Dmitry Ignatov, Sergei Kuznetsov, Jonas Poelmans (Eds.)

CDUD'11 – Concept Discovery in Unstructured Data

Workshop co-located with the 13th International Conference on Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing (RSFDGrC-2011)
June 2011, Moscow, Russia

The proceedings are published online in the CEUR-Workshop series (ISSN 1613-0073) and the volume Vol-757 has a unique URN: urn:nbn:de:0074-757-4.

Volume Editors

Dmitry Ignatov
School of Applied Mathematics and Informatics
National Research University Higher School of Economics, Moscow, Russia

Sergei Kuznetsov
School of Applied Mathematics and Informatics
National Research University Higher School of Economics, Moscow, Russia

Jonas Poelmans
Faculty of Business and Economics
Katholieke Universiteit Leuven, Belgium

Printed in Russia by the National Research University Higher School of Economics.

Preface

Concept discovery is a Knowledge Discovery in Databases (KDD) research field that uses human-centered techniques such as Formal Concept Analysis (FCA), Biclustering, Triclustering, Conceptual Graphs etc. for gaining insight into the underlying conceptual structure of the data. Traditional machine learning techniques are mainly focusing on structured data whereas most data available resides in unstructured, often textual, form. Compared to traditional data mining techniques, human-centered instruments actively engage the domain expert in the discovery process.

This volume contains the contributions to CDUD 2011, the International Workshop on Concept Discovery in Unstructured Data (CDUD) held in Moscow. The main goal of this workshop was to provide a forum for researchers and developers of data mining instruments working on issues with analyzing unstructured data.

We are proud that we could welcome 13 valuable contributions to this volume. The majority of the accepted papers described innovative research on data discovery in unstructured texts. Authors worked on issues such as transforming unstructured into structured information by amongst others extracting keywords and opinion words from texts with Natural Language Processing methods. Multiple authors who participated in the workshop used methods from the conceptual structures field including Formal Concept Analysis and Conceptual Graphs. Applications include but are not limited to text mining police reports, sociological definitions, movie reviews, etc.

Last but not least, we would like to thank the administration of the Higher School of Economics who took care of all arrangements to make this conference pleasant and enjoyable.

June 2011, Moscow

Dmitry Ignatov
Sergei Kuznetsov
Jonas Poelmans

Organization

This CDUD'11 workshop was held in June 2011 in Moscow, Russia co-located with the 13th International Conference on Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing (RSFDGrC-2011) at the National Research University Higher School of Economics.

Program Chairs

Dmitry Ignatov	State University Higher School of Economics, Russia
Sergei Kuznetsov	State University Higher School of Economics, Russia
Jonas Poelmans	Katholieke Universiteit Leuven, Belgium

Program Committee

Guido Dedene	Katholieke Universiteit Leuven, Belgium Amsterdam Business School, The Netherlands
Paul Elzinga	Amsterdam-Amstelland Police, The Netherlands
Bernhard Ganter	Dresden University of Technology, Germany
Richard Hill	University of Derby, UK
Alex Neznanov	State University Higher School of Economics, Russia
Simon Polovina	University of Sheffield, UK
Henrik Scharfe	Aalborg University, Denmark
Vladimir Selegey	ABBYY, Russia
Stijn Viaene	Katholieke Universiteit Leuven, Belgium
Laszlo Szathmary	University of Quebec in Montreal, Canada

Sponsoring Institutions

ABBYY, Moscow
Russian Foundation for Basic Research, Moscow
Poncelet Laboratory (UMI 2615 du CNRS), Moscow
State University Higher School of Economics, Moscow
Yandex, Moscow
Witology, Moscow
Dynasty Foundation, Moscow

Table of Contents

Automatic Entity Detection Based on News Cluster Structure

Aleksey Alekseev, Natalia Loukachevitch

Research Computing Center of
Lomonosov Moscow State University, Russia

a.a.alekseevv@gmail.com, louk_nat@mail.ru

Abstract. In this paper we consider a method for extraction of alternative names of a concept or a named entity mentioned in a news cluster. The method is based on the structural organization of news clusters and exploits comparison of various contexts of words. The word contexts are used as basis for multiword expression extraction and main entity detection. At the end of cluster processing we obtain groups of near-synonyms, in which the main synonym of a group is determined.

Keywords. Entity Detection, Lexical Cohesion, News Clusters.

1 Introduction

An important step in news processing is thematic clustering of news articles describing the same event. Such news clusters are the basic units of information presentation in news services.

After a news cluster is formed, it undergoes various kinds of automatic processing:

- Duplicates are removed from the cluster. Duplicate is a message that almost completely repeats the content of an initial document,
- A cluster is categorized to a thematic category,
- A summary of a cluster is created, usually containing the sentences from different documents of the cluster (multi-document summary) etc.

The formation of a cluster can represent a serious problem. It is especially difficult to form clusters correctly for complex hierarchical events having some duration in time and distributed geographic location (world championships, elections) [1], [2].

A part of news cluster forming and processing problems is due to the fact that in cluster documents the same concepts or entities may be named differently. Lexical chain approaches could partly overcome this problem using thesaurus information [3], [4]. However in a pre-created resource, it is impossible to fix all variants for entity naming in various clusters. For example, the U.S. air base in Kyrgyzstan may be called in documents of the same news cluster as *Manas base, Manas airbase, Manas, base at Manas International Airport, U.S. base, U.S. air base* and etc.

The problem of alternative names for named entities is partly solved by coreference resolution techniques (*Russian President Dmitry Medvedev, President Medvedev, Dmitry Medvedev*) [5], [6]. In Entity Detection and Tracking Evaluations, mainly such entities as organizations, persons and locations are detected and provided with coreferential relations [7]. But main entities of a cluster can be events such as *air base closure* and *air base withdrawal.* Besides, the variability of entity names in news clusters refers not only to concrete entities but also to concepts, which can also be main discussed entities such as ecology or economic problems.

News clusters as sources of various paraphrases are studied in several works. In [8] the authors describe the procedure of corpus construction for paraphrase extraction in the terrorist domain. The study in [9] is devoted to creation of a corpus of similar sentences from news clusters as a source for further paraphrase analysis. These studies are aimed to obtain general knowledge about a domain or linguistic means of paraphrasing, but it is also important to extract near-synonyms or coreferential expressions of various types from a news cluster and to use them to improve the processing of the same news cluster or a corresponding theme.

In this paper we consider a method for extraction of main entities from a news cluster including named entities, activities and concepts. The method is based on the structural organization of news clusters and exploits comparison of various contexts of words. The word contexts are used as a basis for multiword expression extraction and main entities detection. At the end of cluster processing we obtain main entities of a news cluster and their mention expressions presented as a group of near-synonyms, in which the main synonym of a group is determined. Such synonym groups include both single words and multiword expressions. In this paper we study only simple features generated from a news cluster without attraction of additional semantic and other types of information as a basic line for future research. The experiments were carried out for Russian news flows.

2 Principles of Cluster Processing

Processing of cluster texts is based on the structure of coherent texts, which have such properties as the topical structure and cohesion.

Van Dijk [10] describes the topical structure of a text, the macrostructure, as a hierarchical structure in a sense that the theme of a whole text can be identified and summed up to a single proposition. The theme of the whole text can usually be described in terms of less general themes, which in turn can be characterized in terms of even more specific themes. Every sentence of a text corresponds to a subtheme of the text.

The macrostructure of a connected text defines its global coherence: “Without such a global coherence, there would be no overall control upon the local connections and continuations” [10]. Sentences must be connected appropriately according to the given local coherence criteria, but the sequence would go simply astray without some constraint on what it should be about globally.

Cohesion, that is surface connectivity between text sentences, is often expressed through anaphoric references (i.e. pronouns) or by means of lexical or semantic repetitions. Lexical cohesion is modeled on the basis of lexical chains [11].

The proposition of the main theme, that is an interaction between theme participants, should be represented in specific text sentences, which should refine and elaborate the main theme. This means that if a text is devoted to description of relations between thematic elements $C_1 \ldots C_n$, then references to these participants should be met in different roles to the same verb in text sentences.

Thus if even very semantically close entities C_1 and C_2 often co-occur in the same sentences of a text, it means that the text is devoted to consideration of relations between these entities and they represent different elements of the text theme [12], [13]. At the same time, if two lexical expressions C_1 and C_2 are rarely met in the same sentences but occur very frequently in neighbor sentences then we can suppose that they are elements of lexical cohesion, and there is a semantic relation between them.

A news cluster is not a coherent text but cluster documents are devoted to the same theme. Therefore statistical features of the topical structure are considerably enhanced in a thematic cluster, and on such a basis we try to extract unknown information from a cluster.

To check our idea that near-synonyms can be more often met in neighbour sentences than in the same sentences we have carried out the following experiment. More than 20 large news clusters have been matched with terms of Sociopolitical thesaurus [14] and thesaurus-based potential near-synonyms have been detected. Such types of near-synonyms include (these examples are translations from Russian, in Russian the ambiguity of expressions is absent):

- nouns – thesaurus synonyms (*Kyrgyzstan – Kirghizia*),
- adjective – noun derivates (*Kyrgyzstan – Kyrgyz*),
- hypernym and hyponym nouns(*deputy – representative*),
- hypernym–hyponym noun - adjective (*national – Russia*),
- part-whole relations between nouns (*parliament – parliamentarian*),
- part-whole relations for adjective and noun (*American – Washington*),

For each cluster we considered all these pairs of expressions with a frequency filter: the frequencies of the expressions in a cluster should be more than a quarter of the number of documents in the corresponding cluster. For these pairs we computed the ratio between their co-occurrence in the same sentence clauses F_{segm} and in neighbour sentences F_{sent}. Table 1 shows the results of our experiment.

Table 1. Frequency ratio of related expressions within segments of sentences and neighbour sentences

Type of relation	F_{segm}/F_{sent} ratio	Number of pairs
Synonymic Nouns	0.309	31
Noun-adjective derivation	0.491	53
Hyponym – Hypernym (nouns)	1.130	88
Hyponym – Hypernym (noun – adjective)	1.471	28
Meronym- holonym (nouns)	0.779	58
Meronym- holonym (noun – adjectives)	1.580	29
Other	1.440	21483

From the table we can see that the most closely-related expressions (synonyms, derivates) are much more frequent in neighbour sentences than in the same clauses of the same sentences. Further, the more the distance in a sense between expressions is the more the ratio F_{segm}/F_{sent} is until stabilization near the value equal 1.5.

We can also see that noun-noun and noun-adjective pairs have different values of the ratio. We suppose that in many cases adjectives are elements of noun groups, which can play own roles in a news cluster. Therefore the first step in detection of main entities should be extraction of multiword expressions denoting main entities of the cluster.

3 Stages of Cluster Processing

Cluster processing consists of three main stages. At the first stage noun and adjective contexts are accumulated. The second stage is devoted to multiword expression recognition. At the third stage the search of near-synonyms is performed.

In next sections we consider processing stages in more detail. As an example we use the news cluster, which is devoted to Kyrgyzstan and the United States agreement denunciation on U.S. air base located at the Manas International Airport (19.02.2009). This news cluster contains 195 news documents and is assembled on the basis of the algorithm described in [1].

3.1 Extraction of Word Contexts

Sentences are divided into segments between punctuation marks. Contexts of word W include nouns and adjectives situated in the same sentence segments as W. The following types of contexts are extracted:

– Neighboring words: neighboring adjectives or nouns situated directly to the right or left from W (*Near*),

- Across verb words: adjectives and nouns occurring in sentence segments with a verb, and the verb is located between W and these adjectives or nouns (*AcrossVerb*),
- Not near words: adjectives and nouns that are not separated with a verb from W and are not direct neighbors to W (*NotNear*).

In addition, adjective and noun words that co-occur in neighboring sentences are memorized (Ns). For this context extraction only sentence fragments from the beginning up to a segment with a verb are taken into consideration. It allows us to extract the most significant words from neighboring sentences.

3.2 Extraction of Multiword Expressions

We consider recognition of multiword expressions as a necessary step before near-synonym extraction. An important basis for multiword expression recognition is the frequency of word sequences [15]. However, a news cluster is a structure where various word sequences are repeated a lot of times. We supposed that the main criterion for multiword expression extraction from clusters is the significant excess in co-occurrence frequency of neighbor words in comparison with their separate occurrence frequency in segments of sentences (1):

$$\text{Near} > 2 * (\text{AcrossVerb} + \text{NotNear}) \quad (1)$$

In addition, the restrictions on frequencies of potential component words are imposed.

Search for candidate pairs is performed in order of the value "*Near - (AcrossVerb + NotNear)*" reducing. If a suitable pair has been found, its component words are joined together into a single object and all contextual relationships are recalculated. The procedure starts again and repeats until at least one join is performed.

As a result, such expressions as *Parliament of Kyrgyzstan, the U.S. military, denunciation of agreement with the U.S., Kyrgyz President Kurmanbek Bakiyev* were extracted from the example cluster.

3.3 Detection of Near-Synonyms

At the third stage, search for near-synonyms is produced. For assuming a semantic relationship between expressions U_1 and U_2, the following factors are exploited:

- U_1 and U_2 have formal resemblance (for example, words with the same beginning),
- U_1 and U_2 co-occur more often in neighboring sentences than within segments of the same sentences; here we use results of the experiment described in section 2;
- U_1 and U_2 have similar contexts based on Near, AcrossVerb, NotNear and Ns features, which are determined by calculating scalar products of corresponding vectors (NearScalProd, AVerbScalProd, NotNearScalProd, NsentScalProd),
- U_1 and U_2 should be enough frequent in a cluster to present main entities.

Note that if the comparison of word contexts is a well known procedure for synonym detection and taxonomy construction [16], but the generation of contexts from neighboring sentences has not been described in the literature.

Near-synonyms detection consists of several steps. A different set of criteria is applied at each step. The lookup is performed in order of frequency decreasing: for every expression U_1, all expressions U_2 having a lower frequency than U_1, are considered. If all conditions are satisfied, then less frequent expression U_2 is postulated as a synonym of U_1 expression, all U_2 contexts are transferred to U_1 contexts, the expressions U_1 and U_2 become joined together. As a result the sets of near-synonyms (synonym groups) are produced, i.e. linguistic expressions that are equivalent with respect to the content of the cluster.

We assume that U_1 and U_2 expressions, when they are enclosed in such a synonym group, are closely related in sense, or their referents in current cluster are closely related to each other, so that U_2 does not represent separate thematic significance with respect to U_1. For example, such words as *parliament* and *parliamentarian* have a close semantic relationship between them in general context, but they are not synonyms. But within a particular cluster, e.g., in which decision-making process in a parliament is discussed, these words may be classified as near-synonyms.

At the first step (3.1) semantic similarity between expressions consisting of similar words is sought, e.g. *Kyrgyzstan - Kyrgyz, Parliament of Kyrgyzstan - Kyrgyz Parliament*. We used simple similarity measure – the same beginning of words.

