

Application of Formal Contexts in the Analysis of Heterogeneous Biomedical Data

Mikhail Bogatyrev^a, Dmitry Orlov^a

^aTula State University, Tula, Russia

Abstract

The paper proposes a method of conceptual modeling based on the use of formal contexts. Formal context is the main notion in the Formal Concept Analysis (FCA), the lattice-based data analysis approach. Biomedical data and the tasks of their analysis under Biomedical Natural Language Processing are discussed. Two variants of formal contexts constructed on natural language texts are considered. They are the contexts constructed with the use of keywords and n -grams. Textual n -grams are acquired by using conceptual graphs and Abstract Meaning Representation (AMR) schemata. Both contexts are used in the text clustering task. It is shown that the classical FCA clustering on keyword based context entails appearing bag-of-words in clusters. It is proposed the clustering approach for n -grams based multidimensional contexts which avoid appearance of a bag-of-words in clusters. The method was tested on the texts of annotations of scientific articles from PubMed databases.

Keywords

conceptual modeling, conceptual graphs, Biomedical Natural Language Processing, polyadic formal context

1. Introduction

Data analysis tasks are diverse. A task that is often a priority in data analysis is clustering. Clustering allows one to combine data into subsets-clusters according to some proximity measure of data, which simplifies their further analysis. This work relates to two areas that are directly and indirectly related to clustering: Formal Concept Analysis (FCA) [1] and Text Mining. In the classical FCA, the sets of data related by “object-attribute” relationship is studied. This data forms the *formal context* on which the concept lattice is built. The formal concepts that make up the lattice are linked by a general-private relationship and form a hierarchical conceptual data model. A special feature of the FCA is the mathematical rigor of the proposed solutions and their universality. The formal context, the central notion of FCA, is defined on arbitrary sets, so it can be applied to data of any nature. This advantage of the FCA has the opposite side – the need to adapt the FCA to specific tasks, which often requires special research. As a result of such research, many applications of FCA are known in a variety of fields, from biology and natural sciences, software engineering and public networks to computational linguistics. The FCA application review [15] contains a structured analysis of them and is still relevant. The current state of FCA is characterized by the use of multidimensional formal contexts and


Russian Advances in Artificial Intelligence: selected contributions to the Russian Conference on Artificial intelligence (RCAI 2020), October 10-16, 2020, Moscow, Russia

✉ okkambo@mail.ru (M. Bogatyrev)

ORCID 0000-0001-8477-6006 (M. Bogatyrev)

© 2020 Copyright for this paper by its authors.

Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

corresponding generalizations of FCA concepts in data analysis problems. Formal concepts based on two-dimensional formal contexts represent a special solution to the data clustering problem: biclustering [10]. The use of multidimensional formal contexts extends cluster analysis to three-dimensional and n -dimensional clustering. An important result here is progress in solving the well-known problem of cluster interpretability: n -dimensional clusters are more representative than normal ones, since they are built simultaneously on several sets.

In this paper, FCA clustering on heterogeneous biomedical data is investigated. The heterogeneity of biomedical data consists in the use of texts along with numerical values. This can be text designations of drugs, genes, bacteria, etc., as well as natural language texts. We consider variants of formal contexts for text clustering problems, including contexts have been constructed using n -grams. We propose a method for constructing formal contexts on textual data, in which n -grams are obtained from conceptual graphs and correspond to the model of Abstract Meaning Representation of text [6]. Then constructed formal contexts are used in the task of clustering biomedical data.

It is shown that the use of standard FCA clustering algorithms on such contexts leads to the appearance “bag-of-words” in clusters, which makes it difficult to interpret them. A version of clustering using n -gram data associations is proposed, which allows avoiding “bag-of-words” and allows interpreting clusters in the context of data queries in the form of meaningful phrases.

The paper is organized as follows. Section 2 briefly introduces the main definitions of Formal Concept Analysis. In the Section 2.1 polyadic formal contexts and multimodal clusters are described. Section 3 contains brief review of biomedical data analysis including Biomedical Natural Language Processing in the Section 3.1. Section 4 is devoted to constructing polyadic formal contexts on natural language texts. The use of keywords as attributes in formal context is described in the Section 4.1 and in the Section 4.2 we discuss semantic features of formal contexts. In Section 5 results of experimental study of application of two variants of formal contexts for clustering are presented. Section 6 devoted to comparing our results with some ones in related work. In Section 7 we conclude and discuss the future work.

