Conceptual Modeling with Formal Concept Analysis on Natural Language Texts †

© Mikhail Bogatyrev
Tula State University,
Tula
okkambo@mail.ru

Abstract

The paper presents conceptual modelling technique on natural language texts. This technique combines the usage of two conceptual modeling paradigms: conceptual graphs and Formal Concept Analysis. Conceptual graphs serve as semantic models of text sentences and the data source for concept lattice – the basic conceptual model in Formal Concept Analysis. With the use of conceptual graphs the Text Mining problems of Named Entity Recognition and Relations Extraction are solved. Then these solutions are applied for creating concept lattice. The main problem investigated in the paper is the problem of creating formal contexts on a set of conceptual graphs. Its solution is based on the analysis of semantic roles and conceptual patterns in conceptual graphs. Concept lattice built on textual data is applied for knowledge extraction. Knowledge, sometimes interpreted as 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.

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

1 Introduction

Knowledge extraction from textual data requires more in-depth intensive analysis of this data. In the area of Text Mining, some variants of knowledge extraction have been realized by solving such problems as *sentiment analysis*, *fact extraction* and *decision making support*. To solve these problems it is necessary to have models that reflect semantics of textual data. It is especially urgent when this data is presented as unstructured natural language texts.

Proceedings of the XVIII International Conference «Data Analytics and Management in Data Intensive Domains» (DAMDID/RCDL'2016), Ershovo, Russia, October 11 - 14, 2016

Conceptual modeling is one of the ways of modeling semantics in the Natural Language Processing (NLP) [22]. Conceptual modeling is the process of conceptualization of real world phenomena and creating conceptual models as a result of conceptualization. Conceptual model is a graph which vertices are concepts and arrows or edges are links between concepts. Every conceptual model has its own semantics which represents the meanings of concepts and links.

Conceptual modeling has long been applied for databases and software modeling [19] and this term is also used in other fields including NLP. Entity Relationship Diagram (ERD) [19] is well known representative of conceptual models. It describes the structure of database in terms of *entities*, *relationships*, and *constraints*. These terms of entities, relationships, and constraints are explicitly or implicitly present at many other conceptual models including ones discussed in this paper.

Formal Concept Analysis (FCA) [13] 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 [21], [22]. Several successful implementations of FCA methods on fact extraction on textual data [8] and Web data are known [15]. 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.

The main problem in creating concept lattice on textual data is building so called *formal contexts* on this data. Formal context is matrix representation of the relation on the sets of objects and attributes. So it is needed to acquire words or word combinations from texts which are interpreted as objects and attributes. To restrict all possible combinations of words of such meanings we need to select from them those ones which are valued for solving concrete problem or the class of problems. As a result a concept lattice created on texts

becomes domain specific. This is similar to the design of ontologies and concept lattice is often considered as framework of ontology [21].

Another paradigm of conceptual modeling is Conceptual Graphs (CGs) [24]. 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 [3], [5].

The main purpose of this paper is to show how two paradigms of conceptual modeling - Conceptual Graphs and Formal Concept Analysis - can be united in one modeling technique. The idea of joining these two paradigms seems very attractive but not elaborated much enough [22], [26].

Proposed technique is used in on-going project of creating fact extraction system working on biomedical data. Experimental investigation of it is performed on the annotated textual corpus consisted of descriptions of biotopes of bacteria.

2 CGs-FCA modeling

The proposed modeling technique named briefly as CGs-FCA modeling is based on using conceptual graphs and concept lattice. It may be applied for knowledge extraction from textual data. In CGs-FCA modeling conceptual graphs serve as semantic models of text sentences and the data source for formal context of concept lattice. Concept lattice built on textual data is applied for knowledge extraction. Knowledge, sometimes interpreted as facts, can be extracted by using navigation in the lattice and interpretation its concepts and hierarchical links between them.

To illustrate CGs-FCA modeling, consider some FCA basics.

2.1 Formal Concept Analysis basics

There are two basic notions FCA deals with: *formal context* and *concept lattice* [13]. 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 P, correspondingly: $G = (G, \phi)$, $M = (M, \mathrm{P})$. 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 $B' : B \to A$ 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.

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 ϕ and P 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 $B_2 \, \mathrm{P} \, B_1$ 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 [13], 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.

