphs. It is a matrix denoting a relation on two sets of objects and their attributes. These sets must be determined on the set of conceptual graphs. This stage is a crucial step in the technology. The number of formal contexts and their content depends on many factors and is domain-specific.
3. Formal context contains formal concepts which are combinations of objects and attributes that meet certain conditions known as Galois connection and constitute a lattice named as concept lattice [17]. Concept lattice is interpreted as storage of facts. Facts can be extracted by processing input textual queries and then navigating in the lattice and interpreting its concepts and hierarchical links between them.

### 2.1 Acquiring and Implementing Conceptual Graphs

The method of acquiring conceptual graphs from natural language texts is considered in [5]. Some peculiarities of conceptual graphs created with this method are illustrated in [6, 7].

The method has standard phases of lexical, morphological and semantic analysis extended with the solution of the problem of semantic role labeling [8]. This problem is non-trivial since semantic roles do not belong to the sentence processed and must be discovered from existing roles by means of morphological analysis.

Semantic analysis on the stage of acquiring conceptual graphs is domain-specific. For example, working with biological domain and not considering its specificity we will not acquire correct conceptual graph for the following sentence:

*"H12424 is characterized as a representative of the B. cenocepacia PHDC clonal lineage".*

Wrong conceptual graph is a graph which has isolated concepts do not linked with any other concepts as it is shown on Fig. 2.

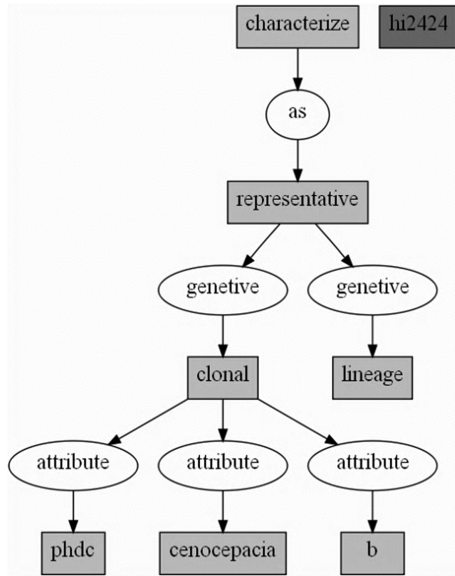


Fig. 2. Example of conceptual graph with isolated concept.

For the sentence above and for similar sentences being characteristic for biological domain we use supervised learning and external resources in the form of textual corpus. Then, after learning, the algorithm of acquiring conceptual graphs knows that *B. cenocepacia* is a shortcut name of the *Burkholderia cenocepacia* bacterium, *HI2424* is the code of this bacterium and *PHDC* is the name of the clone of bacteria.

Extracting facts is performed on the same stage of creating conceptual graphs. Some isolated concepts appearing on applying the algorithm before its learning may indicate facts. Figure 2 illustrates this showing conceptual graph for the sentence discussing above before the algorithm learning. Here the presence *HI2424* code in the sentence is the fact that marks this sentence as having information about *Burkholderia cenocepacia* bacterium which will be used later to filter non-informative sentences.

The next stage of the fact extraction technology is creating formal contexts and concept lattice as the main conceptual model serving as the source of facts. Conceptual graphs and FCA models are closely related when they are applied as conceptual models in text processing. One of the first mentioning of this relation is in [30]. Now it is used in connection with the problem of aggregation of conceptual graphs.

## 2.2 Conceptual Graphs and Formal Concept Analysis

There are two basic notions FCA deals with: *formal context* and *concept lattice* [17]. 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  $\sqsubseteq$  and  $\Subset$ , correspondingly:  $G = (G, \sqsubseteq)$ ,  $M = (M, \Subset)$ . Formal context is represented by  $[0, 1]$  matrix  $\mathbf{K} = \{k_{ij}\}$  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 exist mappings (which may be functions also)  $A': A \rightarrow B$  and  $B': B \rightarrow A$  with properties of  $A' = \{\exists m \in M \mid \langle g, m \rangle \in I \forall g \in A\}$  and  $B' = \{\exists g \in G \mid \langle g, m \rangle \in I \forall m \in B\}$  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  $\mathbf{K} = (G, M, I) \in$ , respectively.