2. Elements of Formal Concept Analysis

Briefly consider the main definitions of the FCA. Classical FCA [1] deals with two basic notions: *formal context* and *concept lattice*. Formal context is a triple $\mathbf{K} = (G, M, I)$ where G is a set of objects, M is a set of their attributes, $I \subseteq G \times M$ – binary relation which represents facts of belonging attributes to objects. Formal context may be represented by $[0, 1]$ -matrix $\mathbf{K} = \{k_{i,j}\}$ in which units denote relationship between objects $g_i \in G$ and attributes $m_j \in M$. The concepts in the formal context are defined in 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 the properties of $A' := \{m \in M \mid \langle g, m \rangle \in I \text{ for all } g \in A\}$ and $B' := \{g \in G \mid \langle g, m \rangle \in I \text{ for all } m \in B\}$ then the pair of subsets (A, B) like that $A' = B, B' = A$ are called formal concepts. The sets A and B called the *extent* and the *intent* of a formal context $\mathbf{K} = (G, M, I)$ respectively.

In 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. If for formal concepts (A_1, B_1) and (A_2, B_2) , $A_1 \subseteq A_2$ and $B_2 \subseteq B_1$ 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 makes a lattice, which is called *concept lattice*. A lattice is a partially ordered set in which every two elements have a *supremum* (also called a least upper bound or *join*) and a *infimum* (also called a greatest lower bound or *meet*).

2.1. Polyadic FCA

Polyadic or multidimensional FCA is based on the notion of multidimensional formal context. A multidimensional, n -ary formal context is defined by a relation $R \subseteq D_1 \times D_2 \times \dots \times D_n$ on data domains D_1, D_2, \dots, D_n . The context is an $n+1$ set:

$$\mathbb{K} = \langle K_1, K_2, \dots, K_n, R \rangle, \quad (1)$$

where $K_i \subseteq D_i$. Every n -ary context begets k -ary contexts, whose number is given by the Stirling formula $S(n, k) = \frac{1}{k!} \sum_{i=0}^k (-1)^i \binom{k}{i} (k-i)^n$ [7]. As it is shown in [19], multidimensional n -ary context also contains formal concepts which also form a lattice.

Already on two-dimensional formal contexts, and especially on multidimensional ones, not only formal concepts are of interest, but also “insufficiently dense concepts” which are two-dimensional, three-dimensional, and n -dimensional clusters. These clusters may contain useful information.

By introducing the notion of bicluster density [11], one can investigate various biclustering options and estimate their significance [12]. An important result here is the statement proved in [11] that each concept contains a cluster, but the opposite is not true.

Clustering on multidimensional formal contexts is called *multimodal clustering*.

According to multimodal clustering, for any dimension of formal context, the purpose of its processing is to find n - sets $H = \langle X_1, X_2, \dots, X_n \rangle$ which have the closure property [7]:

$$\forall u = (x_1, x_2, \dots, x_n) \in X_1, X_2, \dots, X_n, u \in R, \quad (2)$$

$\forall j = 1, 2, \dots, n, \forall x_j \in D_j \setminus X_j \langle X_1, \dots, X_j \cup \{x_j\}, \dots, X_n \rangle$ does not satisfy (2). The sets $H = \langle X_1, X_2, \dots, X_n \rangle$ constitute *multimodal clusters*.

When solving any clustering problem, a proximity measure of the objects being clustered is used. In FCA clustering, the proximity of objects is set by their relation R , so it is actually written in a formal context: those objects are close to each other that have common attributes, and vice versa for attributes and objects.

3. Tasks and Methods of Heterogeneous Biomedical Data Analysis

One of the areas where NLP applications are becoming more in demand is Bioinformatics. Data in Bioinformatics is often heterogeneous: it includes both numeric and symbolic sequences, as

well as texts. The Biomedical Natural Language Processing (BioNLP) [8] is the new area of research in Bioinformatics which appearance was due to the avalanche-like growth of publications in the field of biomedicine.

3.1. Biomedical Natural Language Processing

The main purpose of BioNLP is to obtain new knowledge from published texts, not completely contained in each individual publication. Initially, the main area of application of BioNLP methods was genomic studies. Over time, the subject matter of texts processed by BioNLP has expanded to other areas. BioNLP was formed as research area with its own data, tasks and methods [8, 9]. They are summarized as follows.

BioNLP Data and Resources. The main types of data used in BioNLP are the texts of scientific publications – usually abstracts of these publications. Along with lexical and grammatical elements common for non-structured texts, they have their own specifics: characteristic terms, for example, names of genes, ab-breviations, and the inclusion of numerical data in the text. Distinctive feature of biomedical data is synonymy. The same concept may be expressed using different words in a text. For example, “heart attack” and “myocardial infarction” refer to the same medical problem.

Natural language texts are implied to be unstructured. The peculiarity of biomedical textual data is that it is actually semi-structured. The most used resource of the modern BioNLP systems is the PubMed system [20]. This is online knowledge base that includes many special databases. Simultaneously, PubMed is both ontology and a text corpus with extra linguistic tagging. Tagging options are limited to hyperlinks to other publications cited in one work, as well as to publications of similar content. The tagging makes it easier to solve a number of problems on corpus data, for example, solving clustering problems.