A	B	C	D	E
	Membrane	Nucleus	Replication	Recombination
DNA				X
Virus	X			X
Prokaryotes	X		X	
Eukaryotes	X	X	X	
Bacterium	X		X	

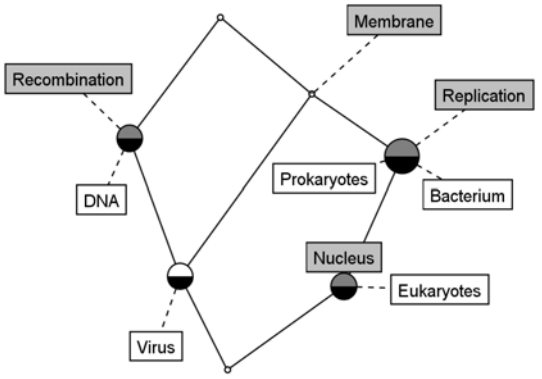

Figure 1 Example of formal context and concept lattice.

To illustrate these abstract definitions consider an example. Figure 1 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. In the concept lattice the bacterium is placed in the concept C_1 = ({*Prokaryotes, Eukaryotes, Bacterium*}, {*Membrane, Replication*}). In this concept, three objects {*Prokaryotes, Eukaryotes, Bacterium*} constitute the extent of the concept; they are united by their mutual attribute {*Membrane, Replication*} which constitute the intent of the concept. The concept C_1 is more general concept than the concept C_2 = ({*Eukaryotes*}, {*Nucleus*}).

Also on the Fig. 1 there are two different branches of concepts characterizing two families: the viruses and DNA and prokaryotes, eukaryotes and bacteria. This concept demonstrates the fact of separation of objects from the set *G* into two important branches. The link between them is the attribute "*Membrane*". It is known [7] 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. 1 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 sub lattices in the general case.

2.1.1 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 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 [16],
- verb-object dependencies from texts [10],
- words and their lexico-syntactic contexts [20].

These lexical elements must be distinguished 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,
- using corpus tagging and semantic models of texts [10].

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

2.2 The modeling process

Consider in general the process of CGs – FCA modeling. It includes the following steps.

1. *Acquiring a set of conceptual graphs from input texts*. As it is mentioned above conceptual graphs can be acquired from texts by existing information systems. For example they can be created by our system CGs Maker [1]. Some details about it can be found in [3], [5]. We use verb-centered approach for creating conceptual graphs. According to this approach, a conceptual graph is constructed so that there is the central concept in it which is realized as a verb. If there are no verbs in a sentence then method also creates conceptual graph. Verb-centered approach is important for us since it provides predicate forms in the structures of conceptual graphs. These forms are mostly used for representing conceptual graph semantics.

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 following ways of conceptual graphs aggregation: conceptual graphs clustering and using corpus tagging together with support of concept types in conceptual graphs. Clusters of conceptual graphs need to be semantically interpreted which may lead to additional investigations. The second method is more constructive since it selects those conceptual graphs which concepts have mappings to certain domain. Such domain of terms may be presented by corpus tagging or by thesaurus. Some details of aggregation are below.

3. *Creating formal contexts*. This is the central point of CGs – FCA modeling. One or several formal contexts have been built on the aggregated conceptual graphs. The number of formal concepts and the method of building them depend on the problem being solved with CGs – FCA modeling.

4. *Building concept lattice*. Having a formal context as input data, a concept lattice may be created by using various algorithms. There is a field of research in FCA devoted to creating and developing algorithms for concept lattice creation [21]. On the current stage of CGs – FCA modeling technique we use standard solution of creating concept lattice realized in the open source tool [27]. Nevertheless, here there are certain possibilities to create new algorithms oriented on specific structure of formal contexts acquired from conceptual graphs. One of such structure is block-diagonal structure which arises namely on using textual data as input.

5. *Knowledge extraction from concept lattice*. In concept lattice it is possible to identify connections

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

between its concepts according to the principle of "common – particular". Each concept may be interpreted as the set of potential facts of certain level, which is associated with other facts. So the knowledge extracted from concept lattice may be interpreted a s facts.

2.3 Aggregation of conceptual graphs

In the theory of conceptual graphs aggregation means replacing conceptual graphs by more general graphs [24] . 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.

Clustering is a way of aggregation of conceptual graphs. 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 [2] using various similarity measures. There are two known similarity measures proposed in [17], the conceptual similarity

$$s_c = \frac{2n(\gamma_c)}{n(\gamma_1)+n(\gamma_2)} \quad (1)$$

and relative similarity

$$s_r = \frac{2m(\gamma_c)}{m_{\gamma_c}(\gamma_1)+m_{\gamma_c}(\gamma_2)} \quad (2)$$

Here γ_1, γ_2 - conceptual graphs, $\gamma_c = \gamma_1 \cap \gamma_2$ is their common sub graph, $n(\gamma_i)$ - number of concepts of graph γ_i, $m(\gamma_i)$ - number of relations of graph γ_i, $m_{\gamma_c}(\gamma_i)$ is the number of relations of conceptual graph γ_i, at least one of which belongs to the common sub graph γ_c.