To connect words with the same beginning in synonym groups, the following conditions are required: the co-occurrence frequency in neighboring sentences is significantly higher than co-occurrence frequency in the same sentences (2, 3) (see section 2); both expressions should have sufficient frequencies in the cluster. The procedure is iterative:

$$Ns > 2 * (AcrossVerb + Near + NotNear) \quad (2)$$

$$Ns > 1 \quad (3)$$

If expressions are rarely located in neighboring sentences ($Ns < 2$), then the scalar product similarity of contexts is required:

$$NearScalProd + NotNearScalProd + AVerbScalProd + NSentScalProd > 0.4 \quad (4)$$

At the second step (3.2) semantic similarity between expressions, one of which is included into another, is sought, for instance, *Parliament - Parliament of Kyrgyzstan, airbase - Manas airbase*. The meaning of this step lies in the fact that a cluster might not mention any other parliaments, except of the *Kyrgyz Parliament*, i.e. in both cases the same object is mentioned. Similarity of neighbor contexts is required here:

$$NearScalProd > 0.1 \quad (5)$$

At the third step (3.3) we are looking for semantic similarity between the expressions with equal length and including at least one the same word, for example, *Manas Base*

- *Manas Airbase, the U.S. military - the U.S. side*. High frequency of co-occurrence in neighboring sentences is required (6, 7):

$$NS > 2 * (AcrossVerb + Near + NotNear) \quad (6)$$

$$NS > 1 \quad (7)$$

Finally, at last step (3.4) semantic similarity between arbitrary linguistic expressions, mentioned in cluster documents, is searched, e.g. *USA - American, Kyrgyzstan - Bishkek*. An assumption on semantic similarity between arbitrary expressions requires the maximum number of conditions: high frequency of co-occurrence in neighboring sentences (8, 9); restrictions on occurrence frequencies of candidates, context similarity:

$$NS > 2 * (AcrossVerb + Near + NotNear) \quad (8)$$

$$NS > 0.1 * MaxAcrossVerb \quad (9)$$

The following synonym groups were automatically assembled for the example cluster as a result of described stages (the main synonym of a group, which was automatically determined, is highlighted with bold font):

- ***Manas base:*** *base, Manas Air Base, Air Base, Manas;*
- ***USA:*** *American, America;*
- ***Kyrgyzstan:*** *Kirghizia, Kyrgyz, Kyrgyz-American, Bishkek;*
- ***Parliament of Kyrgyzstan:*** *Kyrgyz parliament, parliament, parliamentary, parliamentarian;*
- ***Manas International Airport:*** *airport, Manas airport;*
- ***Bill:*** *law, legislation, legislative, legal* and etc.

4 Evaluation of Method

To test the introduced method we took 10 news clusters on various topics with more than 40 documents in each cluster.

Two measures of quality were tested for multiword expression extraction. Firstly, we evaluated the percentage of syntactically correct groups among all extracted expressions. Secondly, we have attracted a professional linguist and asked her to select the most significant multiword expressions (5-10) for each cluster, and to arrange them in descending order of importance.

So for the example cluster, the following expressions were considered significant by the linguist:

- *Manas Airbase*
- *Parliament of Kyrgyzstan*
- *Manas base*
- *Kyrgyz Parliament*
- *Denunciation of agreement*

– *Government's decision*

Note that such an evaluation task differs from evaluation of automatic keyword extraction from texts [17], when experts are asked to identify the most important thematic words and phrases of a text. In our case we tested exactly multiword expression extraction. In addition, a list created by the linguist could contain semantic repetitions (*Parliament of Kyrgyzstan - Kyrgyz Parliament*).

364 multiword expressions were automatically extracted from test clusters, 312 (87.9%) of which were correct syntactic groups. With account of phrase frequencies, correct syntactic expressions achieved 91.4% precision. The linguist chose 70 most important multiword expressions for clusters and 72.6% of them were automatically extracted by the system.

We tested extracted synonym groups by evaluating semantic relatedness of every synonym in a group to its main synonym. Every occurrence of supposed synonyms was tested. If more than a half of all occurrences of such a synonym in a cluster were related to the main synonym in the group, the synonymic relation was considered as correct.

Table 2 contains information about the quality of generated synonym groups calculated in number of expressions and in their frequencies.

Table 2. Test results for automatic detection of synonym groups in news clusters

Step	Number of joins	Total join frequency	Percent of correct joins	Percent of correct joins by frequency
3.1. The same beginning expressions	155	4383	87.9%	91.4%
3.2. Embedded expressions	99	9131	91.4%	92.9%
3.3. Intersecting expressions	8	677	85.7%	80.8%
3.4. Arbitrary expressions	38	4822	62.5%	62.4%

To assess the contribution of co-occurrence in neighboring sentences, we conducted detailed testing of the same beginning expression joining (step 3.1) for the example cluster (Table 3). Table 3 shows that Ns factor adding, as it is done in step 3.1, improves precision and recall of near-synonym recognition. The proposed method has not the absolutely best F-measure value, but the precision less than 80% is inadmissible for the near-synonym detection task. Therefore, the BasicLine should not be considered as the best approach.

Table 3. Test results for different methods of detection of near-synonyms with the same beginning

Method	Number of joined expressions	Total joining frequency	Correct joining frequency	Precision by frequency (%)	Recall by frequency (%)	F-measure (%)
Expressions with the same beginning (BasicLine)	383	2266	1472	65%	100%	78.8%
Expressions with the same beginning + scalar products (threshold 0.1)	38	996	834	83.7%	56.7%	67.6%
Expressions with the same beginning + scalar products (threshold 0.4)	36	976	814	83.4%	55.3%	66.5%
Step 3.1 conditions	36	965	873	**90.5%**	**59.3%**	**71.7%**

5 Conclusion

In this paper we have described two experiments on news clusters: multiword expression extraction and detection of near-synonyms presenting the same main entity of a news cluster. In addition to known methods of context comparison, we exploited co-occurrence frequency in neighboring sentences for near-synonym detection. We conducted the testing procedure for the introduced method.

In future we are going to use extracted near-synonyms in such operations as cluster boundaries correction, automatic summarization, novelty detection, formation of sub-clusters and etc. We also intend to study methods of combination automatically extracted near-synonyms, methods of coreference resolution and thesaurus relations.

6 References

1. Dobrov, B., Pavlov, A.: Basic line for news clusterization methods evaluation. In: Proceedings of the 5-th Russian Conference RCDL-2010 (2010) (in Russian)
2. Allan, J.: Introduction to Topic Detection and Tracking. In: Topic detection and tracking, Kluwer Academic Publishers Norwell, MA, USA,. pp. 1-16 (2002)

3. Li, J., Sun, L., Kit, C., Webster, J.: A Query-Focused Multi-Document Summarizer Based on Lexical Chains. In: Proceedings of the Document Understanding Conference DUC-2007 (2007)
4. Dobrov, B., Loukachevitch, N.: Summarization of News Clusters Based on Thematic Representation. In: Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference Dialog'2009, pp. 299-305 (2009) (In Russian)
5. Duame, H., Marcu, D.: A large Scale Exploration of Global Features for a Joint Entity Detection and Tracking Model. In: Proceedings of Human Language Conference and Conference on Empirical Methods in Natural Language Processing, pp. 97-104 (2005)
6. Ng, V.: Machine learning for coreference resolution: from local classification to global ranking. In: Proceedings of ACL-2005 (2005)
7. Doddington, G., Mitchell, A., Przybocki, M., Ramshaw, L., Strassel, S., Weishedel, R.: The Automatic Content Extraction (ACE): Task, Data, Evaluation. In: Proceedings of Fourth International Conference on Language Resources and Evaluation, LREC 2004 (2004)
8. Barzilay, R., Lee, L.: Learning to Paraphrase: an Unsupervised Approach Using Multiple Sequence Alignment. In: Proceedings of HLT/NACCL-2003 (2003)
9. Dolan, B., Quirk, Ch., Brockett, Ch.: Unsupervised Construction of Large Paraphrase Corpora: Exploiting Massively Parallel News Sources. In: Proceedings of COLING-2004 (2004)
10. Dijk, van T.: Semantic Discourse Analysis. In: Teun A. van Dijk, (Ed.), Handbook of Discourse Analysis, vol. 2., pp. 103-136, London: Academic Press (1985)
11. Hirst, G., St-Onge, D.: Lexical Chains as representation of context for the detection and correction malapropisms. In: WordNet: An electronic lexical database and some of its applications / C. Fellbaum, editor. Cambrige, MA: The MIT Press (1998)
12. Hasan, R.: Coherence and Cohesive harmony. J. Flood, Understanding reading comprehension, Newark, DE: IRA, pp. 181-219 (1984)
13. Loukachevitch, N.: Multigraph representation for lexical chaining. In: Proceedings of SENSE workshop, pp. 67-76 (2009)
14. Loukachevitch, N., Dobrov, B.: Evaluation of Thesaurus on Sociopolitical Life as Information Retrieval Tool. In: M.Gonzalez Rodriguez, C. Paz Suarez Araujo (Eds.), Proceedings of Third International Conference on Language Resources and Evaluation (LREC2002), Vol.1, pp.115-121 (2002)
15. Witten, I., Paynter, G., Frank, E., Gutwin, C., Newill-Manning, C.: KEA: practical automatic keyphrase extraction. In: Proceedings of the fourth ACM conference on Digital Libraries (1999)
16. Yang, H., Callan, J.: A metric-based framework for automatic taxonomy induction. In: Proceedings of ACL-2009 (2009)
17. Su Nam Kim, Medelyan, O., Min-Yen Kan, Baldwin, T.: Automatic Keyphrase Extraction from Scientific Articles. In: Proceedings of the 5-th International Workshop on Semantic Evaluation, ACL -2010, pp. 21-26 (2010)

Application of Conceptual Structures in Requirements Modeling

Michael Bogatyrev, Vadim Nuriahmetov

Tula State University,
Lenin ave. 92, 300600 Tula, Russia

okkambo@mail.ru, vadim-nuriahmetov@yandex.ru

Abstract. Requirements modeling has been applied in CASE technologies to formalize knowledge needed for constructing models of information systems. The problem is to acquire knowledge from requirements texts and represent it as intermediate requirements model for entity-relationships or object oriented modeling. Proposed approach is based on formalization of entities and their attributes as formal contexts. It is shown that formal contexts created on the set of conceptual graphs extracted from requirements text may serve as data source for requirements models have been applied in real CASE technologies.

Keywords: CASE technology, requirements modeling, conceptual graphs, conceptual structures, conceptual requirements model, Sybase PowerDesigner.

1 Introduction

In one of early works of John Sowa [1] conceptual graphs were discovered as intermediate models between natural language and database interfaces. Following this idea in this paper conceptual graphs are used as intermediate model between natural language and *requirements models* which have been applied in database CASE technologies.

Requirements Modeling has been applied in *Requirements Engineering* [3] to formalize a knowledge needed for constructing models of information systems in CASE technologies. Modern CASE technologies, for example technology of Sybase PowerDesigner [5], realize Requirements Modeling as a real working tool. Here a text of requirements of a project is a data source which contents (words or phrases) beget requirements. Every requirement is an object in requirements model. It has a name and attributes - *type, status, priority, risk*, etc. In the requirements model every requirement is connected with elements of other CASE-models, for example with elements of Entity-Relationship Diagrams (ERD) or UML diagrams. Connection means that when a CASE-model is processed it must be done by meeting demands of requirements. The instrument of Requirements Modeling is actual in big projects with complex textual requirements. It is also important in supporting life cycle of the system to be designed [4].

A challenging problem in Requirements Modeling is the problem of creating requirements model from natural language text of requirements.
Significant numbers of works in the area of Requirements Modeling have been devoted to this problem. The most of them treat it as direct mapping text to CASE-models and a requirement considered as a text. All such works can be divided into two sets: one set of works is devoted to derive a family of Entity Relationship models (plain or extended ERD) from natural language texts ([6] - [8]); another set of works is about object oriented models represented by class diagrams ([9], [12]). These works are based on the assumption that meaning of concepts being extracted from a text can be derived from grammar structures of natural language. Heuristic rules of implementing properties of parts-of-speech and their functions in sentences are applied here. Besides English language, decisions for some other languages including German [8] and Japanese [9] have been presented. Modeling by analyzing contexts in requirements texts is presented in [10]. Some examples of real requirements modeling systems are presented in [16].
In spite of many existing results here including ones oriented on grammars of concrete languages, full automation of CASE-models design from requirements texts is fundamentally impossible. The text of requirements actually contains more or less portion of information needed for creating a CASE-model and textual data could not be mapped exactly to the data of CASE-model.
Therefore the central Requirements Modeling problem needs to be formulated in its natural form – as a problem of creating requirements model from natural language text. This requirements model has to be treated as separate intermediate model between requirements texts and CASE-models.
This paper is based namely on that approach to Requirements Modeling. It is shown that formal contexts created on the set of conceptual graphs extracted from requirements text may serve as data source for requirements models have been applied in real CASE technologies.

2. Conceptual Requirements Modeling

The term *Conceptual Requirements Modeling* is appropriate to denote the fact of applying *Conceptual Structures* in *Requirements Modeling*. Domain of Conceptual Structures combines *conceptual graphs* [2] and Formal Concept Analysis [13] techniques and now can be considered as general approach for modeling many problems in Data Mining and Text Mining areas.

2.1 Conceptual Structures as Requirements Model

Both Entity Relationships and Object Oriented CASE-models use *objects* and *attributes*. Attributes belong to *entities* in ERD and to *objects* in Object Oriented models (OOM).

An entity is an object from real world having finite set of attributes. Entity name denotes this set of attributes, for instance, $Student\{Name, Date\_Birth, ...\}$.

One of the crucial principles of Entity Relationship modeling claims that *every entity has only generic attributes* i.e. attributes which characterize only entity itself. Describe this by the following way. Consider the set of data types $T = \{D_1, D_2, ..., D_n\}$ consisted of domains D_i. Every domain is ordered set of data of certain type, for example *character*, *numeric*, *date*, etc. The set of attributes $A = \{a_i\}, a_i \in D_j$ is multiset, which contains examples of domain elements. We denote every i - th example of entity as $e_i = \{A_i\}, A_i \subset A$. All examples e_i constitute an *entity type* $Ee_j = \{\}_i$, which attributes are the same for all examples of entities. Then the *generic feature* of attributes is described by condition:

$$for\ E_i \neq E_j\ \ A_i \bigcap A_j = \varnothing \qquad (1)$$

Very often this demand is not met in practice. For example: $E_1 = Student\{Name, Date\_Birth, ...\}$ and $E_2 = Teacher\{Name, Date\_Birth, ...\}$ have similar subsets of attributes. To hold the condition (1) in CASE-technology one must rename attributes of entity in the example above. Note that the mapping sets of attributes to entities (or vice versa) could not be rigorously formalized as a function - it is *a relation*. An appropriate way of expressing it is formal context [13].

Consider a formal context (*E*, *A*, *R*), where $E = \bigcup_i E_i$ and *R* is a relation which establishes the facts of belonging attributes to entities. Formal context (*E*, *A*, *R*) may be represented by [0, 1] - matrix in which units mark correspondence between entities *E* and attributes *A*. If the set *A* is ordered by its subsets $A = \{A_1, A_2, ..., A_k\}$ and the condition (1) is hold then the context matrix has block-diagonal structure

$$\mathbf{C} = diag[\mathbf{C}_1, \mathbf{C}_2, ..., \mathbf{C}_k] \qquad (2)$$

as it is shown on Figure 1. Every sub matrix $\mathbf{C}_j$ represents a relation on subsets of entities where entities are grouped into *associated entities* which are associated by closed subsets of attributes. An example of associated entity $\tilde{E}_1 = Human$ {*Student*, *Teacher*, *Dean*} is shown on Figure 1.

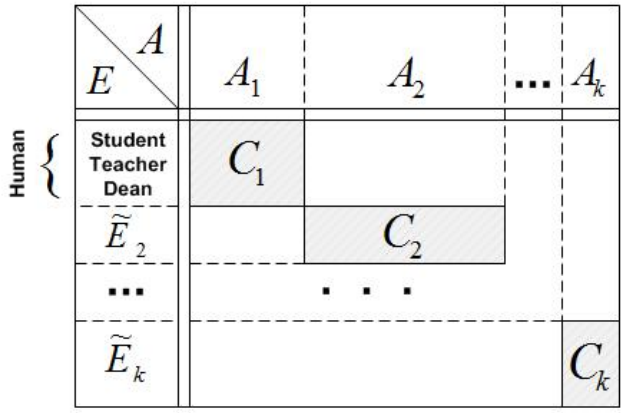


Fig. 1. Formal context for associated entities.