Knowledge Sources. In addition to the data itself, biomedical information resources contain the so called sources of knowledge. Among them there is the Medical Subject Headings (MeSH) [13], which contains controlled vocabulary terms organized as tree structure. Another important resource is the Unified Medical Language System [18]: the compendium of controlled vocabularies. It has knowledge source databases and associated software tools for use by BioNLP systems developers.

Thus, almost any task of BioNLP is solved not only directly on the texts, but also with the involvement of external resources.

BioNLP Tasks and Methods. All the BioNLP tasks may be classified as more or less general. The general task of knowledge extraction is transformed to the tasks of fact extraction and event extraction. Often these terms are not distinguished [2]. As atomic tasks, being solved as a part of the solution of more general task there are tasks of Named Entity Recognition and Relation Extraction.

Named Entity Recognition. Named Entity Recognition (NER) is the standard task of BioNLP. It consists in automatically identifying occurrences of biological or medical terms in unstructured text. As named entities, there are the names of genes, proteins, living organisms or diseases – it is depended on the domain to which processed text belongs to.

NER is typically consisted of three-stages process [9] that involves:

- determining an entity’s substring boundaries within the text,

- assigning the entity to a predefined class or category, and
- selecting the preferred name or unique identifier of the concept that the entity names.

The performance of NER solutions is measured in terms of precision, recall, and F-score [9] since this task can be interpreted as classification task.

Relation Extraction. Relation Extraction (RE) is another standard task of BioNLP. Relations are associations among biomedical entities. The simplest relations are binary, involving only the pair-wise associations between two entities. But biomedical relationships can involve more than just two entities. This kind of relationship is actual in the task of event extraction. In our time, named as genomic era, much of BioNLP work has focused on automatically extracting interactions between genes and proteins. Other associations include interactions between proteins and mutations, proteins and their binding sites, genes and diseases, genes and phenotypic context [8, 9].

Events (Facts) Extraction. As it was noticed events and facts are often not distinguished in the BioNLP literature. But strictly it is appropriate to consider a fact as a static object and to attribute some duration to an event. Additional distinction is that events can be nested.

It is also known from BioNLP literature that events are typically characterized by verbs or nominalized verbs in the text [6, 10]. This is roughly true because there exists a verb-centric model of the meaning, according to which the meaning of the sentence is primarily reflected by verbs. But in general, facts or events are identified by means of objects that are external to the text.

BioNLP Methods. All methods for solving BioNLP problems can be divided into two types: methods that work at the level of individual words and sentences of the text, and methods that use models that are external to the text. Such models include syntactic models, for example, parse trees, as well as text semantics models. The methods working "inside" the text are typical for the tasks of linguistics, where, for example, the peculiarities of the use of certain lexical elements in the texts are studied.

The tasks of information retrieval usually require the involvement of both types of methods. For example, the solution of the NER problem described in the previous paragraph includes three stages. The second and third stages of the solution involve the use of external data and conceptual models.

The extraction of the boundaries of the entity influence in the text is based on the linguistic concept of context. A context is a region of text that surrounds selected elements of a sentence, and is usually associated with a specific content of a fragment of text. The notion of formal context used in FCA is not tied to specific boundaries in the text that significantly expands the possibilities of text analysis, using this notion.

Text Mining as general technique is applied in BioNLP systems. However, special approaches and methods are being developed here [2, 8, 9].

Statistical approach is the oldest one in BioNLP and has been applied as in the NER as in the RE tasks. It is based on the idea that if the entities are repeatedly mentioned together, then there is a greater chance that they may be related in some way. But the type and direction of this relation cannot be determined by co-occurrence statistics only.

Rule-based approach uses the linguistic patterns connected with particular relations. Unlike the systems based on statistical term co-occurrences, rule-based approach demonstrates high

precision and low recall [6]. The rules used for relation extraction can be manually defined by domain experts, or they can be derived from annotated corpora by machine learning algorithms

Classification-based approach together with Dictionary-based methods is also frequently used to identify relations involving medical entities. Dictionaries, thesauri and ontologies constitute the set of external resources which have been applied here [21, 22].

4. Constructing Polyadic Formal Contexts

Consider two approaches to building formal contexts on textual data: using keywords and using n -grams based on conceptual graphs. These approaches are not depended on biomedical data domain, they use features of any text but standard tasks of BioNLP.

4.1. The Use of Keywords

Keywords are a long-standing and frequently used tool in linguistics and Text Mining. They are still used in modern models, for example, in the Word2Vec model [14] and in thematic text representation models. They construct vectors containing the frequency of occurrence of keywords in texts, which are compared using a proximity measure – often the cosine of the angle between the vectors. Similar proximity measures are also used in text clustering tasks.