By other words, 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  $\sqsubseteq$  and  $\Subset$  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  $(A_2, B_2)$ ,  $A_1 \sqsubseteq A_2$  and  $B_2 \Subset B_1$  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 [17] 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.

To illustrate these abstract definitions consider an example. Figure 3 shows simple formal context and concept lattice composed on the sets  $G = \{DNA, Virus, Prokaryotes, Eukaryotes, Bacterium\}$  and  $M = \{Membrane, Nucleus, Replication, Recombination\}$ . The set  $G$  is ordered according to sizes of its elements: DNA is smallest and bacterium is biggest ones. The set  $M$  has relative order: one part (*Membrane, Nucleus*) characterizes microbiological structure of objects from  $G$ , but another part (*Replication, Recombination*) characterizes the way of breeding, and these parts are incomparable.

	Membrane	Nucleus	Replication	Recombination
DNA				X
Virus	X			X
Prokaryotes	X		X	
Eukaryotes	X	X	X	
Bacterium	X		X	

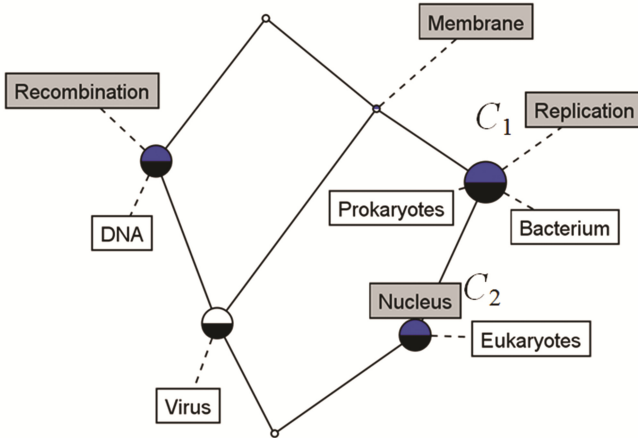


Fig. 3. Example of formal context and concept lattice.

The lattice for formal context on Fig. 3 is drawn compact and is interpreted in the following way. There are empty concepts on the top and on the bottom of the lattice diagram. Every formal concept lying on the path from top to bottom contains attributes (shown dark) which are gathered from the concepts lying before. Vice versa, every formal concept lying on the path from bottom to top contains objects (shown bright) which are gathered from the concepts lying before. That is why the concept  $C_1 = (\{Prokaryotes, Eukaryotes, Bacterium\}, \{Membrane, Replication\})$  contains the object *Eukaryotes* and the attribute *Membrane*. The concept  $C_1$  is more general than the concept  $C_2 = (\{Eukaryotes\}, \{Membrane, Replication, Nucleus\})$ .

Also on the Fig. 3 there is the fact of existing two different branches of concepts characterizing two families:  $\{viruses, DNA\}$  and  $\{prokaryotes, eukaryotes, bacteria\}$ . The link between them is the attribute “*Membrane*”. It is known [11] that viruses can have a lipid shell formed from the membrane of the host cell. Therefore, the membrane is positioned in the formal context on the Fig. 3 as an attribute of the virus.

This example demonstrates specific ways of extracting knowledge from conceptual lattice:

- analyzing formal concepts in concept lattice;
- analyzing conceptual structures in concept lattice – its paths and sub lattices in the general case.

These ways are applied in our previous [9] and current research of bacteria biotopes.

**FCA on Textual Data.** The main problem in applying FCA to textual data is the problem of building formal context. If textual data is represented as the natural language texts then this problem becomes acute. 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 is the context in which the objects are text documents and the attributes are the terms from these documents. Another variant is building formal context directly on the texts and the formal context may represent various features of textual data:

- semantic relations (synonymy, hyponymy, hypernymy) in a set of words for semantic matching [20],
- verb-object dependencies from texts [14],
- words and their lexico-syntactic contexts [21, 24].

These lexical elements must be distinguished in texts as objects and attributes. There are following approaches to solve this problem:

- creating corpus tagging by adding special descriptions in texts which mark objects and attributes [10],
- using semantic models of texts [14].