If two conceptual graphs have identical concepts then their conceptual similarity has non zero value. Relative similarity is non-zero when two conceptual graphs have identical structures of patterns of conceptual relations.

We used conceptual and relative similarities (1), (2) and their combination in the experiments of conceptual graphs clustering [2]. Except traditional algorithms of clustering such as *K*-means, we used genetic clustering algorithm with special encoding. The peculiarity of implementing genetic algorithms for clustering is that there may be several final solutions i.e. several different variants of clustering.

All numerical characteristics of conceptual graphs clustering results (number of clusters, dimensions of clusters, etc.) are not informative. Clusters of conceptual graphs need to be semantically interpreted. The way of that interpretation depends on the nature of the problem to be solved with conceptual graphs.

Both conceptual and relative similarity measures share a common sub graph γ_c. But two conceptual graphs may have no common sub graph but may be similar "semantically". That means that their concepts have the same type. For example different names of bacteria belong to the type "bacterium" or the type "the name of bacteria".

The second way of conceptual graphs aggregation is based on supporting types of concepts by using external resources. Thesaurus or corpus tagging may be such resource. Section 3 contains additional details.

2.4 Creating formal contexts

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 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.

Fig. 2 shows an example of conceptual graph for the sentence "*Burkholderia phytofirmans belongs to the beta-proteobacteria and was isolated from surface-sterilized glomus vesiculiferum-infected onion roots*."

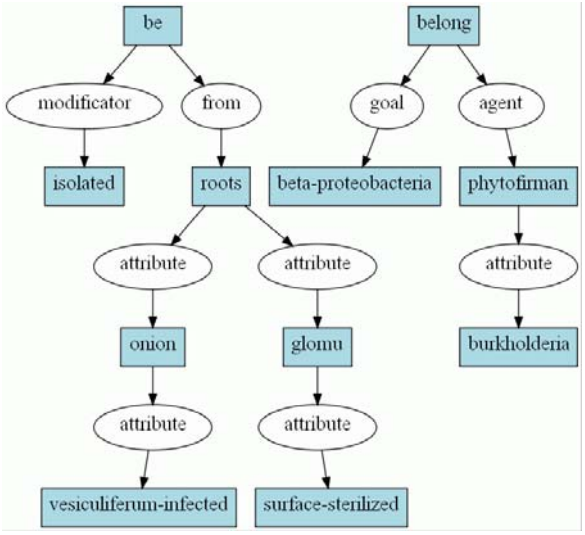

Figure 2 Conceptual graph for the sentence "*Burkholderia phytofirmans belongs to the beta-proteobacteria and was isolated from surface-sterilized glomus vesiculiferum-infected onion roots*."

This graph has five conceptual relations "attribute" but only four of them indicate the objects and attributes valid for formal context. Using "*phytofirman*" as object and "*Burkholderia*" as attribute in the formal context is wrong way because "*Burkholderia phytofirmans*" is known full name of this bacterium [6] and full names of bacteria have to be objects in a formal context devoted to bacteria. Word combinations denoting the names of bacteria must be recognized before the building of conceptual graphs. There is no other way of doing this than to use an external source of information, for example, the corpus tagging. So in this example the sub

graph <*phytofirman*>- (*attribute*) - < *burkholderia* > is useless for bacteria names recognizing.

Remaining elements of conceptual graph on the Fig. 2 are not useless and play significant roles in creating formal context. Conceptual graph on the Fig. 2 represents two facts:

1. bacterium *Burkholderia phytofirmans* belongs to beta-proteobacteria;
2. this bacterium infects the onion.

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 predicate forms <object> - <predicate> - <subject> which in conceptual graphs look as the template <concept>- (patient) - < verb > - (agent) - <concept>. Not only agent and patient semantic roles but also other similar to them (goal on the Fig. 2) roles are allowed in templates.
2. The hierarchy of conceptual relations in conceptual graphs is fixed and taken into account when creating formal context. This hierarchy exists on the Fig.2: relations “agent”, “goal”, “from”, “modificator” are on the top level and relations "attribute" belong to underlying levels. Using this hierarchy of conceptual relations we can 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 [12]. 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. The relative similarity (2) is mostly close to be similarity operation for patterns.

The CGs – FCA modeling technique was tested in various levels of its realization for classification messages in technical support services [3], modeling requirements for information systems [4] and classifying queries to biomedical systems [5].