Consider a formal context $\mathbf{K}_w = (T, W, I)$ that is built using keywords. Let T be a set of texts and W be a set of keywords. The context matrix is binary, with elements that reflect the fact that keywords belong to certain texts. Each formal concept (A, B) in this context is a combination of elements of sets T and W : $A \subseteq T, B \subseteq W$. It reflects the division of texts into subsets according to the occurrence of keywords. The set of formal concepts $\bigcup_i (A_i, B_i)$ forms a lattice that defines the hierarchy of texts according with the presence keywords in them. This solution to the text clustering problem is constructed without the use of traditional linguistic proximity measures. The advantages of this clustering method compared to standard clustering methods are the absence of the need to set the number of clusters in advance and the presence a cluster hierarchy in the form of concept lattice. Compared to standard hierarchical clustering methods, this method works faster because it does not require multiple calculations of the proximity measure.

The disadvantage of using keywords in formal context will be the appearance of a “bag-of-words” in clusters, because, regardless of the method of obtaining keywords from texts, they are sets of words that are not related in meaning. If more than one text is included in the concept, then, having “bag-of-words” in the concept, one can determine what word belongs to what text only by referencing to the original formal context. So the problem of interpreting results of clustering has no solution in this case.

4.2. Preservation of Semantics in Formal Contexts

To avoid “bag-of-words”, we need to apply in formal context the objects that reflect the semantics of texts to some extent. These objects include n -grams, the sets of words in the form of sequences that have a certain meaning.

In this paper, we use n -grams extracted from texts by constructing conceptual graphs [17] and corresponding to the Abstract Meaning Representation (AMR) of the text [6].

Conceptual graphs constitute semantic model of text that belongs to the class of semantic networks. They play an important role as a conceptual modeling tool in the fields of mathematical linguistics, bioinformatics, and mathematical logic.

A conceptual graph is a finite oriented connected bipartite graph [17] which has two different kinds of nodes: concepts and conceptual relations. Figure 1 shows a fragment of conceptual graph for one of the sentences of the processed texts in our experiments together with the marked elements of the AMR scheme. In the conceptual graph in Fig. 1 concepts are represented by rectangles, and conceptual relationships are represented by ellipses. We used conceptual graphs in a number of studies [3, 4]. We obtain conceptual graphs similar to the one shown in Fig. 1 using a method [3] which is based on the solution of the Semantic Role Labeling problem [5]. The algorithm of acquiring conceptual graphs from text has the following main steps.

1. Dividing the text into sentences.
2. Dividing sentences into words, punctuation marks, and other symbols. Deleting stop words.
3. Determining morphological features of words in sentences.
4. Defining semantic roles as conceptual relations in conceptual graph. At this stage, lexico-semantic templates are used.
5. Constructing conceptual graph visualization.

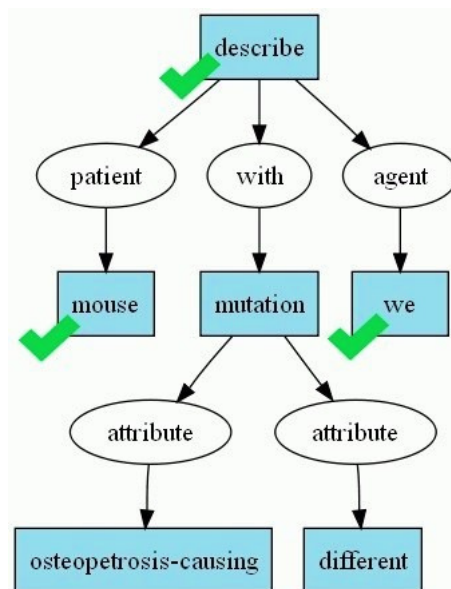


Figure 1: Fragment of the conceptual graph for the sentence “We describe a mouse with a different osteopetrosis-causing mutation.”

The Abstract Meaning Representation of a text [6] is defined as a directed tree graph that

fixes a certain concept in the text in such a way that sentences that have the same meaning from the point of view of this concept have the same AMR graph. The AMR graph usually corresponds to an AMR schema in the form of a phrase, for example: “*who*” – “*what does*” – “*with whom*”. This scheme is three-element one. An AMR diagram can correspond to the entire text or to individual sentences and may have various numbers of elements. There are several approaches to building an AMR of a text.

Using conceptual graphs allows one to build quite complex AMR schemata. An AMR scheme is constructed as a tuple $\langle c_1, c_2, \dots, c_n \rangle$, the elements of which are the concepts of a conceptual graph connected by conceptual relations corresponding to the meaning of the phrase of AMR scheme. So for the AMR scheme “*who*”–“*what does*”–“*with whom*” such relations are the well-known semantic roles of “*agent*” and “*patient*”. In the Fig. 1, the elements of the “*who*”–“*what does*”–“*with whom*” AMR scheme are marked, which is made up with the concepts $\langle \text{“SHP-2”}, \text{“attenuates”}, \text{“function”} \rangle$, together forming a meaningful phrase.

Conceptual graphs allow us to define AMR schemes uniquely in the form of, in which the length of the scheme n is equal to the number of conceptual graph concepts involved in constructing the scheme. As a result, such n -grams constitute meaningful phrases.