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

**Aggregation of Conceptual Graphs and Pattern Structures.** In the theory of conceptual graphs aggregation means replacing conceptual graphs by more general graphs [25]. These general graphs may be created as new graphs or may be graphs or sub graphs from initial set of graphs. Aggregation of conceptual graphs has semantic meaning and general graphs make up the context (not formal context) of initial set of graphs.

One way of aggregation is conceptual graphs clustering. Graphs which are the nearest ones to the centers of clusters have been treated as general graphs.

We have studied several approaches for clustering conceptual graphs using various similarity measures [6] and applied clustering for creating formal concepts on conceptual graphs.

Another way of conceptual graphs aggregation is based on supporting types of concepts of conceptual graphs. Types of concepts have been implied in the model of conceptual graph [17]. To support types of concepts, external resources are needed. They may be thesaurus or textual corpus with tagging or ontology.

According to generalization of FCA [16] conceptual graphs and their external resource may be considered as pattern structures.

### 2.3 Creating Formal Contexts with Conceptual Graphs

The crucial step in the described process of CGs – FCA modeling is creating formal contexts on the set of conceptual graphs.

At the first glance, this problem seems simple: those concepts of conceptual graphs 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 provide the presence information about those and other facts in the formal contexts the following rules are implemented as mostly important when creating formal contexts.

1. Not only individual concepts and relations, but also patterns of connections between concepts in conceptual graphs represented as sub graphs have been analyzed and processed. These patterns are the predicate forms  $\langle object \rangle - \langle predicate \rangle - \langle subject \rangle$  which in conceptual graphs look as the template  $\langle concept \rangle - (patient) - \langle verb \rangle - (agent) - \langle concept \rangle$ . Not only *agent* and *patient* semantic roles but also other roles are allowed in the templates.
2. The hierarchy of conceptual relations in conceptual graphs is fixed and taken into account when creating formal context. Using this hierarchy of conceptual relations it is possible to select for formal contexts more or less details from conceptual graphs.

These empirical rules are related to the principle of pattern structures which was introduced in FCA in the work [16]. A pattern structure is the set of objects with their descriptions (patterns), not attributes. Patterns also have similarity operation. The instrument of pattern structures is for creating concept lattices on the data being more complicated than sets of objects and attributes.

Conceptual graph is a pattern for the object it represents. A sub graph of conceptual graph is projection of a pattern. Namely projections are often used for creating formal contexts. Similarity operation on conceptual graphs is a measure of similarity which is applied in clustering.

### 3 Fact Extraction from Biomedical Data

Bioinformatics is one of the fields where Data Mining and Text Mining applications are growing up rapidly. New term of “Biomedical Natural Language Processing” (BioNLP) has been introduced there [4]. This term is stipulated by huge amount of scientific publications in Bioinformatics and organizing them into corpora with access to full texts of articles via such systems as PubMed [26]. Information resources of PubMed have been united in several subsystems presenting databases, corpora and ontologies.

So called “research community around PubMed” [18] forms data intensive domain in this area. It not only uses data from PubMed but also creates new data resources and data mining tools including specialized languages for effective biomedical data processing [15].

In our experiments we use PubMed vocabulary thesaurus MeSH (Medical Subject Headings) as external resource for supporting types of concepts in conceptual graphs.



### 3.1 Data Structures

Our experiments have been carried out using text corpus of bacteria biotopes which is used in the innovation named as BioNLP Shared Task [10]. Biotope is an area of uniform environmental conditions providing a living place for plants, animals or any living organism. Biotope texts form tagged corpus. The tagging includes full names of bacteria, its abbreviated names and unified key codes in the database. We can add additional tags and we do it.

A BioNLP data is always domain-specific. All the texts in the corpus 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.

The fact extraction technology is realized as experimental modeling framework [7] having DBMS for storing and managing data used in experiments. We use relational database on the SAP-Sybase platform. Database stores texts, conceptual graphs, formal contexts and concept lattices. Special indexing is applied on textual data.

### 3.2 BioNLP Tasks

According to the BioNLP Shared Task initiative [10] there are two main tasks solving on biomedical corpora: the task of Named Entity Recognition (NER) and the task of Relations Extraction (RE).