3 CGs-FCA modeling on biomedical data

3.1 Biomedical data intensive domain

Bioinformatics is the field where Data Mining and Text Mining applications are growing up rapidly. New term of “Biomedical Natural Language Processing” (BioNLP) has been appeared there [1]. This appearing 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 [25]. Information resources of PubMed have been united in several subsystems presenting databases, corpora and ontologies.

So called “research community around PubMed” [14] 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 [11].

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

3.2 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 [6]. 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 CGs – FCA modeling environment has 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.

3.3 BioNLP tasks

According to the BioNLP Shared Task initiative [6] 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.

3.4 NER and anaphora resolution

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:

- 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 bacteria names. We use double-pass algorithm for anaphora resolution which controls so called isolated concepts appeared on the first pass of the algorithm. Isolated concepts are those concepts which are not connected by relation with any other concepts. As a rule they appear when a sentence contains abbreviations or code of bacterium. For example, in the sentence "*Streptococcus thermophilus strain LMG 18311*" there is code of bacterium strain. This code will be presented as isolated concept in conceptual graph. Later in another sentence there is text fragment "*…two yogurt strains of S. thermophiles* …" which has abbreviation of the name of bacterium. Having isolated concept with strain code we can identify it with bacterium using corpus tagging. For resolving abbreviations programming triggers which react to the second word after abbreviation are applied.

To evaluate the quality of this solution of NER the standard characteristics of recall, precision and *F*-score were calculated. To obtain them it was needed to mark named entities manually in the texts used in experiments. The table 1 contains values of recall, precision and *F*-score compared with corresponding values from the work [23]. In this work pattern-based approach is also applied and several external resources were involved in the NER solution. The Alvis system was explored in [23] and SemText is the name of our system which explores CGs – FCA modeling.

Table 1 Recall, precision and *F*-score for NER solutions

	Recall	Precision	*F*-score
Alvis	0,52	0,46	0,59
SemText	0,42	0,53	0,47

The ratio of the values of recall and precision is more informative than their individual ones and is shown on the Fig. 3. According to the table 1 and Fig. 3 we resume that there is medium quality of our solution of NER. It is explained by disability of our algorithm to interpret all possible isolated concepts in conceptual graph. As a result approximately half of marked lexical elements were not recognized as entities.

3.5 Relations 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 on 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 2 shows numerical characteristics of created contexts.

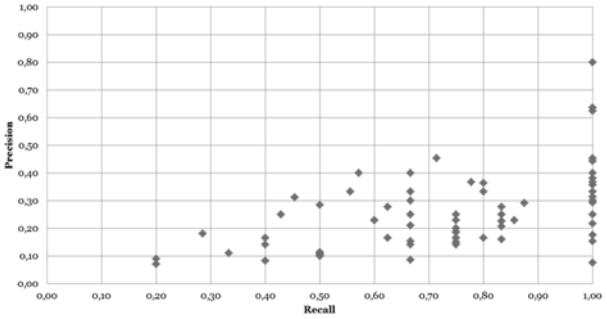

Figure 3 Recall and precision ratio for NER solution on 60 objects

Table 2 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.

For extracting relations we use visualization on the current stage of modeling technique. It allows getting results only for relatively small lattices.

Often relations between concepts in concept lattice may be treated as facts. 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. 4. The lattice on the Fig. 4 contains formal concepts related to the following bacteria: *Borrelia turicatae, Frankia, Legionella, Clamydophila, Thermoanaerobacter tengcongensis, Xanthomonas oryzae*. Highlighted view on the figure illustrates 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 two words in the text according to the rule of marking words denoting extreme situation.

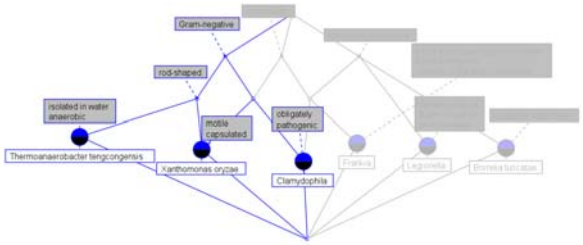

Figure 4 Example of view of gram-negative property of bacteria.

Comparing our results of relations extraction with the known ones from [23] we resume that concept lattice provides principally another variant of solution of this task. In [23] results of relations extraction are presented as marked words in the texts. Visualized concept lattice is more powerful object for investigating relations.

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 as CGs – FCA modeling on textual data show its good potential for knowledge extraction.