Every associated entity $\tilde{E}_i$ unites maximum number of entity attributes and not all of them belong to all entities inside an association, so association context matrices $\mathbf{C}_i$ may be sparse. Their attribute subsets A_i may have some subsets of attributes $\tilde{A}_i \subseteq A_i$ which belong to all entities in association. If we construct another context $(\tilde{E}, \tilde{A}, R)$ regard to associated entities and attributes $\tilde{A} = \{\tilde{A}_i\}$ then this context sub matrices $\tilde{\mathbf{C}}_i$ will be completely filled by units

Sub matrices $\tilde{\mathbf{C}}_i$ may represent formal concepts [13] on the context $(\tilde{E}, \tilde{A}, R)$. Formal concept on the context $(\tilde{E}, \tilde{A}, R)$ is a pair of subsets $X \subseteq \tilde{E}, Y \subseteq \tilde{A}$ together with pair of mappings $\varphi : \tilde{E} \to \tilde{A}, \psi : \tilde{A} \to \tilde{E}$ realizing so called *Galois connection* [14]. A pair (φ, ψ) is a Galois connection between the partially ordered sets (posets) $(\tilde{E}, \sqsubseteq), (\tilde{A}, \Subset)$ if the following conditions hold: for all

$$x \in X,\ y \in Y\ x \sqsubseteq \psi(\varphi(x)), \varphi(\psi(y)) \Subset y\ . \qquad (3)$$

Galois connection is that type of mapping which "synchronously" conserves sets orders or maps sets orders from one poset to another. The set of formal concepts on a context forms a conceptual lattice [13].

Considered conceptual structures – formal context and formal concepts – may serve as an instrument for constructing requirements models. They unite objects and attributes by relations and have important property of completeness: as formal context as formal concepts are complete objects with certain informational content extracted

from the text of requirements. Apparently this content must be represented in CASE-models so the context and formal concepts constitute a kind of requirements. They may be considered as *Conceptual Requirements Model*.

2.2 Conceptual graphs acquisition and processing.

To extract objects and their attributes from requirements text, the approaches mentioned in the Introduction section may be applied. Conceptual graphs are appropriate for it due to the following reasons:

- if successfully acquired from text, conceptual graphs represent compact model for discovering objects and their attributes - there may be a set of conceptual relations in a graph which depict connection between objects and their attributes;
- conceptual graphs naturally belong to Formal Concept Analysis paradigm and have been successfully applied for constructing formal contexts [18].

Using conceptual graphs, another problem becomes actual – the problem of acquisition conceptual graphs from texts.

We use our software for conceptual graphs acquisition from natural language texts [15]. The software is based on existing approaches of lexical, morphological and semantic analysis. *Semantic roles labeling* [19] is applied as the main instrument for constructing relations in acquisition algorithm. The algorithm works with our recently developed *controllable grammatical templates*. Using these templates, it is possible to adapt acquisition algorithm as to certain language grammar (Russian grammar in the current version of the system) as to some peculiarities of concrete language. User interface has also tools for recognizing incorrect conceptual graphs. Incorrect conceptual graph is a graph having *isolated concepts* i.e. concepts which are not connected to other concepts by relations.

Conceptual graphs are acquired from subtitles of requirements text and from text sections. It is interesting to find similarities between graphs acquired from subtitles and graphs acquired from text sections since some terms (objects) declared in a subtitle may be mentioned and concretized in a section text. We apply measures of similarity of conceptual graphs which we used in our experiments of conceptual graphs clustering [20].

All acquired correct conceptual graphs are processed to extract objects and their attributes. The way of extraction is based on fixing certain set of conceptual relations presented in derived graph. There are trivial and non trivial patterns of concepts and corresponding relations which may exist in a graph. If standardized text of requirements is a source for graph acquisition then it is possible to create special templates for graph acquisition algorithm.

2.3 Creating and processing formal contexts

Conceptual graph represents semantics of only one sentence. Important information about objects and their attributes may be presented in several various sentences. To

collect it we use formal context. A context with associated entities having block-diagonal structure (2) contains the needed information.

Formal context is created as [0, 1] matrix in which correspondence between objects and their attributes is supported. That correspondence is established after processing conceptual graphs and it is not enough to say that condition (1) is true on the context's structure created automatically on the acquired sets of objects and attributes. So we apply block-diagonal decomposition of context matrix to find its structures similar to shown on Figure 1. Any algorithm of block-diagonal matrix decomposition works so that it is equivalent to some permutations of rows and columns of matrix. As a result initial correspondence between objects and attributes may be disrupted. The sets of objects and attributes in a context are partially ordered so only those permutations which conserve this feature are allowed.

3. Conceptual Requirements Modeling System Realization

The approach we propose brings additional functionality to those CASE technologies which work with conceptual and object oriented models. The Sybase PowerDesigner CASE system [5] is one of the few systems where Requirements Modeling is really exploring.

Sybase PowerDesigner CASE technology supports requirements modeling with natural language texts as its input. Figure 2 illustrates the principle of requirements modeling in PowerDesigner [5].

As it is shown on the Figure 2 PowerDesigner processes formatted MS Word textual documents. Requirements model on the Figure 2 is consisted of two elements: requirements which are title and headings of sections and subsections of the document and *traceability matrices* which represent various connections between requirements and between requirements and elements of created CASE-models. Title and headings are treated as elements of requirements model. The text between two headings is treated as the requirements object's comment.

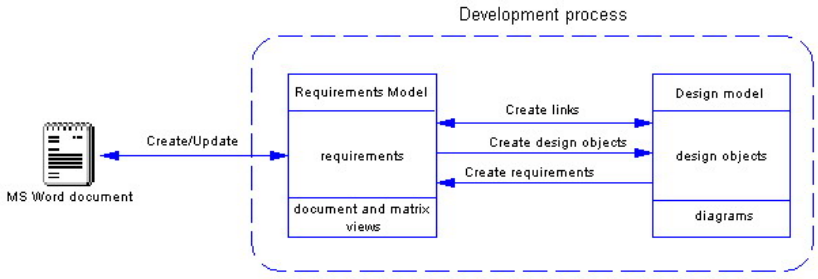


Fig. 2. Principle of requirements modeling in Sybase PowerDesigner

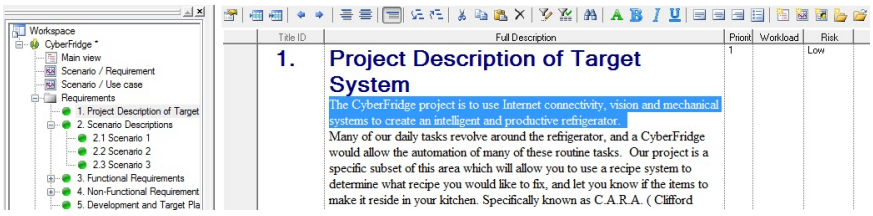
Detecting requirements as title and headings of sections and subsections of a text is the only current function of natural language processing in PowerDesigner. Having such requirements, user manually applies them in constructing CASE-models by

setting entities and relationships, objects and their attributes: type, status, priority, risk. Using traceability matrices user creates tools for checking connections between various objects of models.

Additional functionality to this technology is caused by the fact that requirements text between two headings has been also processed to extract objects and attributes united by formal context. It takes place by the following.

1. Conceptual graphs are acquired from the whole text of requirements. Fixed set of conceptual relations (for example *genitive* and *attribute* relations) in conceptual graphs is applied to select candidate pairs of *objects - attributes* to form a context.
2. Context matrix is formed so that only objects with more than one attribute have been included in the matrix. This is the way to select objects significant for constructing CASE-models.
3. Initial sparse context matrix is transformed by using linear algebra methods for block-diagonal decomposition. The problem of keeping order in the sets which form context is actual here. Associated entities may be ordered by hierarchy relation, for example as *Student*, *Teacher*, *Dean* on Figure 1. As usual attributes are less ordered and can be permuted for block-diagonal decomposition.
4. Interaction with user by special interface and visualization is very important since the process of creating CASE-models still remains closely depended on developer's skill. In the current experimental version of the system there is user interface to show every subset of object and its attributes obtained from conceptual graphs. User can correct this set.

Figure 3 illustrates this technology on the fragment of *CyberFridge* project included in PowerDesigner as an example of requirements modelling. On the figure we combined two windows: on the top of the figure there is PowerDesigner interface window showing how requirements are represented, below there is interface window of our system visualizing conceptual graph corresponded to the first sentence highlighted in the top window. Only *attribute* relation was processed here and candidate pairs of *objects – attributes* are shown. Later, analyzing other graphs and the context created on the whole set of candidate pairs we keep only *refrigerator* and *cyberfridge* entities as requirements.



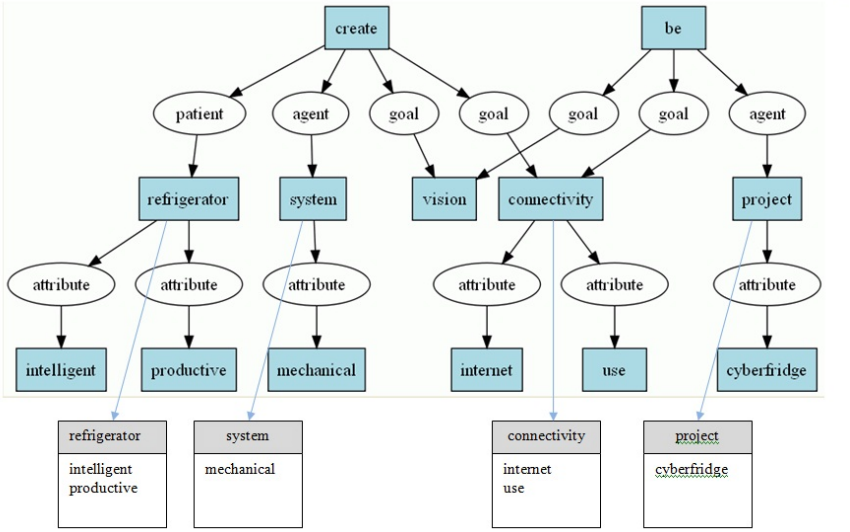


Fig. 3. The example of visualization of conceptual graph and *objects – attributes* pairs of the sentence of requirements text: *"The CyberFridge project is to use Internet connectivity, vision and mechanical systems to create an intelligent and productive refrigerator"*.

.

On the standard way of creating requirements model in PowerDesigner user takes only headlines of requirements text ("*Project Description of Target System*" on Figure 3) as requirements objects and treats remaining text as comments. Conceptual Requirements Modeling System extends functionality of requirements modeling in PowerDesigner realizing more complete text processing.

4. Preliminary Results and Future Work

The first version of the system of Conceptual Requirements Modeling was tested on various Russian texts of requirements being structured according to the standard [17]. We also started to process English texts as it is shown on Figure 3.

First results obtained from experiments demonstrate the following.

1. Conceptual graphs are valid for extracting objects and attributes from natural language texts of requirements and can deliver specific new information for CASE-models developer.

2. Formal context serves as a tool for collecting entities and their attributes for Entity Relationship and Object Oriented Modeling and selects objects significant for constructing CASE-models.

The way of developing proposed approach is mostly experimental and its final effectiveness can be confirmed after series of additional experiments and corresponded changes in the algorithm.

Future work is planned in the following directions.

1. Extending the set of relations which is applied to select candidate pairs of *objects - attributes* to form a context. For example the *goal* relation on Figure 3 is also informative as *attribute* relation.
2. Discovering the way of implementing formal concepts on formal context in the approach. Specifically, if formal concepts exist on the context do they form additional objects significant for CASE modeling?
3. Deep integration proposed approach with CASE technology, particularly with Sybase PowerDesigner.

We also plan to expand the set of our controllable grammatical templates by including English language grammar to it.

References

1. Sowa, J.F.: Conceptual Graphs for a Data Base Interface. IBM Journal of Research and Development 20(4): 336-357 (1976).
2. Sowa, J.F., Knowledge Representation: Logical, Philosophical, and Computational Foundations, Brooks Cole Publishing Co., Pacific Grove, CA, (2000).
3. Young, R.: The Requirements Engineering Handbook. Artech House Publishers (2004)
4. Blanchard B. S., Fabrycky, W. J.: Systems Engineering and Analysis, Fourth Edition. Prentice Hall. (2006)
5. PowerDesigner 15.2 Requirements Modeling. Sybase documentation, DC 00121-01-1520-01. February 2010
6. Chen, P.: English Sentence Structure and Entity-Relationship Diagrams. Information Science, 29(2-3): 127–149 (1983)
7. Hartmann, S., Link, S.: English Sentence Structures and EER Modeling. In Proc. of 4th Asia-Pacific Conference on Conceptual Modelling (APCCM2007), 27–35 (2007)
8. Tjoa, A. M., Berger, L.: Transformation of Requirement Specifications Expressed in Natural Language into an EER Model. LNCS, vol. 823, 206-217. Springer-Verlag, Berlin, Heidelberg (1994)
9. Hasegawa, R., Kitamura, M., Kaiya, H. and Saeki, M.: Extracting Conceptual Graphs from Japanese Documents for Software Requirements Modeling. In Proc. Sixth Asia-Pacific Conference on Conceptual Modelling (APCCM 2009), Wellington, New Zealand. CRPIT. 87-96.(2009)
10. Lee, S., Kim, N., Moon, S.: Context-Adaptive Approach for Automated Entity Relationship Modeling. Journal of Information Science and Engineering v. 26, 2229-2247 (2010)
11. Saeki, M., Horai, H., Enomoto, H.. Software Development Process from Natural Language Specification. In Proc. of 11th International Conference on Software Engineering, 64–73. (1989)
12. Overmyer, S. Lavoie, B., Rambow, O.: Conceptual Modeling through Linguistic Analysis Using LIDA. In Proc. of 23rd International Conference on Software Engineering (ICSE'01), 401–410. (2001)

13. Ganter, B., Stumme, G., Wille, R.: Formal Concept Analysis, Foundations and Applications. LNCS. 3626, Springer. (2005)
14. Birkhoff, G.: Lattice Theory. Providence, RI: Amer. Math. Soc. (1967)
15. Bogatyrev, M. .Y, Mitrofanova, O. A., Tuhtin, V. V.: Building Conceptual Graphs for Articles Abstracts in Digital Libraries. - Proceedings of the Conceptual Structures Tool Interoperability Workshop (CS-TIW 2009) at 17th International Conference on Conceptual Structures (ICCS'09), 50-57. Moscow. (2009)
16. Omar, N., Paul, H., Paul, M. K.: Heuristics-Based Entity-Relationship Modelling through Natural Language Processing, Proceedings of the Fifteenth Irish Conference on Artificial Intelligence and Cognitive Science, Galway-Mayo Institute of Technology (GMIT), 302-313. Castlebar, Ireland. (2004)
17. Information technology. Set of standards for automated systems. Technical directions for automated system making. (GOST 34.602-89). http://www.vniiki.ru/document/4144866.aspx
18. 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)
19. Gildea D., Jurafsky D.: Automatic labeling of semantic roles. Computational Linguistics, 2002, v. 28, 245-288. (2002)
20. Bogatyrev, M. Y., Terekhov, A. P.: Framework for Evolutionary Modeling in Text Mining. - Proceedings of the SENSE'09 - Conceptual Structures for Extracting Natural language Semantics. Workshop at 17th International Conference on Conceptual Structures (ICCS'09), 26-37 (2009)

Abstracting Concepts from Text Documents by Using an Ontology

E.Chernyak[1], O.Chugunova[1], J.Askarova[1], S. Nascimento[2], B.Mirkin[1,3]

[1]Division of Applied Mathematics and Informatics, National Research University Higher School of Economics, Moscow, Russian Federation
[2]Department of Informatics, New University of Lisbon, Caparica, Portugal
[3]Department of Computer Science, Birkbeck University of London, London, UK

Abstract. A method for computationally visualizing and interpreting a text or corpus of texts in a taxonomy of the field is described. The method involves such stages as matching taxonomy topics and text(s) by using annotated suffix trees (ASTs), combining multiple information such as text abstracts, key-words and taxonomy cross-references, building clusters of taxonomy topics and their profiles, and lifting the profiles to higher ranks of the taxonomy hierarchy.