The task of Named Entity Recognition on the corpus of bacteria descriptions is formulated as seeking bacteria names presented directly in the texts or as co-references (anaphora).

Relations Extraction means seeking links between bacteria and their habitat and probably diseases it causes. The task of Named Entity Recognition has direct solution with conceptual graphs. The only problem which is here is anaphora resolution.

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.

Biotopes texts we work with contain several types of anaphora:

- hypernym defining expressions (“bacterium” - “organism”, “cell” - “bacterium”),
- higher level taxa often preceded by a demonstrative determinent (“this bacteria”, “this organism”),
- sortal anaphoras (“genus”, “species”, “strain”).

For anaphora detection and resolution we used a pattern-based approach. It is based on fixing anaphora items in texts and establishing relations between these items and bacteria names. Additional details may be found in [6, 9].

### 3.3 Fact Extraction with Concept Lattices

Conceptual graphs represent relations between words. Therefore they can be applied for relations extraction but only in one sentence. For extracting relations between bacteria in several texts we applied concept lattices.

We had selected 130 mostly known bacteria and have processed corresponding corpus texts about them. All the texts were preliminary filtered for excluding stop words and other non-informative lexical elements.

Three formal contexts of “Entity”, “Areal” and “Pathogenicity” were 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.

**Table 1.** 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

Among attributes there are bacteria properties (gram-negative, rod-shaped, etc.) for “Entity” context, mentions of water, soil and other environment parameters for “Areal” context and names and characteristics of diseases for “Pathogenicity” context.

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.

**Visualization in Fact Extraction.** Visualization plays significant role in FCA [28] and in fact extraction since not only formal concepts but also relations between concepts in a concept lattice may be treated as facts, and visualization helps to find them fast. But it allows getting results only for the relatively small lattices. For extracting facts we use visualization together with other ways including database technologies. A possibility was created to visualize sub lattices of a concept lattice to form special views constructed on the lattice corresponding to certain property (intent in the lattice) or entity (extent in the lattice) on the set of bacteria. We applied open source tool [29] which was modified and built in our system.

Consider the example demonstrating the work of the system. One of the problems solving in investigations of bacteria biotopes is the problem of bacteria classification: it is needed to classify bacteria according to their properties characterizing them as the entities, characterizing their areal and pathogenicity. Various bacteria may have similar properties or may not. It is interesting to find clusters of bacteria containing ones having similar properties. This clustering task may be solved with a concept lattice.

Figure 4 shows a fragment of the formal context with the attributes related to some properties of bacteria: Gram staining, the property of being aerobic, etc.

	gram-negative	Gram-positive	acid-fast	rod-shaped	spiral	aerobic	anaerobic	non-spore	bacteria
<i>Mycobacterium bovis</i>		×		×		×			×
<i>Mycoplasma agalactiae</i>		×							
<i>Bifidobacterium longum</i>		×					×		
Corynebacteria		×							×
Mycobacteria		×	×	×					×
Streptomycetes		×							×
<i>Mycobacterium tuberculosis</i>		×				×			×
Clostridium		×					×		×
<i>Clostridium sticklandii</i>		×					×		
<i>Deinococcus radiodurans</i>		×							
<i>Deinococcus deserti</i>	×								
<i>Cupriavidus metallidurans</i>	×							×	
<i>Thermoanaerobacter tengcongensis</i>	×			×			×		
<i>Thermus thermophilus</i>	×					×			
<i>Sinorhizobium meliloti</i>	×					×			
<i>Bdellovibrio bacteriovorus</i>	×			×					
<i>Xylella fastidiosa</i>	×								
<i>Chlamydia pneumoniae</i>	×								
<i>Helicobacter hepaticus</i>	×				×				
<i>Legionella pneumophila</i>	×			×		×			

Fig. 4. A fragment of the formal context for 20 bacteria.

It is evident directly from the context that these 20 bacteria constitute two clusters according to the Gram staining: there is no bacterium which is simultaneously Gram-positive and Gram-negative. Lattice diagrams on the Fig. 5 confirm this fact.