In spite of advantage of CGs – FCA modeling there are some problems which need to be solved for improving the quality of modeling technique.

1. Conceptual graphs acquired from texts contain many noise elements. Noise is constituted by the text elements that contain no useful information or cannot be interpreted as facts. Noise elements significantly decrease efficiency of algorithms of CGs – FCA modeling. To exclude noise we need to distinguish textual data that can be excluded from consideration, for example, information about when and by whom a bacterium was first identified.
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 patterns structures and their projections is that way of formalizing CGs – FCA modeling technique.
3. The next stage of developing CGs – FCA modeling is creating fledged information system which process user queries and produce 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.

References

[1] BioNLP 2014. Workshop on Biomedical Natural Language Processing. Proceedings of the Workshop. The Association for Computational Linguistics. Baltimore, 2014. 155 p. http://acl2014.org/acl2014/W14-34/W14-34-2014.pdf

[2] Bogatyrev M., Latov, V., Stolbovskaya. Application of Conceptual Graphs in Digital Libraries. – Proc. RCDL (Digital libraries: advanced methods and technologies, digital collections), pp. 464-468. 2007.

[3] Bogatyrev M., Kolosoff A. Using Conceptual Graphs for Text Mining in Technical Support Services. Pattern Recognition and Machine Intelligence. - Lecture Notes in Computer Science, 2011, Volume 6744/2011, pp. 466-471. Springer-Verlag, Heidelberg, 2011. http://link.springer.com/chapter/10.1007%2F978-3-642-21786-9_75

[4] Bogatyrev, M., Nuriahmetov, V., 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.

[5] 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). http://www.matbio.org/2013/Bogatyrev_8_340.pdf

[6] 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. http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-13-S11-S3

[7] Campbell, N. A., etc., Biology: Concepts and Connections. Benjamin-Cummings Publishing Company, 2005.

[8] Carpineto, C., & Romano, G. Exploiting the Potential of Concept Lattices for Information Retrieval with CREDO. Journal of Universal Computing, 10, 8, 985-1013. 2004.

[9] 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.

[10] 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. http://arxiv.org/pdf/1109.2140.pdf

[11] Edhlund, B., McDougall, A., Pubmed Essentials, Mastering the World's Health Research Database. Form & Kunskap AB, 2014. https://www.amazon.com/PubMed-Essentials-Mastering-Research-Database/dp/1312289457

[12] Ganter, B., Kuznetsov, S.O. Pattern structures and their projections. In: ICCS, pp. 129–142. 2001

[13] Ganter, B., Stumme, G., Wille, R., eds., Formal Concept Analysis: Foundations and Applications, Lecture Notes in Artificial Intelligence, No. 3626, Springer-Verlag. 2005

[14] Hunter L, Cohen KB. Biomedical language processing: what's beyond PubMed? Mol. Cell. 2006. V. 21. P. 589–594.

[15] Ignatov, Dmitry I. and Kuznetsov, Sergei O. and Poelmans, Jonas, Concept-based Biclustering for Internet Advertisement. In: Vreeken, J and Ling, C and Zaki, MJ and Siebes, A and Yu, JX and Goethals, B and Webb, G and Wu, X , Eds., Proc. of 12th IEEE International Conference On Data Mining Workshops (ICDMW 2012), pp. 123-130, 2012.

[16] Meštrović, A. Semantic Matching Using Concept Lattice. Concept Discovery in Unstructured Data, CDUD 2012, pp. 49-58. http://ceur-ws.org/Vol-871/paper_6.pdf

[17] Montes-y-Gomez, Gelbukh, Lopez-Lopez, Baeza-Yates, Flexible Comparison of Conceptual Graphs. Lecture Notes in Computer Science 2113. Springer-Verlag, 2001.

[18] Obitko, M., Snasel,V., Smid, Jan. Ontology Design with Formal Concept Analysis. – Proc. of the CLA 2004 International Workshop on Concept Lattices and their Applications. http://ceur-ws.org/Vol-110/paper12.pdf

[19] Olivé, Antoni, Conceptual Modeling of Information Systems. Springer-Verlag. Berlin, Heidelberg, 2007. https://www.amazon.com/dp/3540393897/ref=rdr_ext_tmb

[20] 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.

[21] 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.

[22] 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

[23] 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. http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-13-S11-S8

[24] Sowa, J.F., Conceptual Structures: Information Processing in Mind and Machine. Addison-Wesley, London, UK. 1984

[25] U.S. National Library of Medicine. http://www.ncbi.nlm.nih.gov/pubmed .

[26] 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

[27] ConExp-NG. https://github.com/fcatools/conexp-ng