A polyadic formal context based on AMR schemata is constructed as follows.

1. A conceptual graph is constructed for each sentence of the processed text.
2. A specific AMR schema is created based on the elements of the conceptual graph.
3. The formal context is constructed as a multidimensional tensor. Its points $k_{i,j,\dots,n} = \{c_i, c_j, \dots, c_n\}$ are the elements $c_k, k = 1, 2, \dots, N$ of the AMR scheme for each sentence, N – the total number of concepts obtained on the processed text.

The number of points in the formal context matches the number of AMR schemata found in the text.

The vast majority of points in the formal context are meaningful phrases, which is an important feature of this method.

5. Experimental studies

Experimental studies of the developed approach were performed on the texts of the Active Gene Annotation Corpus (AGAC), which contains abstracts of scientific articles on the biomedical topics of the PubMed system. The corpus was created for the BioNLP Shared Tasks 2019 competition and was offered as a data set for NER and RE extraction tasks. The corpus contains 1000 unprocessed abstracts, and its size is about 300,000 tokens.

The considered approach to clustering using formal contexts was applied in the task of studying the interrelations of texts.

5.1. Clustering using keywords

The first variant of clustering was performed using keywords. The experiment included the following stages.

1. Finding keywords in the set of texts.
2. Constructing formal context $\mathbf{K} = (G, M, I)$ where G is the set of text names, M is the set of keywords.

3. Generating concept lattice for the context being used.

4. Comparing results of clustering with the standard k -means clustering.

Options for limiting the number of keywords in the range from 5 to 20 words were studied.

As a result the following regularity is obtained. Increasing the number of keywords leads to appearing more links between texts. The link configuration in the concept lattice has a hierarchy that allows one to evaluate the generality/particularity of texts in terms of their keywords.

In order to avoid bulky presentation, we will limit it to presenting results for five texts. Figures. 2.2 a,b show in the form of concept lattices the examples of clustering constructed for five texts and two sets of 5 and 20 keywords. The figures show the increasing the number of links between texts depending on the number of keywords: concept lattice on the Fig. 2.2 b) is more multilinked than one on the Fig. 2.2 a).

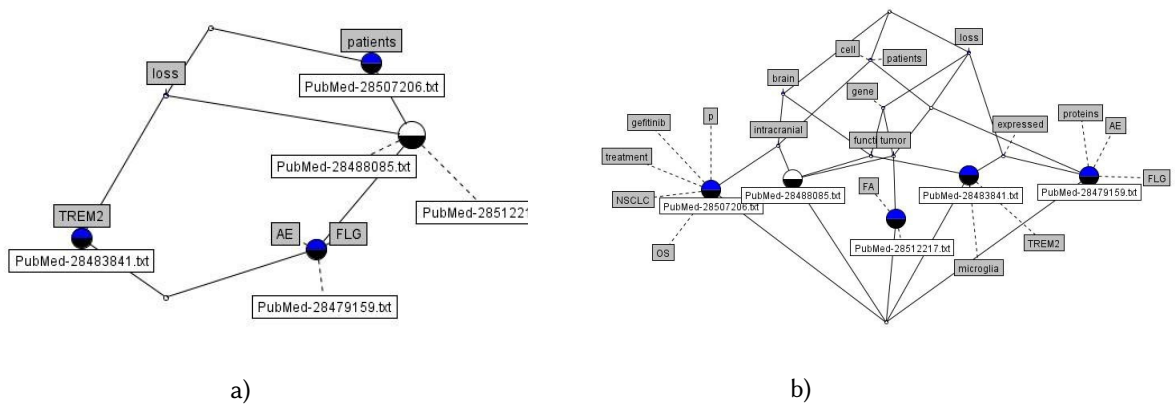


Figure 2: Concept lattices constructed for five texts and two sets of 5 (a) and 20 (b) keywords.

Formal concepts in both lattices contain “bag-of-words”. So for solving the task of NER additional analysis of words belonging to formal concepts is required. The task of relation extraction (RE) has general solution acquired from the lattices demonstrating how texts are linked by keywords.

Comparing with k -means clustering. As it is known, k -means clustering method produces as many clusters as the k -variable specifies. In our example $k_{min} = 1, k_{max} = 5$. The formal context on the Fig. 3 demonstrates what words belong to what texts.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
	TREM2	patients	loss	FLG	AE	OS	p	gefitinib	FA	intracranial	gene	microglia	brain	cell	expressed	treatment	NSCLC	tumor	function	proteins	
PubMed-28479159.txt		X	X	X	X	X									X	X				X	X
PubMed-28483841.txt	X																				
PubMed-28488085.txt		X	X																		
PubMed-28507206.txt			X				X	X	X	X	X						X	X			
PubMed-28512217.txt			X							X	X								X		

Figure 3: Formal context for five texts and 20 keywords.

It is clear from the context that all five variants of clustering are possible. The concept lattice

in figure 2.2. b) also shows all possible variants of clustering.