1 Introduction

The concept of ontology as a computational device for handling domain knowledge is one of the points of growing interest in machine intelligence. Initially main efforts of the researchers concentrated on building ontologies; currently the research interests are shifting towards the usage of ontologies (see, for example, [5], [6], [10], [4]). This paper is aimed at the latter perspective. Our ultimate goal is to devise a system that would allow the user to use a domain ontology for computational interpretation of a text or a set of texts from this field. The paper presents some initial stages of our work on the long path towards achieving the goal. These stages include the following: (a) selection of the conceptual hierarchy (taxonomy) as a formalization of the concept of ontology, (b) representation of both texts and taxonomy topics in a unified framework that facilitates and channels the sifting of the taxonomy topics through the texts to score matches between them in a comprehensive way, (c) developing quantitative profiles of the texts, (d) clustering them in a way that does not require much input from the user, and, in the very end, (e) lifting the profiles of clusters or individual texts to higher ranks of the hierarchy to visualize and interpret them.

When applying this approach to texts in Russian, additional issues emerge due to lack of adequate tools for both linguistic analysis and taxonomy development in Cyrillic alphabet.

The remainder describes the techniques that are being under development and, in part, is an adaptation of a method in [5]. Our preliminary attempts at applying the techniques to real data are described too. The conclusion states what has been already done and issues yet to be tackled.

This paper comprises research findings obtained in the framework of the research project "Development and adaptation of clustering methods to automate analysis of unstructured texts using domain ontologies" supported by The NRU Higher School of Economics 2011-2012 Academic Fund Program. The project was partly supported by the Program of Fundamental Studies of the NRU Higher School of Economics in 2011.

2 Method's description

2.1 Input information

There are two inputs to the method: (1) a domain ontology and (2) a domain related text collection.

We consider an ontology to be a rooted tree-like structure of topics in the domain, with the parental nodes corresponding to more general topics than the children. Besides the hierarchical relation between the topics, other relations might exist. There can be links between topics from different parts of the tree.

We work with two such sets: (1) the ACM-CCS ontology [1] and a collection of ACM journal abstracts; (2) the VINITI ontology of mathematics and informatics [3] and a collection of teaching syllabuses of mathematics and informatics in The National Research University Higher School of Economics Moscow (NRU HSE, in Russian).

The ACM-CCS ontology is a four-layer rooted tree in which three upper layers are coded as usual, whereas the fourth layer is not coded and can be considered as consisting of descriptions of the third layer topics. The tree has eleven major topics on the first layer, such as B. Hardware, C. Computer Systems Organization, etc. They are subdivided in 81 second-layer topics, which are further divided into third-level topics or so-called leaf topics. Almost all leaf topics are accomplished by topic descriptors that are sets of common phrases or terms corresponding to the topic. There are some cross-references between topics in different partitions. Here is a part of ACM–CCS ontology related to one of the eleven main subjects, D. Software:

D. Software
D.0 GENERAL
D.1 PROGRAMMING TECHNIQUES (E)
D.1.0 General
D.1.1 Applicative (Functional) Programming
D.1.2 Automatic Programming (I.2.2)
D.1.3 Concurrent Programming
Distributed programming
Parallel programming

Three coded layers above are presented by topics D., D.0, D.1.0, etc. The topic D.1.3 is supplied with the topic description involving two terms. The topics D.1 and D.1.2 have references to topics E and I.2.2, respectively.

The VINITI ontology of mathematics and informatics is the most extensive ontology of mathematics domain that is available in Russian [3]. It is an unbalanced rooted tree of mathematics and informatics topics supplied with many cross-references.

Usually, these ontologies are used to annotate documents or publications in large collections such as the ACM portal library or VINITI journals library. Here we concentrate on a different aspect of using ontologies – a procedure for abstracting concepts from text documents.

Accordingly, we consider two collections of texts.

First, we have taken an issue of ACM Journal on Emerging Technologies in Computing Systems (JETC), which is a free access journal [2, 8]. Each publication is represented by three items: 1) an abstract; 2) a set of keywords provided by authors; 3) a set of index terms that are ACM–CCS ontology topics, used on the journal's web site to manually index the article. We use both the abstract and keywords to represent the contents of an article.

Second, we have NRU HSE teaching syllabuses for the courses involving Mathematics and/or Informatics as they are taught in the School of Applied Mathematics and Informatics at NRU HSE. They can be easily downloaded from the NRU HSE web-site (`http://www.hse.ru`).

2.2 Method's composition

The method takes in a text, generates its profile, and then proceeds to further stages described below. The profile is a list of ontology topics generated for the input text. This is based on estimations of the degree of similarity between ontology topics and the text derived by using the so-called annotated suffix tree (AST) techniques [8].

This is the sequence of the method's steps:

1. texts and ontology preprocessing
2. presenting the texts as annotated suffix trees (ASTs)
3. evaluating similarity between the ontology topics and the texts according to texts' AST features
4. constructing the text profiles
   (a) computing the similarity matrix of the ontology topics according to the text corpus
   (b) computing the similarity matrix between the texts
5. finding and analyzing text clusters
6. finding clusters of ontology topics
7. mapping the clusters into higher layers of the ontology structure.

Steps 5 and 6 can be skipped so that the following applies to both individual texts and text corpora.

2.3 Texts preprocessing

Each text is split into sentences and each ontology topic usually consists of one sentence. We represent both the texts and ontology as sets of sentences that are taken as strings. To construct a simple machine representation, two stages need to be completed:

1. extracting meaningful parts from texts;
2. removing from them the unnecessary symbols such as html tags, punctuation marks, etc, and transforming them to the lower case.

While the latter step can be easily done automatically, the first one is conventionally manual. For example, NRU HSE teaching syllabuses include not only subjects but some administration issues as well. Another obstacle to automation of the process is the fact that the teaching syllabuses have no unified template but rather are formatted and stored in different styles and formats.

2.4 Annotated suffix tree representation of a text

An annotated suffix tree (AST) is a data structure used for computing and representing of all the text fragments frequencies. AST for a string is a rooted tree, where each node is labeled with one character and one number. Each path from the root to a leaf reads/encodes one of the string suffixes. Frequency of a node is the frequency of fragment occurrences in the string which is read/encoded by the corresponding path from the root to the node. AST for a collection of strings reads/ encodes every suffix of each string and their occurrence frequency in all the strings.

Examination of a set consisting of an ontology and a text corpus is done according to procedures described in [8]. It involves constructing an AST for each text and evaluating the relevance of each ontology topic to the text. The details can be found in [8]. Therefore, the ontology topics assigned with the highest estimations for a text are selected to form the text's profile either as a fuzzy set of the estimates or a crisp set of the selected topics.

In some cases, a text can be seen as a more complicated entity than just a set of strings. If, as it happens, keywords for a text are provided, one AST may be not enough to represent the keywords-text combination. The ontology, being a hierarchic structure with cross-references, should not be treated as a primitive set of strings too. Here we come to an advanced model of the query set.

2.5 Generating profiles: abstracts, keywords, cross-references

Consider a journal publication that is represented by its abstract together with keywords. On the one hand, keywords may be considered as part of the abstract. Hence after building an AST for the strings of the abstract, keywords can be added one by one to the tree. On the other hand, keywords can be treated as being apart from the abstract as a different constituent of the publication. In this case, one should build two ASTs. First is constructed for strings in the abstract, the second is built for the key-

words. Thus the process of ontology topics evaluation has to be repeated twice, using both of the created ASTs. These estimations are to be summed, possibly with different weights, to form the total ontology topics estimation.

To take into account the third constituent, the cross-references, let us first imagine an ontology as a graph structure. It is composed of two parts: 1) a tree structure that is the hierarchic relation between ontology topics; 2) random references between ontology topics at various layers that can be interpreted as edges of the ontology graph. Hence let us define the distance between two topics as the length of the path between them. If there is no such a path, the distance is set to zero. Now suppose that scores for all the ontology topics are computed. The score of the topic N is amended with the score of the topic N_1 related to N by using the distance between N and N_1. Denoting the distance between N and N_1 by $distance(N, N_1)$, the score of the topic N can be set as

$TotalTopicEval(N) := TopicEval + \alpha^{distance(N, N_1)} TopicEval(N_1)$, where α is a constant such that $0 \le \alpha \le 1$ and $TopicEval$ is the ontology topics scoring function.

2.6 Similarity between ontology topics according to the profiles

The scoring of ontology topics over a text corpus results in a topic-to-text matrix A, where a_{ij} is the total score of topic i in text j. The columns in the matrix A are referred to as profiles of the texts. This matrix can be transformed into a similarity matrix of the ontology topics by computing dot products of rows of matrix A. This allows us to use similarity clustering methods, including spectral clustering as discussed in [11] and [9].

For the sake of simplicity, the procedures of clustering and cluster lifting further described are based on the similarity matrix (and therefore the clusters) involving only leaf topics. Therefore, all texts profiles are to be cleaned of upper layer topics before forming the similarity topic-to-topic matrix.

2.7 Spectral clustering

Additive Fuzzy Spectral Clustering method (FADDIS) [5] combines the Additive Fuzzy Clustering Model and the Spectral Clustering approach. The Spectral Clustering approach relies on the eigenstructure of the similarity matrix. Additive Fuzzy Clustering Method finds one cluster at a time by subtracting the similarities taken into account by preceding clusters from the initial similarity matrix [5]. Therefore FADDIS method sequentially finds the cluster membership vector and its intensity using the maximum eigenvalue and corresponding eigenvector of the residual similarity matrix. A special attention should be given to the data pre-processing stage: FADDIS involves the pseudo-inverse Laplace transformation of the initial similarity

matrix. It was shown by experiments that such a transformation may make more clear the structure of clusters to be extracted.

2.8 Query set lifting over ontology

After a fuzzy, or crisp, topic set is extracted as a cluster or single text profile, this set is considered as an abstraction query to the ontology: a few topics of the higher rank are to be found so that the query set is covered, to an extent, by these high-rank nodes, or "head subjects" representing the query set in as general way as possible. We refer to such a result, and the process, as "lifting" the query set over the hierarchically organized ontology.

The lifting algorithm [5,6] proceeds according to the assumption that if all or almost all elements of a set are covered by a high-layer topic, then the set has been lifted to that very topic.

To conform to this hypothesis we introduce a penalty function to be minimized by lifting the query topic set to the ontology root. It is defined as the weighted sum of different types of nodes that do not fit. The "odd" nodes are determined during the lifting procedure. At the level of leaves we have leaves that either belong to the query set or not. A topic that generalizes most of the topics in a cluster is algorithmically interpreted at the head subject for the query set. Those nodes that are covered by a head subject but do not belong to the set are referred to as gaps. Those nodes that are not covered by a head subject but do belong to the query set are referred to as offshoots. The problem is to minimize the total number of head subjects, gaps and offshoots.

We denote the number of head subjects by H, the number of offshoots, by O, and the number of gaps, by G. Then we recurrently minimize the penalty function $P = h * H + off * O + g * G$ at each step of the lifting process; here h, off and g are the corresponding penalty weights.

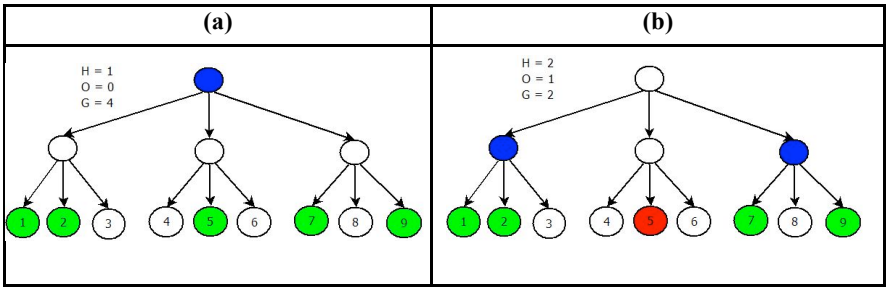


Fig. 1. An illustrative example of mapping a query set to ontology at different penalty weights (see explanation in the text).

Consider the following example of a three-layer ontology and a query set consisting of leaves 1, 2, 5, 7 and 9 (Fig. 1). On Fig. 1(a) there is only one head subject that covers leaves 1, 2, 5, 7 and 9 as well as leaves 3, 4, 6, 8 which are gaps. On Fig. 1(b)

there are 2 head subjects, one offshoot (leaf 5) and 2 gaps: leaves 3 and 8. The opti-

Table 1. Profile A. The spin-wave nanoscale reconfigurable mesh and the labeling problem

AST–profile					ACM index terms
TE	ID	Ontology topic	#	ID	Ontology topic
133.072	C.1.4	Parallel Architectures	1	C.1.4	Parallel Architectures
128.647	C.1.1	Single Data Stream Architectures			
121.059	C.1.2	Multiple Data Stream Architectures (Multiprocessors)			
107.253	D.2.11	Software Architectures	3	C.1.2	Multiple Data Stream Architectures (Multiprocessors)
105.72	C.1.3	Other Architecture Styles			
...	...	...			

mal lifting is determined now by the minimal value of penalty in both cases that depends on the relation between the gap and offshoot penalties.

3 Examples of experimental studies

3.1 ACM Journal abstracts

As mentioned above, we downloaded and examined a number of journal publications. Each of them is represented by an abstract, several keywords and manually indexed ACM-CCS topics.

This last item gives us a tool to evaluate the machine-constructed profiles, based on AST-evaluation of ACM–CCS ontology topics, so that we are able to find how well the estimated topics match those manually selected. Here is an example of profiles for two journal publications [2, 8], one good and the other poor, in Tables 1 and 2.

Each of the tables consists of two parts. The left part presents our machine generated annotated suffix tree profile (AST–profile), the right one stands for the index terms, which were used by publications' authors to annotate the publication. In the tables: TE is the total score of the ontology topic, ID is the index of the ontology topic. '#" denotes the place of the ontology topic in the descending sorted order of the profile. We expect that index terms are to be high scored according to their abstracts by the AST–procedure and to be placed on the top of AST–profile.

The profile A was constructed for the publication that can be previewed on the following web page `http://portal.acm.org/citation.cfm?id=1265951`. The profile B was generated for the publication on `http://portal.acm.org/citation.cfm?id=1265956`. Only five best scoring ontology topics are present here. However, the AST-profile consists of the all leaves of the ACM CCS ontology.

Table 2. Profile B. A self-organizing defect tolerant SIMD architecture			
AST–profile			ACM index terms

TE	ID	Ontology topic	#	ID	Ontology topic
127.503	C.4	PERFORMANCE OF SYSTEMS	40	C.1.2	Multiple Data Stream Architectures (Multiprocessors)
102.03	B.8.1	Reliability, Testing, and Fault-Tolerance			
79.475	B.4.5	Reliability, Testing, and Fault-Tolerance	108	B.4.3	Interconnections (Subsystems)
76.611	B.8.2	Performance Analysis and Design Aids			
			135	B.6.1	Design Styles
72.382	B.3.4	Reliability, Testing, and Fault-Tolerance			
...	...	...			