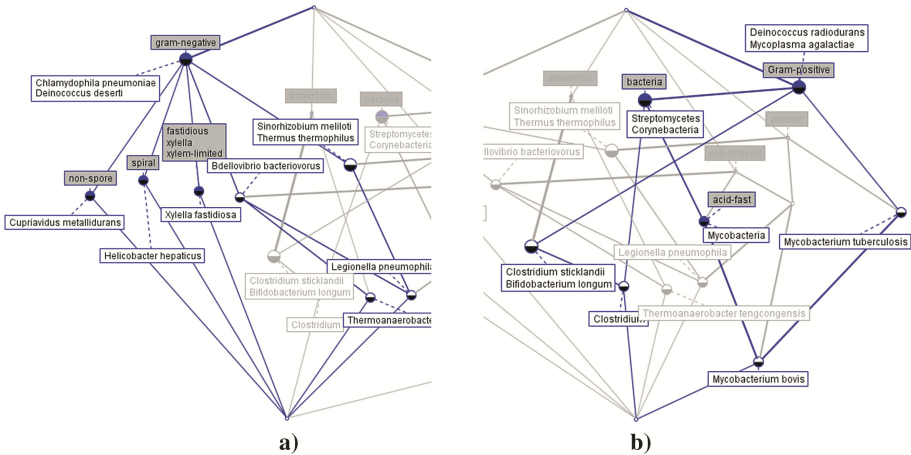
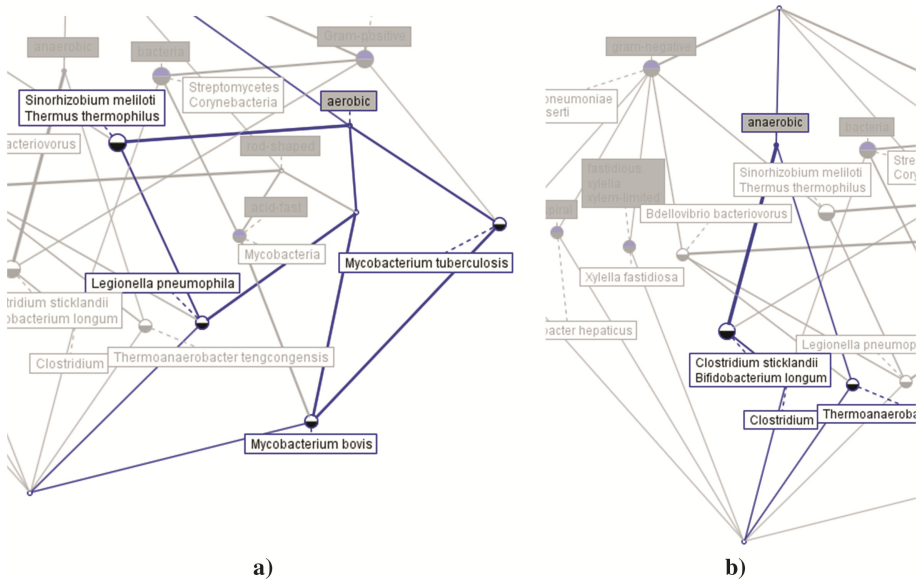


Fig. 5. Views of concept lattice demonstrating Gram staining: (a) – Gram-negative property, (b) – Gram-positive property.

Interpreting views on Fig. 5 as we did it for the example on Fig. 3 we resolve that bacteria are clustered according to their Gram staining because the views on Fig. 5(a) and (b) do not intersect.

Clustering bacteria according to the property of being aerobic is not evident from the context on Fig. 4. Lattice diagrams on Fig. 6 confirm the clustering bacteria according to this property in the same manner as for Fig. 4.



**Fig. 6.** Views of concept lattice demonstrating the property of bacteria to be aerobic.

However, the number of bacteria in Figs. 5 and 6 is not the same: Fig. 5 contains all 20 bacteria (10 in Fig. 5-a and 10 in Fig. 5-b.) and Fig. 4 - contains only 9 bacteria. This is due to the fact that the relevant texts do not contain information about the property of being aerobic for some bacteria.

**Comparing Results.** We can compare our results with two known similar solutions related to fact extraction problem. The first solution of extracting events is presented in [3] and is based on using special framework of EventMine. This solution is realized as marking of the text by highlighting its lexical elements as elements of event.