Indeed, the five concepts in figure 2.2 b) that have text names in light rectangles make up five clusters. Next, we see that, for example, two concepts with the names of texts *PubMed-28507206.txt* and *PubMed-28488085.txt* can form a single cluster if we take into account their common keyword *intracranial*, located higher in the lattice. This principle of fixing clusters is applied to the whole concept lattice: moving up the lattice, we add keywords to the concepts in the next node, leaving unchanged the text names that were in the lower nodes.

Thus, the FCA clustering with the use of keywords demonstrates the advantage over *k*-means clustering: FCA clustering potentially reveals all variants of clustering.

5.2. Clustering using polyadic formal contexts

In the next experiments, formal contexts constructed using *n*-grams, as it is described in Section 4.2 were also studied. Three-, four-, and five-element *n*-grams were used, which are constructed according to the conceptual graphs and correspond to AMR schemata. Fig. 4 shows five-element AMR scheme which was used to construct the formal context. Formal contexts obtained on such *n*-grams are *n*-dimensional tensors whose points are combinations of words that have a certain meaning. FCA clustering of such contexts is possible by known FCA algorithms, for example OAC [12] or Data Peeler [7] ones. However, such clustering will again lead to the appearance of “bag-of-words” in clusters. In fact, if a point $k_{i,j,\dots,n} = \{c_i, c_j, \dots, c_n\}$ in a formal context falls into a cluster, its elements (words) are combined into subsets with words from other points, forming the following cluster structure:

$$C = \underbrace{\{\{c_i, \dots, c_j\}, \dots, \{c_k, \dots, c_l\}\}}_n \quad (3)$$

In expression (3), the sublists contain “bag-of-words”, *n* is the length of the *n*-gram. Therefore, instead of the standard FCA clustering, a different version of clustering was used, focused on application in question-answering systems. In such systems, the formal context is the basis of the information resource that the system users access. User queries are texts in the form of phrases that may correspond to AMR schemes. To process such queries, the structure of an *n*-dimensional formal context is transformed into a set of associations.

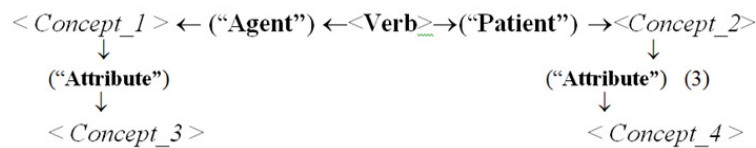


Figure 4: Example of a five-element AMR scheme. Verb – the main verb, the concept in conceptual graph. “Agent”, “Patient”, “Attribute” – semantic roles; Concept_1, 2, 3, 4 – concepts of the conceptual graph.

An Association is a set of points of formal context ordered relative to the selected position of the AMR scheme. This corresponds to the logic of the AMR scheme: associations combine certain

semantic elements in it. The Association includes all words in the selected position of the AMR scheme. Therefore, an Association is a cluster built on the basis of the proximity measure “belong to position of the AMR scheme as a certain grammatical element of a sentence”. On the other hand, Association is a function $A(x_1, \dots, x_p)$ whose argument can be a given word or a set of p words belonging to the k -th position of the AMR scheme.

Associative queries are made to associations – queries that fix one of the variables x_1, \dots, x_p . Responses to such queries contain data belonging to points in the formal context according with semantic meanings of the variable words x_1, \dots, x_p . The use of associations allows to avoid “bag-of-words” in text processing.

The experiments included the following stages.

1. Building associations based on the selected positions of the formal context AMR scheme.
2. Creating queries to associations based on keywords that interest the user.
3. Getting query results as clusters containing points in the formal context.
4. Interpretation of the clusters.

Fig. 5 shows a fragment of the association built for the second position (Concept_1) of the AMR scheme on the Fig. 4, in which the keyword “mutation” is highlighted. The natural numbers in the Association lists are the numbers of the texts that AMR schemata are based on. Most of the corpus texts are devoted to the study of various manifestations of mutation and its impact on organisms. Therefore, queries to associations were made using the keyword “mutation”. The query results are clusters that were also used for building nested associations.

```

mutation → {{ novo, mutation, play, important, role, 27},
             {de, mutation, play, important, role, 27},
             {gof, mutation, have, cardio-protective, effect, 33},
             {loss-of-function, mutation, suppress, jnk, signaling, 34},
             {novo, mutation, disrupt, rna, splicing, 38},

```

Figure 5: Fragment of an association based on the AMR scheme.

The question that determines further actions with the resulting clusters is: *How does the mutation manifest itself?*. Implementation of this query on clusters was performed by building associations relative to the fourth position for the five-element AMR scheme. The keywords and text numbers obtained in the constructed associations were then presented for analysis. The answers to the queries of the associations are generated in table form. If the result of a query to associations for two elements is presented as a cross-table, it is interpreted as a two-dimensional formal context. In this case, it can be visualized as a concept lattice according to the classic FCA.