There are two ontology topics for the publication A. One can see them among the top five ontology topics, on the first and on the third place correspondingly. The publication B is annotated with three ontology topics that are placed on 40th, 108th and 135th places of the AST-profile. While the profile A should be regarded as more or less satisfactory, the profile B is totally inadequate. This difference comes from the difference in the abstracts. The AST–procedure takes into account only matches between the ontology topic's substrings and the abstract's substrings. If no long substrings match, the whole ontology topic will be scored rather low (a long enough subsequence is of 5-7 symbols). In the case of publication B, there are hardly any matches between the ontology topics and the abstract that are longer than 3-4 symbols. In contrast, the abstract of the publication A includes whole words of some ontology topics, such as 'parallel' and 'architecture'. What is more, it is the involvement of the common word 'architecture' that causes so many unrelated ontology topics to be high scored too.

The AST–method is able to detect all the fuzzy matches between an ontology topic and a text. From this point of view the topics "Single Data Stream Architectures" and "Multiple Data Stream Architectures (Multiprocessors)" are identical if only substring "Data Stream Architecture" occurs in the text under examination. The small difference in their total scores may be caused simply by the presence of shorter substrings like 'gle' or even 'e'. Here is the main shortcoming of the AST–method. It is not possible to catch an ontology topic in a text if it is formulated by using other words than in the ontology.

3.2 Syllabuses for HSE courses in Applied Mathematics and Informatics: Preliminary Results

The study of the VINITI Mathematics ontology and the collection of teaching syllabuses showed several shortcomings, both of the ontology and the syllabuses. After

applying the AST–procedure, we derived the topic-to-topic similarity matrix and extracted crisp clusters by means of the FADDIS method mentioned above. Here are a couple of observations. First: Almost each of the crisp clusters contained some topics from the Topology partition. It means that one or two topologic concepts are studied in almost every mathematical course. This should imply that a course in Topology should be included in the curriculum, which is not the case so far. Second: As the VININTY Mathematics ontology has not been updated since 1980's, it was expected that it may have issues in covering more modern topics in mathematics and informatics. Our analysis suggests several nests that should be possibly added to the ontology. For example, the topic 'Lattices' is a leaf in the current ontology. According to our results, it should be a parental node with three children: 'Modular lattices', 'Distributive lattices' and 'Semimodular lattices'. Third, the VININTY Mathematics ontology has been found as being rather imbalanced in the coverage. The profiles of the 'Differential Equations' and 'Calculus' courses according to the ontology are covering all details. This is no wonder because these two constitute almost half of the ontology. Yet branches for less classical subjects such as 'Game Theory' or 'Optimization' are small and not informative.

One more observation is that the main teaching subjects have no matches among the VININTY Mathematics ontology higher ranks. Such is, for instance, 'Discrete Mathematics'.

4 Conclusion

An idea and some initial stages of a different method for abstracting concepts from text documents are presented. It is based on using ontologies as representation of knowledge. We try to simulate the process of abstraction of texts in three coherent steps. First, we match ontology topics to the texts and construct texts profiles by employing text mining techniques. Next step is performed for a corpus of documents: considering leaf topics as the base for abstraction, we find clusters of topics. Finally, we lift cluster query sets to higher layers of the hierarchy to find and visualize head subjects, along with their gaps and offshoots. The head subjects represent the abstraction sought by the method. The method is being developed as an adaptation of the method from [5]. However, a number of novel procedures have been developed in this work as well. Such are using sentence-by-sentence AST modeling, combining different aspects of the texts (such as key-words and cross-referencing) into the scoring system and the like.

The computation experiments lead us to a number of issues that are to be subjects for further developments. The lack of an adequate taxonomy of Mathematics and Informatics in Russian is among them. The AST technology suffers from the effects of repetitive terms such as "architecture", "method", "system" and the like, that act as noise to falsely raise the similarity scores. On the other hand, the scores are dropping down when the texts use slightly different terms for the taxonomy topics. This latter aspect could be treated by using neighbors of the taxonomy topics found in texts retrieved by search engines when queried with the topics. We expect that the neighbors

would allow not only better scoring in the cases of different terminology, but also would be useful in filling in the gaps generated by lifting the head subjects. The other directions for development would be extension of the concept of ontology from the hierarchy to (semi) lattice structures and finding adequate formalisms for dealing with situations at which there are several ontologies related to the texts.

5 List of references

1. ACM Computing Classification System (1998), http://www.acm.org/about/class/1998 (Cited 9 September 2008)
2. Eshaghian-Wilner M. M., Khitun A., Navab S., Wang K. L.: "The spin-wave nanoscale reconfigurable mesh and the labeling problem". ACM Journal on Emergering Technologies in Computing Systems (JETC) 3(2) (2007)
3. I. Yu. Nikol'skaya, V. M. Yefremenkova: "Mathematics in VINITI RAS: From Abstract Journal to Databases". Scientific and Technical Information Processing 35(3) 128-138 (2008)
4. J. Mercadé, A. Espinosa, J-E. Adsuara, R. Adrados, J. Segura and T. Maes: "Orymold: ontology based gene expression data integration and analysis tool applied to rice", BMC Bioinformatics, 10:158 (2009) doi:10.1186/1471-2105-10-158.
5. Mirkin B., Nascimento S., Fenner T., Pereira L. M.: Fuzzy Thematic Clusters Mapped to Higher Ranks in a Taxonomy. International Journal of Software and Informatics 4(3), 257—275 (2010)
6. Mirkin B., Nascimento S., Pereira L.M.: Cluster-lift method for mapping research activities over a concept tree. Recent Advances in Machine Learning II, 245-247 (2010)
7. Pampapathi R., Mirkin B., Levene M.: A suffix tree approach to anti-spam email filtering. Machine Learning 65(1), 309-338 (2006)
8. Patwardhan J., Dwyer C., Lebeck A. R.: "A self-organizing defect tolerant SIMD architecture". ACM Journal on Emergering Technologies in Computing Systems (JETC) 3(2) (2007)
9. Sato M., Sato Y., Jain L.C.: Fuzzy Clustering Models and Applications. Physics-Verlag (1997). ISBN:3790810266
10. V. Karkaletsis, P. Fragkou, G. Petasis and E. Iosif: Ontology Based Information Extraction from Text, Lecture Notes in Computer Science, V. 6050, Knowledge-Driven Multimedia Information Extraction and Ontology Evolution, 89-109 (2011)
11. Von Luxburg, U.: A tutorial in Spectral Clustering. Statistics and Computing, 17 (4), 395-416 (2006)

Extraction and Use of Opinion Words for Three-Way Review Classification Task

Ilia Chetviorkin[1], Natalia Loukachevitch[2],

[1] Faculty of Computational Mathematics and Cybernetics, Lomonosov Moscow State University
ilia2010@yandex.ru
[2] Research Computing Center Lomonosov Moscow State University
louk_nat@mail.ru

Abstract. In this paper, we consider a three-way classification approach for Russian movie reviews. All reviews are divided into groups: “thumbs up”, “so-so” and “thumbs down”. To solve this problem we use various sets of words together with such features as opinion words, word weights, punctuation marks and polarity influencers that can affect the polarity of the following words. Besides, we estimate the maximum upper limit of automatic classification quality in this task.

Keywords: Opinion words, rating-inference problem

1 Introduction

The web is full of customers’ opinions on various products. Automatic collection, processing and summarization of such opinions are very useful for future users. Opinions about the products are often expressed using evaluative words and phrases that have a certain positive or negative sentiment. Therefore, important features in the qualitative classification of opinions about a particular entity are opinion words and expressions used in this domain. The problem is that it is impossible to compile a list of opinion expressions, which will be applicable to all domains, as some opinion phrases are used only in a specific domain; while the others are domain-oriented.

There are two main approaches to the automatic identification of opinion words in texts. The first approach is based on information from a dictionary or a thesaurus. In this approach a small initial set of words is usually chosen manually, and then expanded with the help of dictionaries and thesauruses entries. The basic principle of this approach is that if a word has sentiment polarity, then its synonyms and antonyms have polarity too (orientation may change). Therefore, from the initial set of words, a new, more complete set of opinion words can be constructed [5]. In [4], dictionary definitions are used for opinion words extraction. The basic idea is that words with the same orientation have "similar" glosses.

The second approach – corpus based training. This approach is based on finding rules and patterns in the texts. In [14] word polarity is calculated by comparing the co-occurrence statistics of various words with words “excellent” and “poor”. Authors

assume that words with similar semantic orientation tend to co-occur. The resulting opinion orientation of the words is used to classify reviews to positive and negative.

In this article we will use our method for automatic opinion words extraction based on several text collections, which can be automatically built for many domains. The set of text collections includes: a collection of products reviews with author evaluation scores, a text collection of product descriptions and a contrast corpus (for example, general news collection).

The task of review ranking according to their sentiment has different subtasks. The easiest subtask is to classify reviews into two classes: positive and negative. The quality of two-way classification using topic-based categorization approach for reviews exceeds 80% [11]. In [15] the quality of review classification, based on the so-called appraisal taxonomy, was described as 90.2%.

However, when we turn to the problem of review division into three classes («thumbs up», «thumbs down», «so-so»), the quality of automatic classification decreases significantly [9]. This is partly due to the subjectivity of human evaluation. In [10] the authors conducted a study on the possibility of a human to distinguish reviews rated on a ten-point scale. They describe that if the difference between review scores is more than three points, the accuracy is 100%, two – 83%, one point – 69% and zero points, correspondingly, 55%. Thus, if we classify reviews into a large number of classes, even a human will show low classification accuracy.

In addition, in that paper the difference between evaluation styles of various people was indicated: a review estimated in 5 points (on a ten-point scale) by one person, may express the same opinion and be estimated as 7 points by the other [10]. It was shown that after adjustment to an individual author's style, the quality of the classification increased significantly and reached 75%. But in the classification of 5394 reviews from a large number of authors (494), the achieved accuracy was 66.3%.

In this paper, we analyze various features to improve three-way classification of movie reviews in Russian. For Russian language, studies of this task practically *do not exist*.

We used the following classification features:

- extracted opinion words,
- word weights based on different sources,
- use of polarity influencers: they may reverse or enhance (not, very) polarity of other words,
- length and structure of reviews,
- punctuation marks – as for example in [13] authors used punctuation to reveal sarcastic sentences.

2 Extraction of Opinion Words

For our experiments, we chose movie domain. We collected 28773 film reviews of various genres from online recommendation service *www.imhonet.ru*. For each review, user's score on a ten-point scale was extracted. We called this collection the ***review corpus***.

Example of the review:

Nice and light comedy. There is something to laugh - exactly over the humor, rather than over the stupidity... Allows you to relax and gives rest to your head.

We also needed a contrast collection of texts for our experiments. In this collection concentration of opinions should be as little as possible. For this purpose, we had collected 17680 movie descriptions. This collection was named ***description corpus***.

One more contrast corpus was a collection of one million news documents. We had calculated document frequency of each word in this collection and used only this frequency list further. This list was named ***news corpus***.

2.1 Collection with higher concentration of opinion words

We suggested that it was possible to extract some fragments of the reviews from **review corpus**, which had higher concentration of opinion words. These fragments include:

- Sentences ending with a «!»;
- Sentences ending with a «…»;
- Short sentences, no more than 7 word length;
- Sentences containing the word «movie» without any other nouns.

We call this corpus – ***small corpus***.

2.2 Proposed features

The main task was automatic creation of the opinion word list based on the calculation of various features. Further we exploit the following set of features.

Weirdness. To calculate this feature two collections are required: one with high concentration of opinion words and the other – contrast one. The main idea of this feature is that opinion words will be «strange» in the contexts of the contrast collection. This feature is calculated as follows:

$$\text{Weirdness} = \frac{w_s / t_s}{w_g / t_g}, \text{ where}$$

w_s – frequency of the word in special corpus, t_s – total count of words in special corpus, w_g – frequency of the word in general corpus, t_g – total count of words in general corpus. Instead of frequency one can use the number of documents where the word occurs.

TFIDF. There are many varieties of this feature. We used TFIDF variant described in [1] (based on BM25 function [8]):

$$\text{TFIDF} = \beta + (1 - \beta) \cdot tf \cdot idf \quad (1)$$

$$tf_D(l) = \frac{\text{freq}_D(l)}{\text{freq}_D(l) + 0.5 + 1.5 \cdot \frac{dl_D}{\text{avg\_dl}}} \quad idf(l) = \frac{\log\left(\frac{|c| + 0.5}{df(l)}\right)}{\log(|c| + 1)}$$

- freq(l) – number of occurrences of l in a document,
- dl(l) – length measure of a document, in our case, it is number of terms in a review,
- avg_dl – average length of a document,
- df(l) – number of documents in a collection (e.g. movie descriptions, news collection) where term l appears,
- β = 0.4 by default,
- |c| – total number of documents in a collection.

Deviation from the average score. As we mentioned above we had collected user's numerical score (on a ten point scale) for each review. The main idea of this feature is to calculate average value for each word (sum of review ratings where this word occurs divided into their count) in the collection and then subtract average score of all reviews in the collection from it. Absolute value of this variable is what we need.

$$dev(l) = \left| \frac{\sum_{i=1}^{n} m_i k_i}{k} - \frac{\sum_{i=1}^{n} m_i}{n} \right|$$

$$\sum_{i=1}^{n} k_i = k$$

where l – considered lemma, n – total count of the reviews in the collection, mi – i-th review score, k_i – frequency of the lemma in the i-th review (may be 0).

Frequency of words, which start with the capital letter. The meaning of this feature is the frequency (in the review corpus) of each word starting with the capital letter and does not located at the beginning of the sentence. With this feature we are trying to identify potential proper names, which are always neutral.

2.3 Feature and collection combinations

For our experiments we took top ten thousand words ordered by frequency from the review corpus. We were not interested in words, which occur in the collection only a few times.

Then we divided these words into two groups: adjectives and not-adjectives. The sense of such division is that the majority of opinion words are adjectives and quality evaluation of our approach on them was our special interest. The not-adjective group of words contains nouns, verbs and adverbs. All features were calculated for two above mentioned groups independently.

Thus, we had the following combinations of features and collections:

- TFIDF calculation using the pairs of collections: *small-news*, *small-description*, *opinion-news*, *opinion-description*;
- Weirdness calculation using the pairs of collections: *opinion-news and opinion-description* with document count and *small-description*, *opinion-description* with frequency;
- Deviation from the average score;
- Word frequency in *opinion* and *small collections*;
- Total number of documents in the *opinion corpus*, where the word occurs;
- Frequency of capitalized words.

In addition, separately for description corpus we calculated the following features: frequency, document count, weirdness using *description-news* collections with document count and TFIDF using the same pair. Thus, each lemma had 17 features.

2.4 Data preparation and algorithms

To train supervised machine learning algorithms we needed a set of labeled opinion words. Firstly we tried to extract a list of domain-independent opinion words (near 500) from RuThes thesaurus [7]. The results were unsatisfactory and did not meet expectations as far as this list of general opinion words did not contain domain-specific words, slang words et al. Therefore, we decided to label the full list of ten thousand words manually and then use cross-validation. We marked up word as opinion one in case we could imagine it in any opinion context in the movie domain. All words were tagged by two authors.

In the issue of our mark up we had the list of 3200 opinion words (1262 adjectives, 296 adverbs, 857 nouns, 785 verbs).

Our aim in this part of work was to classify words into two classes: opinion or not opinion. For this purpose we used Rapid Miner[1] data mining tool. We considered the following algorithms: K nearest neighbors (kNN), Naïve Bayes, Perceptron, Neural Network (2 and 3 layers), Logistic Regression and SVM (with linear and radial kernels). For all experiments we used 10 fold cross-validation.