The second solution [24] is directly connected with BioNLP. The tasks of Named Entity Recognition and Relation extraction were solved in [24] with Alvis framework [27]. In [24] results of relations extraction are also presented as marked words in the texts. Our results of solution of NER are similar to [24] and presented in [6].

Comparing our current results of fact extraction with the known ones we resume that concept lattice provides principally another variant of solution of fact extraction problem. The main distinction of this solution is that it is not realized in the processed text by highlighting its lexical elements but it is realized with new external resource, conceptual model in the form of the concept lattice.

## 4 Conclusions and Future Work

This paper describes the idea of joining two paradigms of conceptual modeling - conceptual graphs and concept lattices. Current results of realizing this idea on textual data show its good potential for fact and knowledge extraction. Concept lattice may

serve as a skeleton of ontology constructed on texts. Its data which may or may not be interpreted as facts constitutes a knowledge stored in the concept lattice being ready to extract.

In spite of the certain useful features of presented technology there are some problems which need to be solved for improving the quality of modeling technique.

1. Conceptual graphs acquired from texts contain many noisy elements. Noise is constituted by the text elements that contain no useful information or cannot be interpreted as facts. Noisy elements significantly decrease efficiency of algorithms of fact extraction.
2. Empirical rules which we use for creating formal contexts cannot embrace all configurations of conceptual graphs. More formal approach to creating formal contexts on the set of conceptual graphs will guarantee the completeness of solution. We guess that using pattern structures and their projections is the way of formalizing our modeling technique.
3. The next stage of developing current technology is creating of fledged information system which processes user queries and produces solutions of certain tasks on textual data. Not only visualization but also special user oriented interfaces to concept lattice will be created in this system.

**Acknowledgments.** The paper concerns with the work partially supported by the Russian Foundation for Basic Research, grant № 15-07-05507.

## References

1. Aggarwal, C.C., Zhai, C.: *Mining Text Data*. Springer, New York (2012)
2. Hogenboom, F., Frasinca, F., Kaymak, U., de Jong, F., Caron, E.: A survey of event extraction methods from text for decision support systems. *Decis. Support Syst. (DSS)* **85**, 12–22 (2016)
3. Miwa, M., Ananiadou, S.: Adaptable, high recall, event extraction system with minimal configuration. *BMC Bioinformatics* **16**(10), 1–11 (2015)
4. BioNLP 2014. Workshop on Biomedical Natural Language Processing. Proceedings of the Workshop. The Association for Computational Linguistics, Baltimore (2014)
5. Bogatyrev, M., Tuhtin, V.: Creating conceptual graphs as elements of semantic texts labeling. In: *Computational Linguistics and Intellectual Technologies, Proceedings of the International Conference “Dialogue”, Moscow*, pp. 31–37 (2009). (in Russian)
6. Bogatyrev, M.: Conceptual modeling with formal concept analysis on natural language texts. In: *Proceedings of the XVIII International Conference “Data Analytics and Management in Data Intensive Domains”, CEUR Workshop Proceeding*, vol. 1752, pp. 16–23 (2016)
7. Bogatyrev, M., Samodurov, K.: Framework for conceptual modeling on natural language texts. In: *Proceedings of International Workshop on Concept Discovery in Unstructured Data (CDUD 2016) at the Thirteenth International Conference on Concept Lattices and Their Applications*. Moscow, CEUR Workshop Proceeding, vol. 1625, pp. 13–24 (2016)
8. Gildea, D., Jurafsky, D.: Automatic labeling of semantic roles. *Comput. Linguist.* **28**, 245–288 (2002)
9. Bogatyrev, M.Y., Vakurin, V.S.: Conceptual modeling in biomedical data research. *Math. Biol. Bioinform.* **8**(1), 340–349 (2013)