Fig. 6 shows classical concept lattice based on the results of a query to associations for two elements.

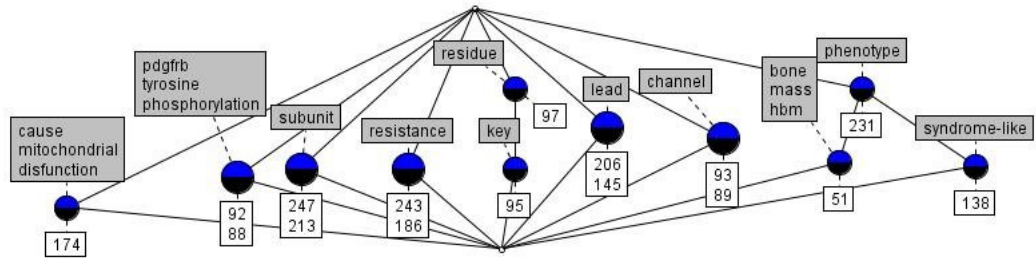


Figure 6: Concept lattice based on the results of a query to associations.

The lattice in Fig. 6 allows one to evaluate the relationships of texts in the context of the word "mutation" and by using the words used in the texts, marked in colored rectangles that are attributes of text objects according to FCA. Some lattice concepts are related hierarchically - these are concepts that include texts 95, 97, 51, 138, and 231. When using standard FCA clustering, the data that makes up such hierarchically related concepts would generate "bag-of-words" in clusters. The use of associations and further visualization them in the form of a concept lattice allows one to correctly investigate the relationship between texts.

Using this kind of clustering, the solution to the Named Entity Recognition (NER) and Relationship Extraction (RE) tasks becomes more defined. These tasks may be solved with the use of corresponding associations. Associations that are built related to the "who" and "with whom" elements of AMR scheme are the most suitable for the NER task and for the RE task the element "verb" of AMR scheme together with the "who" and "with whom" elements is suitable too.

6. Related Work

As it was mentioned previously this work relates to two areas of FCA and BioNLP, where there is a significant number of works devoted to text analysis and clustering. There are not many works devoted to text analysis among them. The work [23] contains a description of the general FCA approach to the problems of linguistics. Other works, for example, work [24], are devoted to solving individual problems there. Our work differs in that it uses a special formal context, which is constructed using n -grams that have a semantic meaning. The use of n -grams is common in text analysis, but the use of conceptual graphs for this purpose and the production of meaningful n -grams, respectively, is not described in the works.

FCA establishes its own approach to clustering based on flat and polyadic formal contexts and supported by various algorithms of constructing concept lattices [7, 11, 12]. Review [25] contains descriptions of almost all FCA models and methods.

Another approach to clustering, often applied in text analysis, is based on the use of vector models Word2Vec, Doc2Vec, etc. [14, 26]. They construct vectors containing the frequency of occurrence of keywords in texts, which are compared using a proximity measure – often the cosine of the angle between the vectors. FCA clustering has the advantage of not requiring an Euclidean proximity measure of objects being clustered. The clustering used in this work is distinguished by the fact that it has “semantic coloring”. Associations are built in such a way that they reveal the contents of data, fixing their topic in the form of the semantic position of the AMR scheme.

Abstract Meaning Representation is applied in BioNLP works, for example, in [27], where this method is used to extract combinations of words interpreted as events. This paper develops this approach towards application it in question-answering systems.

7. Conclusion

This paper describes a method for clustering multidimensional formal contexts built on natural language tests. The method uses conceptual graphs as data source for AMR schemes. It should be noted that the use of conceptual graphs allows building AMR schemata of greater length than those considered in the paper. This will allow one to implement multidimensional formal contexts that reflect the content of the modelled text more fully and, accordingly, extract more complete information from it. This method can be used in question-answering systems where queries in natural language correspond to the logic of AMR schemes.

The method for constructing formal contexts on textual data, in which n -grams are obtained from conceptual graphs and correspond to the model of abstract meaning representation of the text is proposed.

The novelty of the work is as follows. First, we applied a special type of formal context based on the use of n -grams. Second, conceptual graphs are used to extract meaningful n -grams from the text. Third, non-standard clustering in the form of building associations is used to avoid “bag-of-words” as a result of texts clustering.

The future of this work is oriented to realization its results in prototype of question-answering system.

Acknowledgments

The reported study was funded by Russian Foundation of Basic Research according to research project № 19-07-01178 and RFBR and Tula Region according to research project № 19-47-710007.

References

- [1] Ganter, Bernhard; Stumme, Gerd; Wille, Rudolf. *Formal Concept Analysis: Foundations and Applications*. Lecture Notes in Artificial Intelligence, No. 3626, Springer-Verlag, Berlin, 2003.