As a result the best algorithms were Logistic Regression for adjectives (F-measure: 68.1%) and Neural Net with 2 layers for not-adjectives (F-measure: 50.9%, unbalanced data). Using them we obtained term lists (adjectives and not adjectives), ordered by the predicted probability of their opinion orientation.

Let us look at some examples of opinion words with high probability value:

Adjectives: *dobryj (kind), zamechatel'nyj (wonderful), velikolepnyj (gorgeous), potrjasajushij (stunning), krasivyj (beautiful), smeshnoj (funny), ljubimyj (love)* etc.,

Not-adjectives*: fuflo (trash), naigranno (unnaturally), fignja (junk), fil'm-shedevr (masterpiece film), tufta (rubbish)* etc.

Obtained opinion word lists with probabilities we use in our review classification task.

[1] http://rapid-i.com/

3 Features for review classification.

In this section we will consider some additional features, which can help us to solve three-way classification problem.

3.1 Word weights

As the main elements of a feature set we used lemmas (words in the normal form) mentioned in the reviews. Word weights can be binary and reflect only word presence in a review or TFIDF formula can be used.

TFIDF is the most popular method of word weighting in information retrieval [8]. For each term in a text, its TFIDF weight can be represented by multiplication of two factors: TF that defines the frequency of this term in the text and IDF specifying occurrence of the term in documents of a text collection. The more frequently such occurrences are, the smaller resulting IDF will be [8]. TF and IDF factors can be defined by various formulas. We used two variants of TFIDF for calculation.

First, we used the simplest form of TFIDF [8]:

$$\mathrm{TF} = \frac{n_i}{\sum_k n_k} \qquad \mathrm{IDF} = \log \frac{|D|}{|(d_i \supset t_i)|} \quad (2)$$

- n_i is the number of occurrences of a term in a document, and the denominator is the sum of occurrence number of all terms in the document,
- |D| – total number of documents in a collection,
- $|(d_i \supset t_i)|$ – number of documents where term t_i appears (that is $n_i \neq 0$).

In addition, we used TFIDF variant, which was described in Section 2.2.

3.2 Polarity influencers

Intuitive is the fact that there are some words, which can affect polarity of other words – *polarity influencers*. To find them the manually compiled set of opinion words (3200 units) was used (see Section 2.4). From the review corpus, we automatically extracted words directly preceding the manually labeled opinion words and ordered them by decreasing frequency of their occurrence.

Then from the first thousand of words from this list, potential polarity influencers were manually chosen (74 words). To assess how significant the effect of these polarity influencers can be, the following procedure was made: we calculated the average score of opinion words in two cases, when they follow the potential polarity influencers and when they occur without them. The average score of a word is the average value of numerical scores of reviews where this word occurs.

After comparison of these average scores, two significant groups of polarity influencers were discriminated. If an opinion word had the high average score (>8) and changed it to the lower when used after a given polarity influencer, and an opinion word with the low average score (<6.7) changed it to the higher one, it means that this polarity influencer *reverses* word polarity (operator –).

If after a polarity influencer, an opinion word with the high score increased its average score, and an opinion word with the low average score decreased its score, it means that this polarity influencer *magnifies* polarity of other words (operator +).

In our review corpus, we found the following polarity influencers:

operator (–): *net (no), ne (not);*

operator (+): *polnyj (full), ochen' (very), sil'no (strongly), takoj (such), prosto (simply), absoljutno (absolutely), nastol'ko (so), samyj (the most).*

On the basis of this list of polarity influencers we substituted sequences "*polarity influencer_word*" using special operator symbols («+» or «–») depending on an influencer, for example:

NE HOROSHIJ (NOT GOOD) → –HOROSHIJ (– GOOD)

SAMYJ KRASIVYJ (THE MOST BEAUTIFUL) → + KRASIVYJ (+ BEAUTIFUL)

NASTOL'KO KRASIVYJ (SO BEAUTIFUL) → + KRASIVYJ (+ BEAUTIFUL)

Modified lemmas were added to the feature set. Now if in a text a word with a polarity influencer occurs, then only the corresponding modified lemma would be added to the review's vector representation, but not both words. This allows us to take into account the impact of polarity influencers.

3.3 Review length and structural features

Movie reviews can be long or short. We chose a threshold on the review length to be 50 words. If a review is long, it often contains overall assessment for a movie at the beginning or at the end. This was the basis for separate consideration of short and long reviews and dividing long reviews into three parts: the beginning (first sentences of a review with total length less than 25 words), the end (last sentences of a review with total length less than 25 words) and the middle (all that is left). We classified each part separately and then aggregated obtained scores in various ways (voting, average).

3.4 Punctuation marks

In addition we included punctuation marks «**!**», «**?**», «**…**» as elements of the feature set.

4 Review classification experiments

Reviews in the working dataset (***opinion collection***) are provided with authors' scores from 1 to 10 points. To map from the ten-point scale to the three-point scale we used the following function: {1-6} → «**1»** (thumbs down), {7-8} → «**2»** (so-so), {9-10} → «**3»** (thumbs up). Thus, the number of reviews belongs to class «3» is approximately 45% of the total.

All reviews from the collection were preprocessed by a morphological analyzer and lemmas with part of speech tagging were extracted.

Authors of previous studies almost unanimously agreed that Support Vector Machine (SVM) algorithm works better for text classification tasks (and review classification task in particular). We also decided to use this algorithm. In view of the fact that we had a large amount of data and features, library LIBLINEAR was chosen [12]. This library had sufficient performance for our experiments. To obtain statistically significant results five fold cross-validation was used. All other parameters of the algorithm were left in accordance with their default values.

We used the following word sets in our classification experiments:

1. An optimal set of opinion words produced by the method described in Section 2. From the list of adjectives and not adjectives (ordered by the probability of their opinion orientation – ***opinweight***) we selected the optimal adjectives and not-adjectives combination. We iterated over the words in these lists and compared quality of classification. We denote this experimental set ***OpinCycle***,
2. Set of words, which was used in [3] to achieve the best results (***OpinContrast***). This set contains near 500 the most frequent words with high opinion probability weight and 400 words with the highest TFIDF score calculated using review and news collections (see Section 2.2),
3. Set of opinion words (3200 units), obtained by manual labeling by two experts (see Section 2.4) (***OpinIdeal***),
4. Set of all words occurring in the review corpus four or more times (***BoW***). The set includes prepositions, conjunctions and particles as well.

From all these word sets, we chose one set, which yields the best classification accuracy, and analyzed the effect of other features: word weights (***tfidf***), opinion weights (***opinweight***), punctuation marks (***punctuation***), polarity influencers (***operators***), review length (***long*** and ***short***).

TFIDF word weights were calculated relying on two formulas: the most well known formula (2) (***tfidf simple***) and formula (1) (***tfidf)*** (see Section 3.1). IDF factor was calculated on the basis not only the *review corpus*, but also two other collections: the *news corpus* (***tfidf news)*** and the *description corpus* (***tfidf descr)***.

To assess the quality of classification we used *Accuracy measure*. It is calculated as the ratio of correct decisions taken by the system to the total number of decisions [2].

The results of the algorithm using different sets of words and features are listed in Table 1. It is worth mentioning that different sets have different coverage area. All reviews without any features from the set were considered as strongly positive ("thumbs up") in accordance with the review distribution between classes. The basic weight of each word is its presence in a review.

Table 1. The classification results using various features

Feature Set	Feature Count	Accuracy %
OpinCycle	1000 *adj* + 1000 *not-adj*	58.00
OpinContrast	884	60.33
OpinIdeal	3200	57.62
BoW	19214	57.37
OpinCycle + tfidf simple	1000 *adj* + 1000 *not-adj*	59.13
OpinContrast + tfidf simple	884	59.43
OpinIdeal + tfidf simple	3200	59.72
BoW + tfidf simple	19214	*62.52*
BoW + tfidf	19214	61.71
BoW + tfidf descr	19214	61.74
BoW + tfidf news	19214	62.90
BoW + tfidf news + operators	22218	63.46
BoW + tfidf news + punctuation + operators	22221	63.17
BoW + tfidf news + opinweight + operators	22218	**64.48**
BoW + tfidf news+ opinweight + operators + short	22218	63.56
BoW + tfidf news + opinweight + operators + long	22218	62.37
BoW + tfidf news + opinweight + operators + avg	22218	63.14

The results obtained by using ***BoW + tfidf simple*** were taken as a *basic line*. The best results were obtained using bag of words (***BoW***) with TFIDF, opinion weights

and polarity influencers. This is clear improvement over *62.52* where ***BoW + tfidf simple*** is applied; indeed the difference is highly statistical significant ($p < 0.001$, $\alpha = 0.05$, Wilcoxon signed-rank test/Two-tailed test). Punctuation marks did not give any quality improvement, although their usage gave slightly better coverage. Formula (1) usage gives slightly better quality than the second one (2). The choice of the news corpus for IDF calculation in (1) draws better results than using the description corpus (***BoW + tfidf descr***) and the review corpus (***BoW + tfidf***).

To increase weights of opinion words in contrast with the other words we used the list of opinion words with probability weights from 0 to 1 (see Section 2.4). We took 800 the most probable adjectives and 200 not-adjectives (we have tried another combinations also) as opinion words. All other words from the feature set were considered with zero ***opinweight***. We modified the weight of each word in the feature vectors in the following manner:

$$\text{wordweight}(x) = \text{TFIDF}(x) \cdot e^{(\text{opinweight}(x) - 0.5)} \quad (3)$$

Thus, we want to increase weights of the words with high ***opinweight***, and decrease for the other words.

The classification accuracy for short reviews (***BoW + tfidf news + opinweight + operators + short***) is better than for long one (***BoW + tfidf news + opinweight + operators + long***). Although, in average (in accordance with review number in each part) the results were not improved (***BoW+ tfidf news + opinweight + operators + avg***).

For the method with the best results of classification ***BoW + tfidf news + opinweight + operators***, we made additional evaluation with so-called *soft borders*, that is if in the basic scale the author of a review puts a boundary score («8» or «6»), then classification of this review as either class «3» or «2» in case of basic «8», and class «2» or «1» in case of basic «6», was not considered as an error. Such weakening of conditions was made on the assumption that even a human distinguishes boundary classes unsatisfactory. The classification accuracy with *soft borders* reaches **76.48%.**

5 Evaluation of reviews by assessors

We also studied the human's ability in three-way review classification. We wanted to know what the maximal quality of classification we could expect from automatic classification algorithms. Significance of such quality upper bound evaluation is declared, for example, in [6]. For a benchmark, we selected one hundred short reviews (with length less than 50 words) and one hundred long reviews (with length more than 50 words) from the review corpus. Assessors did not know the initial score of a review set by its author. Reviews were extracted in such a manner, as to retain original class distribution. All explicit references to the initial score were removed.

Two assessors evaluated the selected reviews. The results of their evaluation are given in Table 4. The last row of the table indicates the agreement in scores between two assessors.

Table 2. The results of humans‘ estimation

Assessor	Assessors accuracy relative to the author of the review	Accuracy with soft borders	Accuracy of the best classification algorithm relative to the assessor
1	72.5	86.5	69.5
2	72.5	78.5	63.5
1 AND 2	71.5	–	–

Thus, we see that human assessors can reproduce the original scores or be consistent with each other only at the level of 71-72%, which is the absolute upper limit to improve the quality of automatic algorithms. Note that the quality of the automatic classification with soft borders, taking into account the possible ambiguity of the border scores, is 76.48%, which is very close to the classification quality of the second assessor (78.5%).

The percentage of coincident scores between the best algorithm and assessor’s scores confirms the results obtained by cross-validation.

6 Conclusion

In this paper, we described a method for opinion word extraction for any domain on the basis of several domain specific text collections. We studied the role of obtained opinion words in three-way classification of movie reviews in Russian. The most significant impact on the quality of classification had the choice of TFIDF formula, polarity influencers accounting and opinion words information usage. We estimated the upper limit of classification quality, which is very close to the results of the best automatic algorithm. This fact makes it difficult to reach further quality improvement of automatic three-way review classification.

Acknowledgements. This work is partially supported by RFBR grant N11-07-00588-a.

References

1. Ageev M., Dobrov B., Loukachevitch N., Sidorov A.: Experimental algorithms vs. basic line for web ad hoc, legal ad hoc, and legal categorization in RIRES2004 (in Russian). In: Proceedings of Russian Information Retrieval Evaluation Seminar (2004)
2. Ageev M., Kuralenok I. Nekrestyanov I.: Official RIRES Metrics (in Russian). In: Proceedings of Russian Information Retrieval Evaluation Seminar. Kazan (2010)

3. Chetviorkin I., Loukachevitch N.: Automatic review classification based on opinion words (in Russian). In: Proceedings of Conference on Artificial Intelligence. Tver (2010)
4. Esuli A., Sebastiani F.: Determining the Semantic Orientation of Terms through Gloss Classification. In: Conference of Information and Knowledge Management (2005)
5. Hu M., Liu B.: Mining and Summarizing Customer Reviews. In: KDD (2004)
6. Kilgarriff A., Rosenzweig J.: Framework and results for English Senseval Computers and Humanities. In: Special Issue on SENSEVAL. pp. 15-48 (2000).
7. Loukachevitch N.V., Dobrov B.V., Development and Use of Thesaurus of Russian Language RuThes. In: Proceedings of workshop on WordNet Structures and Standardization, and How These Affect WordNet Applications and Evaluation. (LREC2002) / Dimitris N. Christodoulakis – 2002. pp 65-70. Gran Canaria, Spain (2002).
8. Manning C., Raghavan P., Schütze H.: Introduction to Information Retrieval. Cambridge University Press (2008)
9. Pang B., Lee L.: Opinion mining and sentiment analysis. Foundations and Trends® in Information Retrieval. Now Publishers (2008)
10. Pang B., Lee L.: Seeing stars: Exploiting class relationships for sentiment categorization with respect of rating scales. In: Proceedings of the ACL (2005)
11. Pang B., Lee L.: Thumbs Up? Sentiment Classification using Machine Learning Techniques. In: Proceedings of EMNLP (2002)
12. Fan R.-E. , Chang K.-W., Hsieh C.-J., Wang X.-R., and Lin C.-J.: LIBLINEAR: A Library for Large Linear Classification. Journal of Machine Learning Research 9. pp. 1871-1874 (2008). Software available at http://www.csie.ntu.edu.tw/~cjlin/liblinear
13. Tsur O., Davidov D., Rappoport A.: ICWCM – a Great Catchy Name: Semi-Supervised Recognition of Sarcastic Sentences in Online Product Reviews. In: International AAAI Conference on Weblogs and Social Media (2010)
14. Turney P.D.: Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In: Proceedings of ACL. pp. 417-424. (2002)
15. Whitelaw C., Garg N., Argamon S.: Using Appraisal Taxonomies for Sentiment Analysis. In: Proceedings of CIKM (2005)

Constructing Galois Lattice in Good Classification Tests Mining

Naidenova, X.A.

Military Medical Academy, Saint-Petersburg, Russian Federation

ksennaid@gmail.com

Abstract. A large class of machine learning algorithms based on mining good classification tests is described. The Galois lattice is used for constructing good classification tests. Special rules are determined for constructing Galois lattices over a given context. All the operations of lattice construction take their interpretations in human mental acts.

Keywords. Good classification test, the Galois lattice, machine learning, human mental operations

1 Introduction

This paper provides a framework for solving diverse and very important problems of constructing machine learning algorithms based on the concept of good classification test. Good classification tests (GCTs) are item sets of a special kind. They serve as a basis for mining implicative logical rules from the data sets. The lattice theory is used as a mathematical language for constructing GCTs. The definition of GCTs is based on correspondences of Galois on $S \times T$, where S is a given set of objects and T is a set of attributes' values (items). Any classification test is a dual element of the Galois Lattice generated over a given context (S, T). All the operations of lattice construction take their interpretations in human mental acts.