10. 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*, 13:S8, 1–15 (2012)
11. Campbell, N.A., et al.: *Biology: Concepts and Connections*. Benjamin-Cummings Publishing Company (2005)
12. Carpineto, C., Romano, G.: Exploiting the potential of concept lattices for information retrieval with CREDO. *J. Univ. Comput.* **10**(8), 985–1013 (2004)
13. Carpineto, C., Romano, G.: Using Concept Lattices for Text Retrieval and Mining. In: Ganter, B., Stumme, G., Wille, R. (eds.) *Formal Concept Analysis. LNCS (LNAD)*, vol. 3626, pp. 161–179. Springer, Heidelberg (2005). doi:[10.1007/11528784\\_9](https://doi.org/10.1007/11528784_9)
14. Cimiano, P., Hotho, A., Staab, S.: Learning concept hierarchies from text corpora using formal concept analysis. *J. Artif. Intell. Res.* **24**, 305–339 (2005)
15. Edhlund, B., McDougall, A.: *Pubmed Essentials, Mastering the World’s Health Research Database*. Form & Kunskap AB (2014)
16. Ganter, B., Kuznetsov, S.O.: Pattern structures and their projections. In: Delugach, H.S., Stumme, G. (eds.) *ICCS-ConceptStruct 2001. LNCS (LNAI)*, vol. 2120, pp. 129–142. Springer, Heidelberg (2001). doi:[10.1007/3-540-44583-8\\_10](https://doi.org/10.1007/3-540-44583-8_10)
17. Ganter, B., Stumme, G., Wille, R. (eds.): *Formal Concept Analysis. LNCS (LNAI)*, vol. 3626. Springer, Heidelberg (2005). doi:[10.1007/978-3-540-31881-1](https://doi.org/10.1007/978-3-540-31881-1)
18. Hunter, L., Cohen, K.B.: Biomedical language processing: what’s beyond PubMed? *Mol. Cell* **21**, 589–594 (2006)
19. Ignatov, D.I., Kuznetsov, S.O., Poelmans, J.: Concept-based biclustering for internet advertisement. In: Vreeken, J., Ling, C., Zaki, M.J., Siebes, A., Yu, J.X., Goethals, B., Webb, G., Wu, X. (eds.) *Proceeding of 12th IEEE International Conference on Data Mining Workshops (ICDMW 2012)*, pp. 123–130 (2012)
20. Meštrović, A.: Semantic matching using concept lattice. In: *Proceeding Concept Discovery in Unstructured Data (CDUD-2012)*, pp. 49–58 (2012)
21. Otero, P.G., Lopes, G.P., Agustini, A.: Automatic acquisition of formal concepts from text. *J. Lang. Technol. Comput. Linguist.* **23**(1), 59–74 (2008)
22. Poelmans, J., Kuznetsov, S.O., Ignatov, D.I., Dedene, G.: Formal concept analysis in knowledge processing: a survey on models and techniques. *Expert Syst. Appl.* **40**(16), 6601–6623 (2013)
23. Priss, U.: Linguistic applications of formal concept analysis. In: Ganter, B., Stumme, G., Wille, R. (eds.) *Formal Concept Analysis. LNCS (LNAI)*, vol. 3626, pp. 149–160. Springer, Heidelberg (2005). doi:[10.1007/11528784\\_8](https://doi.org/10.1007/11528784_8)
24. Ratkovic, Z., Golik, W., Warnier, P.: Event extraction of bacteria biotopes: a knowledge-intensive NLP-based approach. - *BMC Bioinformatics*, 13(Suppl 11):S8, pp. 1–11 (2012)
25. Sowa, J.F.: *Conceptual Structures: Information Processing in Mind and Machine*. Addison-Wesley, London (1984)
26. U.S. National Library of Medicine. <http://www.ncbi.nlm.nih.gov/pubmed>
27. Alvis framework. <http://bibliome.jouy.inra.fr>
28. Yevtushenko, S.A.: System of data analysis “Concept Explorer”. In: *Proceeding of the 7th National Conference on Artificial Intelligence KII-2000, Russia*, pp. 127–134 (2000). (in Russian)
29. ConExp-NG. <https://github.com/fcatools/conexp-ng>
30. Wille, R.: Conceptual graphs and formal concept analysis. In: Lukose, D., Delugach, H., Keeler, M., Searle, L., Sowa, J. (eds.) *Conceptual Structures: Fulfilling Peirce’s Dream. ICCS-ConceptStruct 1997. LNCS (LNAI)*, vol. 1257, pp. 290–303. Springer, Heidelberg (1997). doi:[10.1007/BFb0027878](https://doi.org/10.1007/BFb0027878)