- [2] Ananiadou, S., Pyysalo, S., Tsujii, J. and D. B. Kell. *Event extraction for systems biology by text mining the literature*. // Trends in Biotechnology, Vol. 28. No 7. 2010.
- [3] Bogatyrev, M.Y., Mitrofanova, O.A., Tuhtin, V.V. *Building Conceptual Graphs for Articles Abstracts in Digital Libraries*. In: Proceedings of the Conceptual Structures Tool Interoperability Workshop (CS-TIW 2009) at 17th International Conference on Conceptual Structures (ICCS'09), pp. 50-57, 2009.
- [4] Bogatyrev, M. *Fact Extraction from Natural Language Texts with Conceptual Modeling*. // Communications in Computer and Information Science. Vol. 706. Springer-Verlag, 2017
- [5] Gildea, D., Jurafsky, D.: *Automatic Labeling of Semantic Roles*. In: Computational Linguistics, 2002, vol. 28. 2002.
- [6] Bos, J.: *Expressive Power of Abstract Meaning Representations*, Computational Linguistics 42(3), 2016.
- [7] Cerf, L., Besson, J., Robardet, C., and Boulicaut, J. F. *Closed patterns meet n-ary relations*. //ACM Trans. Knowl. Discov. Data. 3, 1, 2009.
- [8] Cohen, K. B., Demner-Fushman, D.: *Biomedical Natural Language Processing*. John Benjamins Publishing Company, Philadelphia 2014.
- [9] Demner-Fushman, D., Cohen, K., Ananiadou, S. Tsujii, J. *Proceedings of the 18th BioNLP Workshop and Shared Task*. Association for Computational Linguistics. 2019.
- [10] Hartigan J A. *Direct clustering of a data matrix*. // Journal of the American statistical association, Vol. 67, no. 337. 1972.
- [11] Ignatov D. I., Kuznetsov S. O., Zhukov L. E., Poelmans J., *Can triconcepts become triclusters?* // International Journal of General Systems, Vol. 42. No. 6, 2013.
- [12] Ignatov D. I., Gnatyshak D. V., Sergei O. Kuznetsov, Boris G. Mirkin, *Triadic Formal Concept Analysis and triclustering: searching for optimal patterns*. In: Machine Learning, April, 2015.
- [13] *Medical Subject Headings*, <https://www.nlm.nih.gov/mesh/meshhome.html>
- [14] Mikolov, T., Chen, K, Corrado, G., Dean, J. *Efficient estimation of word representations in vector space*. In: arXiv preprint arXiv:1301.3781. 2013.
- [15] Poelmans J., Ignatov D. I., Kuznetsov S., Dedene G. *Formal concept analysis in knowledge processing: A survey on applications* // Expert Systems with Applications. 2013. Vol. 40. No. 16.
- [16] Simpson M.S., Demner-Fushman D.: *Biomedical Text Mining: A Survey of Recent Progress*. In: Charu C. Aggarwal and ChengXiang Zhai, Editors. Mining Text Data. Springer. 2011.
- [17] Sowa, J.F., *Knowledge Representation: Logical, Philosophical, and Computational Foundations*, Brooks Cole Publishing Co., Pacific Grove, CA. 2000.
- [18] *18. Unified Medical Language System*, <https://www.nlm.nih.gov/research/umls/>
- [19] Voutsadakis, G. *Polyadic concept analysis*. – Order. Vol. 19 (3). 2000
- [20] *U.S. National Library of Medicine*, <http://www.ncbi.nlm.nih.gov/pubmed>
- [21] *Biomedical natural language processing. Tools and resources*, <http://bio.nlplab.org>
- [22] *MetaMap, a Tool For Recognizing UMLS Concepts in Text*, <https://metamap.nlm.nih.gov>
- [23] Priss, U., *Linguistic Applications of Formal Concept Analysis*. In: Ganter; Stumme;

- Wille (eds.), *Formal Concept Analysis, Foundations and Applications*. Springer-Verlag. LNAI 3626. 2005
- [24] Falk, I., Gardent, C.: *Combining Formal Concept Analysis and Translation to Assign Frames and Thematic Grids to French Verbs*. In: Napoli, A., Vychodil, V. (eds.): *CLA 2011*. INRIA Nancy Grand Est and LORIA. 2011.
- [25] Poelmans J., Kuznetsov S., Ignatov D. I., Dedene G. *Formal Concept Analysis in knowledge processing: A survey on models and techniques // Expert Systems with Applications*. Vol. 40. No. 16. 2013.
- [26] Clark, S.: *Vector Space Models of Lexical Meaning*. In: Lappin, Sh., Fox, Ch. (eds.) *The Handbook of Contemporary Semantic Theory*, pp. 493-522. Blackwell Publishing, Ltd. 2015.
- [27] Sudha Rao, Daniel Marcu, Kevin Knight, Hal Daum´e III. *Biomedical Event Extraction using Abstract Meaning Representation*. //Proc. of the BioNLP 2017 workshop, Vancouver, Canada, August 4, 2017.