2 The Rules of the First and Second Kind

In this paper, we focus on conceptual knowledge the main elements of which are objects, properties (attribute values), and classifications (attributes). Taking into account that implications express the links between concepts (object ↔ class, object ↔ property, property ↔ class) we believe classification reasoning to be based on using and searching for only one type of logical dependencies, namely, implicative dependencies. Implicative assertions are considered as logical rules of the first type including the following ones.

Implication: $a, b, c \rightarrow d$. **Interdiction or forbidden rule**: $a, b, c \rightarrow$ *false* (*never*). This rule can be transformed into several implications such as $a, b \rightarrow$ not c; $a, c \rightarrow$

not b; b, $c \rightarrow$ not a. **Compatibility**: a, b, $c \rightarrow VA$, where VA is the frequency of rule's occurrence. The compatibility is equivalent to the collection of implications as follows: a, b $\rightarrow$ c, VA; a, c $\rightarrow$ b, VA; b, $c \rightarrow b$, VA. Generally, the compatibility rule represents a most frequently observed combination of values. The compatibilities can serve as one of the bases of association rules [1], [2]. **Diagnostic rule**: x, $d \rightarrow a$; x, $b \rightarrow$ not a; d, $b \rightarrow false$. For example, d and b can be two values of the same attribute. This rule works when the truth of 'x' has been proven and it is necessary to determine whether 'a' is true or not. If 'x & d' is true, then 'a' is true, but if 'x & b' is true, then 'a' is false. **Rule of alternatives**: a or $b \rightarrow true$ (*always*); a, $b \rightarrow false$. This rule is a variant of interdiction.

Rules of the second type or classification reasoning rules are the rules with the help of which rules of the first type are used, updated, and inferred from data (instances). They embrace both inductive and deductive reasoning rules. Deductive steps of reasoning consist of inferring consequences from some observed facts with the use of implications. For this goal, the main forms of deductive reasoning are applied: modus ponens, modus tollens, modus ponendo tollens, and modus tollendo ponens.

Let X be a collection of true values of some attributes (or evidences) observed simultaneously. Let r be an implication, left(r) and right(r) be the left and the right parts of r, respectively **Using implication**: if left(r) $\subseteq X$, then X can be extended by right(r): $X \leftarrow X \cup$ right(r). Using implication is based on modus ponens: if A, then B; A; hence B. **Using interdiction:** let r be an implication $y \rightarrow$ not k. If left(r) $\subseteq X$, then k is the forbidden value for all extensions of X. Using interdiction is based on modus ponendo tollens: either A or B (A, B – alternatives); A; hence not B; and either A or B; B; hence not A. **Using compatibility:** let r = 'a, b, $c \rightarrow k$, VA', where VA is the support of r. If left(r) $\subseteq X$, then k can be used to extend X along with the calculated value VA for this extension. Calculating VA requires a special consideration. Using compatibility is based on modus ponens. **Using diagnostic rules:** let r be a diagnostic rule 'X, $d \rightarrow a$; X, $b \rightarrow$ not a', where 'X' is true, and 'a', 'not a' are some alternatives. Using diagnostic rule is based on modus ponens and modus ponendo tollens. There are several ways for refuting one of the hypotheses: (1) to infer either d or b using existing knowledge (with the use of deductive reasoning rules); (2) inferring (with the use of inductive reasoning rules of the second type) new implications for distinguishing between the hypotheses 'a' and 'not a'; (3) to address an expert. **Using rule of alternatives** is based on modus tollendo ponens: either A or B (A, B – alternatives); not A; hence B; either A or B; not B; hence A.

Generating hypothesis or abduction rule. Let r be an implication $y \rightarrow k$. Then the following hypothesis is generated "if k is true, then y may be true". **Using modus tollens**: let r be an implication $y \rightarrow k$. If 'not k' is inferred, then 'not y' is also inferred.

When applied, these rules generate the reasoning, which is not demonstrative. The deductive reasoning rules act by means of extending an incomplete description X of some evidences and disproving impossible extensions. All generated extensions must not contradict with knowledge (the first-type rules) and an observable real situation, where the reasoning takes place. They must be intrinsically consistent (there are no

prohibited pairs of values in such extensions). The inductive reasoning rules deal with known facts and propositions, observations and experimental results to obtain or correct the first-type rules. For this goal, the main inductive cannons stated by a British logician John Stuart Mill [3] are used: the Method of Agreement, Method of Difference, Joint method of Agreement and Difference.

3 The Concept of Good Classification Test

Denote by R a set of objects and by S the set of indices of objects of R. Let $R(+)$ and $S(+)$ be the sets of positive objects and indices of positive objects, respectively. Then $R(-) = R/R(+)$ is the set of negative objects. Denote by T a set of attributes values or items (values, for short) each of which appears in description at least of one of the objects of R.

The definition of good tests is based on correspondences of Galois G on $S \times T$ and two relations $S \to T$, $T \to S$ [4]. Let $s \subseteq S$, $t \subseteq T$. Denote by t_i, $t_i \subseteq T$, $i = 1,\ldots, N$ the description of object with index i. We define the relations $S \to T$, $T \to S$ as follows: $S \to T$: $t = \text{val}(s) = \{\text{intersection of all } t_i: t_i \subseteq T, i \in s\}$ and $T \to S$: $s = \text{obj}(t) = \{i: i \in S, t \subseteq t_i\}$.

Of course, we have $\text{obj}(t) = \{\text{intersection of all } s(A): s(A) \subseteq S, A \in t\}$. Operations $\text{val}(s)$, $\text{obj}(t)$ are reasoning operations related to discovering the general feature of objects the indices of which belong to s and to discovering the indices of all objects possessing the feature t, respectively.

The operation **generalization_of**(t) = $t' = \text{val}(\text{obj}(t))$ gives the maximal general feature for objects the indices of which are in $s(t)$; the operation **generalization_of**(s) = $s' = \text{obj}(\text{val}(s))$ gives the maximal set of objects possessing the feature $t(s)$.

The generalization operations are actually closure operators [4]. A set s is closed if $s = \text{obj}(\text{val}(s))$. A set t is closed if $t = \text{val}(\text{obj}(t))$.

These generalization operations are not artificially constructed operations. One can perform, mentally, a lot of such operations during a short period of time. We give an example of these operations. Suppose that somebody has seen two films (s) with the participation of Gerard Depardieu ($\text{val}(s)$). After that he tries to know all the films with his participation ($\text{obj}(\text{val}(s))$). One can know that Gerard Depardieu acts with Pierre Richard (t) in several films ($\text{obj}(t)$). After that he may discover that these films are the films of the same producer Francis Veber ($\text{val}(\text{obj}(t))$).

Notice that these generalization operations are also used in FCA [5], [6]: a pair C = (s, t), $s \subseteq S$, $t \subseteq T$, is called a concept if $s = \text{obj}(t)$ and simultaneously $t = \text{val}(s)$, i. e., for a concept C = (s, t) both s and t are closed. Usually, the set s is called **the extent** of C (in our notation, it is the set of indices of objects possessing the feature t) and the set t of values is called **the intent** of C.

Let $S(+)$ and $S(-) = S \setminus S(+)$ be the sets of indices of positive and negative objects respectively.

Definition 1. **A classification test** for $R(+)$ is a pair (s, t) such that $t \subseteq T$ ($s = \text{obj}(t) \neq \varnothing$), $s \subseteq S(+)$ & $t \not\subset t'$, $\forall t'$, t' is the description of an object belonging to $R(-)$.

In general case, a set t is not closed for classification test (s, t), i. e., the condition val(obj(t)) = t is not always satisfied; consequently, a classification test is not obligatory a concept of FCA [5].

Definition 2. A classification test (s, t), $t \subseteq T$ (s = obj(t) $\neq \varnothing$) is **good** for $R(+)$ if and only if any extension $s' = s \cup i$, $i \notin s$, $i \in S(+)$ implies that (s', val(s')) is not a test for $R(+)$.

Definition 3. A good classification test (s, t), $t \subseteq T$ (s = obj(t) $\neq \varnothing$) for $R(+)$ is **irredundant** if any narrowing $t' = t \backslash A$, $A \in t$ implies that (obj(t'), t')) is not a test for $R(+)$.

Definition 4. A good classification test for $S(+)$ is **maximally redundant** if any extension of $t' = t \cup A$, $A \notin t$, $A \in T$ implies that (obj($t \cup A$), t') is not a good test for $R(+)$.

It is possible to show that good maximally redundant tests (GMRTs) are closed maximal frequent itemsets and good irredundant tests (GIRTs) are minimal generators [2] of GMRTs.

Generating all types of tests is based on inferring the chains of pairs (s, t) ordered by the inclusion relation. The set of all concepts ordered by the relation ≤, where (s, t) ≤ (s*, t*) is satisfied if and only if $s \subseteq s$* and $t \supseteq t$*, $s \in 2^S$, $t \in 2^t$, is an algebraic lattice with operations $\cap$, $\cup$ [5].

4 Constructing Galois Lattice

Inferring the chains of dual lattice elements ordered by the inclusion relation lies in the foundation of generating all types of classification tests. The following inductive transitions from one element of a chain to its nearest element in the lattice are used: (i) from s_q to s_{q+1}, (ii) from t_q to t_{q+1}, (iii) from s_q to s_{q-1}, and (iv) from t_q to t_{q-1}, where q, q+1, q-1 are the cardinalities of enumerated subsets.

Inductive transitions can be **smooth** or **boundary**. Under smooth transition, extending (narrowing) of collections of values (objects) is going with preserving a given property of them. These properties are, for example, "to be a test for a given class of objects", "to be an irredundant collection of values", "to be a good test for a given class of objects" and some others. A transition is said to be boundary if it changes a given property of collections of values (objects) into the opposite one. For realizing the inductive transitions we use the following rules: **generalization and specification rules, and dual generalization and specification rules**.

The generalization rule is used to get all the collections of objects $s_{q+1} = \{i_1, i_2, \ldots i_q, i_{q+1}\}$ from a collection $s_q = \{i_1, i_2, \ldots i_q\}$ such that (s_q, val(s_q)) and (s_{q+1}, val(s_{q+1})) are tests for a given class of objects. The termination condition for constructing a chain of generalizations is: for all the extension s_{q+1} of s_q, (s_{q+1}, val(s_{q+1})) is not a test for a given class of positive objects. The generalization rule uses, as a leading process, an ascending chain ($s_0 \subseteq \ldots \subseteq s_i \subseteq s_{i+1} \subseteq \ldots \subseteq s_m$) and the operation generalization_of(s) = s' = obj(val(s)) for each obtained collection of objects in case of inferring GMRTs [7].

The specification rule is used to get all the collections of values $t_{q+1} = \{A_1, A_2, \ldots, A_{q+1}\}$ from a collection $t_q = \{A_1, A_2, \ldots, A_q\}$ such that t_q and t_{q+1} are irredundant collections of values and (obj(t_q), t_q) and (obj(t_{q+1}), t_{q+1}) are not tests for a given class of objects. The termination condition for constructing a chain of specifications is: for all the extensions t_{q+1} of t_q, t_{q+1} is either a redundant collection of values or a test for a given class of objects. This rule has been used for inferring GIRTs [8]. The specification rule uses, as a leading process, a descending chain ($t_0 \subseteq \ldots \subseteq t_i \subseteq t_{i+1} \subseteq \ldots \subseteq t_m$). Inferring GIRTs does not require the operation generalization_of(t) = t' = val(obj(t)) for each obtained collection of values.

Both generalization and specification rules realize the Joint Method of Agreement and Difference [3].

The dual generalization (specification) rules relate to narrowing collections of values (objects).

All inductive transitions take their interpretations in human mental acts. The extending of a set of objects with checking the satisfaction of a given assertion is a typical method of inductive reasoning. For example, Claude-Gaspar Bashet de Méziriak, a French mathematician (1581 – 1638) has discovered (without proving it) that apparently every positive number can be expressed as a sum of at most four squares; for example, 5 = 22 + 12, 6 = 22 + 12+ 12, 7 = 22 + 12+ 12+12, 8 = 22 + 22, 9 = 32. Bashet has checked this rule for more than 300 numbers. In pattern recognition, the process of inferring hypotheses about the unknown values of some attributes is reduced to the maximal expansion of a collection of known values of some others attributes in such a way that none of the forbidden pairs of values would belong to this expansion. The contraction of a collection of values is used, for instance, in order to delete redundant (non-informative) values from it. The contraction of a collection of objects is used, for instance, to isolate a certain cluster in a class of objects. Thus, we distinguish lemons in the citrus fruits.

The boundary inductive transitions are used to get:

(1) all the collections t_q from a collection t_{q-1} such that (obj(t_{q-1}), t_{q-1}) is not a test but (obj(t_q), t_q) is a test, for a given set of objects;

(2) all the collections t_{q-1} from a collection t_q such that (obj(t_q), t_q) is a test, but (obj(t_{q-1}), t_{q-1}) is not a test for a given set of objects;

(3) all the collections s_{q-1} from a collection s_q such that (s_q, val(s_q)) is not a test, but (s_{q-1}, val(s_{q-1})) is a test for a given set of objects;

(4) all the collections of s_q from a collection s_{q-1} such that (s_{q-1}, val(s_{q-1})) is a test, but (s_q,val(s_q)) is not a test for a given set of objects.

All the boundary transitions are interpreted as human reasoning operations. Transition (1) is used for distinguishing two diseases with similar symptoms. Transition (2) can be interpreted as including a certain class of objects into a more general one: squares can be named parallelograms, all whose sides are equal. In some intellectual psychological tests, a task is given to remove the "superfluous" (inappropriate) object from a certain group of objects (rose, butterfly, phlox, and dahlia) (transition (3)). Transition (4) can be interpreted as the search for a refuting example. The boundary inductive transitions realize the Methods of Difference and Concomitant Changes [3].

Note that reasoning begins with using a mechanism for restricting the space of searching for tests: (i) for each collection of values (objects), to avoid constructing all its subsets and (ii) to restrict the space of searching only to the subspaces deliberately containing the desired GMRTs or GIRTs. For this goal, admissible and essential values (objects) are used.

First, consider the boundary transition (1): getting all the collections t_q from a collection t_{q-1} such that (obj(t_{q-1}), t_{q-1}) is not a test but (obj(t_q), t_q) is a test for a given set of objects. For this transition, we use **the inductive diagnostic rule** and a method for choosing values to extend t_{q-1}. We extend t_{q-1} by choosing values that appear simultaneously with it in the objects of $R(+)$ and do not appear in any object of $R(-)$. These values are to be said essential ones.

Consider the boundary inductive transition (3): getting all the collections s_{q-1} from a collection s_q such that (s_q, val(s_q)) is not a test, but (s_{q-1}, val(s_{q-1})) is a test for a given set of objects. For this transition, we use **the dual inductive diagnostic rule** and a method for choosing objects to delete them from s_q. By analogy with an essential value, we define an essential object (index of essential object).

Let s be a subset of objects belonging to a given positive class of objects; assume also that (s, val(s)) is not a test. The object t_j, $j \in s$ is to be said an essential in s if ($s\backslash j$, val($s\backslash j$)) is a test for a given set of positive objects. Generally, we are interested in finding the maximal subset sbmax(s) $\subset s$ such that (s, val(s)) is not a test but (sbmax(s), val(sbmax(s)) is a test for a given set of positive objects.

Table 1. Deductive Rules of the First Type Obtained with the Use of Inductive Reasoning Rules

Reasoning rules	Inferred rules
Generalization rule	Implications
Specification rule	Implications
Inductive diagnostic rule	Diagnostic rules
Dual inductive diagnostic rule	Compatibility rules

The dual inductive diagnostic rule can be used for inferring compatibility rules of the first type. The number of objects in sbmax(s) can be understood as a measure of "carrying-out" for an acquired rule related to sbmax(s), namely, val(sbmax(s)) $\rightarrow$ $k(R(+))$ frequently, where $k(R(+))$ is the name of the set $R(+)$.

The inductive rules generate logical rules of the first type (see, please Table 1).

During the lattice construction, the deductive rules of the first type (implications, interdictions, rules of compatibility (approximate implications), and diagnostic rules) are generated and used immediately for pruning the search space.

5 Reducing Inductive Transition to the Second Type Rules

We give some examples of realizing the generalization rule for inferring all GMRTs. Any realization of this rule must allow, for each element s, the following actions: a) to avoid constructing the set of all its subsets, b) to avoid the repetitive generation of it.

Let S(test) be the partially ordered set of elements $s = \{i_1, i_2, \ldots i_q\}$, $q = 1, 2, \ldots, nt - 1$ obtained as a result of generalizations and satisfying the following condition: (s, val(s)) is a test for a given class $R(+)$ of objects. Here nt denotes the number of positive objects. Let $STGOOD$ be the partially ordered set of elements s satisfying the following condition: (s, val(s)) is a GMRT for $R(+)$. Consider some methods for choosing objects admissible for extending s [7].

Method 1. Suppose that S(test) and $STGOOD$ are not empty and $s \in S$(test). Construct the set V: $V = \{\cup s', s \subseteq s', s' \in \{S(\text{test}) \cup STGOOD\}\}$.

If we want an extension of s not to be included in any element of $\{S(\text{test}) \cup STGOOD\}$, we must use, for extending s, the objects not appearing simultaneously with s in the set V. The set of objects, candidates for extending s, is equal to: CAND(s) = $nts \backslash V$, where $nts = \{\cup s, s \in S(\text{test})\}$.

An object $j^* \in$ CAND(s) is not admissible for extending s if at least for one object $i \in s$ the pair $\{i, j^*\}$ either does not correspond to a test or it corresponds to a good test (it belongs to $STGOOD$).

Let Q be the set of forbidden pairs of objects for extending s: $Q = \{\{i, j\} \subseteq S(+)$: ($\{i, j\}$, val($\{i, j\}$) is not a test for $R(+)\}$. Then the set of admissible objects is $select$(s) = $\{i, i \in$ CAND(s): $(\forall j)$ ($j \in s$), $\{i, j\} \notin \{STGOOD$ or $Q\}\}$.

The set Q can be generated before searching for all GMRTs for $R(+)$.

Method 2. In this method, the set CAND(s) is determined as follows. Let $s^* = \{s \cup j\}$ be an extension of s, where $j \notin s$. Then val(s^*) $\subseteq$ val(s). Hence the intersection of val(s) and val(j) must be not empty. The set CAND(s) = $\{j: j \in nts \backslash s$, val($j$) $\cap$ val(s) $\neq \emptyset\}$.

Table 2. The use of reasoning rules of the second type

Process	Rule of the second type
Forming Q	Generating forbidden Rules
Forming CAND(s)	Joint method of Agreement and Difference
Forming select(s)	Using forbidden rules
Forming ext(s)	Method of Agreement
Function to be test(t)	Using implication
Generalization_of($snew$)	Closing operation

The set ext(s) contains all the possible extensions of s in the form $snew = (s \cup j), j \in select(s)$ and $snew$ corresponds to a test for $R(+)$. This procedure of forming $ext(s)$ executes the function generalization_of($snew$) for each element $snew \in ext(s)$.

The generalization rule is a complex process in which both deductive and inductive reasoning rules of the second type are used (please, see Table 2). The knowledge

acquired via a generalization process (the sets Q, L, CAND(s), S(test), *STGOOD*) is used for pruning the search in the domain space.

Searching for only admissible variants of generalization is not an artificially constructed operation. A lot of examples of using this rule in human thinking can be given. For example, if your child were allergic to oranges, then you would not buy these fruits but also orange juice and products containing orange extracts. A good gardener knows the plants that cannot be adjacent in a garden. The problems related to placing individuals, appointing somebody to the post, finding lodging for somebody deal with partitioning a set of objects or persons into groups by taking into account forbidden pairs of them.

6 The Decomposition of Good Test Inferring into Subtasks

To transform good classification tests inferring into an incremental process, we introduce two kinds of subtasks [7], [9]: for a given set of positive examples: 1) Given a positive example t, find all GMRTs contained in t, more exactly, all $t' \subset t$, (obj(t'), t') is a GMRT; 2) Given a non-empty collection of values X such that it is not a test, find all GMRTs containing X, more exactly, all Y, $X \subset Y$, (obj(Y), Y) is a GMRT.

Each example contains only some subset of values from T; hence each subtask of the first kind is simpler than the initial one. Each subset X of T appears only in a part of all examples; hence each subtask of the second kind is simpler than the initial one.

There are the analogies of these subtasks in natural human reasoning. Describing a situation, one can conclude from different subsets of the features associated with this situation. Usually, if one tells a story from his life, then somebody else recalls a similar story possessing several equivalent features. We give, as an example, a fragment of the reasoning of Dersu Usala, the trapper, the hero of the famous book of Arseniev, V. K. [10]. He divided the situation into the fragments in accordance with separate observed facts and then he concluded from each observation independently.

On the shore, there was the trace of bonfire. First of all, Dersu noted that the fire ignited at one and the same place many times. He concluded that here was a constant ford across the river. Then he said that three days ago a man passed the night near the bonfire. It was an old man, the Chinese, a trapper. He did not sleep during entire night, and, in the morning, he did not cross the river and he left. Dersu deduced that only one person was here from the only one track on the sand. He deduced that the person was a trapper on the basis of a wooden rod used for making traps for small animals. That this was the Chinese, Dersu learned from the manner to arrange bivouac. That this was an old man, Dersu deduced after inspecting the deserted footwear: young person first tramples nose edge of foot-wear, but old man tramples heel.

The subtask of the first kind. We introduce the concept of an object's (example's) projection proj(R)[t] of a given positive object t on a given set R(+) of positive examples. The proj(R)[t] is the set $Z = \{z$: (z is non empty intersection of t and t') & ($t' \in R(+)$) & (($obj(z)$, z) is a test for $R(+)$)$\}$.

If the proj(R)[t] is not empty and contains more than one element, then it is a subtask for inferring all GMRTs that are in t. If the projection contains one and only one element t, then (obj(t), t) is a GMRT.

The subtask of the second kind. We introduce the concept of an attributive projection proj(R)[A] of a given value A on a given set $R(+)$ of positive examples. The projection proj(R)[A] = $\{t: (t \in R(+))$ & (A appears in $t)\}$. Another way to define this projection is: proj(R)[A] = $\{t_i: i \in (obj(A) \cap s(+))\}$. If the attributive projection is not empty and contains more than one element, then it is a subtask of inferring all GMRTs containing a given value A. If A appears in one and only one object X, then A does not belong to any GMRT different from X.

Forming the projection of A makes sense if A is not a test and the intersection of all positive objects in which A appears is not a test too, i.e., $obj(A) \not\subset s(+)$ and $t' = t(obj(A) \cap s(+))$ does not correspond to a test for $R(+)$. The procedures using these subtasks for inferring GMRTs can be found in [7], [9].

Restricting the search for tests to a sub-context of given context favors completely separating tests [11], i.e., increases the possibility to find values each of which belongs only to one GMRT in this sub-context. Choosing subcontexts can be controlled by a domain ontology.

We introduce the following operations: choosing object (value) for subtasks, forming and reducing subtasks. The choice of values (objects) for forming subtasks requires a special consideration. It is convenient using essential values in an object and essential objects in a projection for the decomposition of inferring good tests into subtasks of the first or second kind. The following theorem gives the foundation for reducing projections [9].

Theorem 1. Let A be a value from T, (obj(X), X) be a maximally redundant test for a given set $R(+)$ of positive objects and obj(A) $\subseteq$ obj(X). Then A does not belong to any GMRT for $R(+)$ different from (obj(X), X).

Solving subtasks of the first kind initializes deleting object descriptions (item sets), deleting item sets from projection may be followed by deleting values (items) satisfying Theorem 1 or becoming less frequently. Deleting values (items) from item sets may result in deleting item sets not containing any tests for a given class of objects.

7 An Approach to Incremental Inferring Good Tests

Incremental supervised learning is necessary when a new portion of observations becomes available over time. Suppose that each new object comes with the indication of its class membership. The following actions are necessary with the arrival of a new object: 1) checking whether it is possible to perform generalization of some existing rules (tests) for the class to which a new object belongs (a class of positive objects, for certainty), that is, whether it is possible to extend the set of objects covered by some existing rules or not; 2) inferring all good classification tests contained in the new object description; 3) checking the validity of rules (tests) for negative objects, and, if it is necessary, modifying the tests that are not valid (test for negative objects is not valid if it is included in a new (positive) object description). The second act can be

reduced to the subtask of the first kind. The third act can be reduced either to the inductive diagnostic rule followed by the subtasks of the first kind or only to the subtask of the second kind. These acts have been implemented in an incremental algorithm INGOMAR for inferring GMRTs [7].

8 Conclusion

The methodology presented in this paper provides a framework for solving diverse and very important problems of constructing machine learning algorithms based on a unified logical model in which it is possible to interpret any elementary step of logical inferring as a human mental operation. This methodology deals with object classifications and their approximations by the use of classification tests constructed in a given features space. This fact allows managing the procedures of discovering knowledge from data by the aid of domain ontology.

References

1. Jipp, J., Gatzer, U., Nakhaeizadeh, G..: Algorithms for Association Rule Mining – a General Survey and Comparison. ACM SIGKDD Explorations 2(1), 58-64 (2000).
2. Zaki, M.J.: Mining Non-Redundant Association Rules. Data Mining and Knowledge Discovery 9, 223-248 (2004).
3. Mill, J. S.: The System of Logic, Ratiocinative and Inductive, Being a Connected View of the Principles of Evidence, and the Methods of Scientific Investigation, Vol.1. John W. Parker, West Strand, London (1872).
4. Ore, O.: Galois Connexions. Transactions of American Mathematical Society 55(1), 493-513 (1944).
5. Wille, R.: Concept Lattices and Conceptual Knowledge System. Computers & Mathematics with. Applications (Oxford, England) 23(6-9), 493-515 (1992).
6. Ganter, B., Wille, R.: Formal Concept Analysis: Mathematical Foundations. Springer, Berlin/Heidelberg (1999).
7. Naidenova, X .A.: Machine Learning Methods for Commonsense Reasoning Processes. Interactive Models. Inference Science Reference, Hershey, New York (2009).
8. Megretskaya, I. A.: Construction of Natural Classification Tests for Knowledge Base Generation. In: Pecherskij, Y. (ed), The Problem of Expert System Application in National Economy: Reports of the Republican Workshop, pp. 89-93. Mathematical Institute with Computer Centre of Moldova Academy of Sciences, Kishinev, Moldava (1988) (in Russian).
9. Naidenova, X. A., Plaksin, M. V., Shagalov, V. L.: Inductive Inferring All Good Classification Tests. In: Valkman, J. (ed.). Knowledge-Dialog-Solution, Proceedings of International Conference in Two Volumes, Vol. 1, pp. 79-84. Kiev Institute of Applied Informatics, Jalta, Ukraine (1995).
10. Arseniev, V. K.: Dersu, the Trapper: Exploring, Trapping, Hunting in Ussuria. (1st ed.). E. P. Dulton, N.Y. (1941).
11. Dickson, T. J.: On a Problem Concerning Separating Systems of a Finite Set. J. of Comb. Theory 7, 191-196 (1969).

Concept Relation Discovery and Innovation Enabling Technology (CORDIET)

Jonas Poelmans[1], Paul Elzinga[3], Alexey Neznanov[5], Stijn Viaene[1, 2], Sergei O. Kuznetsov[5], Dmitry Ignatov[5], Guido Dedene[1, 4],

[1]K.U.Leuven, Faculty of Business and Economics, Naamsestraat 69,
3000 Leuven, Belgium
[2]Vlerick Leuven Gent Management School, Vlamingenstraat 83,
3000 Leuven, Belgium
[3]Amsterdam-Amstelland Police, James Wattstraat 84,
1000 CG Amsterdam, The Netherlands
[4]Universiteit van Amsterdam Business School, Roetersstraat 11
1018 WB Amsterdam, The Netherlands
[5]National Research University Higher School of Economics (HSE), Pokrovskiy boulvard 11
101000 Moscow, Russia
{Skuznetsov, Dignatov, Aneznanov}@hse.ru
{Jonas.Poelmans, Stijn.Viaene, Guido.Dedene}@econ.kuleuven.be
Paul.Elzinga@amsterdam.politie.nl

Abstract. Concept Relation Discovery and Innovation Enabling Technology (CORDIET), is a toolbox for gaining new knowledge from unstructured text data. At the core of CORDIET is the C-K theory which captures the essential elements of innovation. The tool uses Formal Concept Analysis (FCA), Emergent Self Organizing Maps (ESOM) and Hidden Markov Models (HMM) as main artifacts in the analysis process. The user can define temporal, text mining and compound attributes. The text mining attributes are used to analyze the unstructured text in documents, the temporal attributes use these document's timestamps for analysis. The compound attributes are XML rules based on text mining and temporal attributes. The user can cluster objects with object-cluster rules and can chop the data in pieces with segmentation rules. The artifacts are optimized for efficient data analysis; object labels in the FCA lattice and ESOM map contain an URL on which the user can click to open the selected document.

1. Introduction

In many law enforcement organizations, more than 80 % of available data is in textual form. In the Netherlands and in particular the police region Amsterdam-Amstelland the majority of these documents are observational reports describing observations made by police officers on the street, during motor vehicle inspections, police patrols, interventions, etc. Intelligence Led Policing (ILP) aims at making the shift from a traditional reactive intuition-led style of policing to a proactive intelligence led approach (Collier 2006). Whereas traditional ILP projects are typically based on statistical analysis of structured data, e.g. geographical profiling of street robberies, we go further by uncovering the underexploited potential of unstructured textual data.

In this paper we report on our ongoing research projects on concept discovery in law enforcement and the CORDIET tool that is being developed based on this research. At the core of CORDIET is the Concept-Knowledge (C-K) theory (Poelmans et al. 2009) which structures the KDD process. For each of the 4 transitions in the design square functionality is provided to support the data analyst or domain expert in exploring the data. First, the data source and the ontology containing the attributes used to analyze these data files should be loaded into CORDIET. In the ontology, the user can define temporal, text mining and compound attributes. The text mining attributes are used to analyze the unstructured text in documents, the temporal attributes use these document's timestamps for analysis. The compound attributes are XML rules based on text mining and temporal attributes. The user can cluster objects with object-cluster rules and can chop the data in pieces with segmentation rules. After the user selected the relevant attributes, rules and objects, the analysis artifacts can be created. The tool can be used to create FCA lattices, ESOMs and HMMs. The artifacts are optimized for efficient data analysis; object labels in the FCA lattice and ESOM map contain an URL on which the user can click to open the selected document. Afterwards the knowledge products such as a 27-construction for a human trafficking suspect can be deployed to the organization.

2. Data analysis artifacts

In this section we briefly describe the data analysis and visualisations artifacts that can be created with the CORDIET software. The tool uses Formal Concept Analysis (FCA), Emergent Self Organizing Maps (ESOM) and Hidden Markov Models (HMM) as main artifacts in the analysis process.

2.1 Formal Concept Analysis

Formal Concept Analysis (FCA), a mathematical unsupervised clustering technique originally invented by Wille (1982) offers a formalization of conceptual thinking. The intuitive visualization of concept lattices derived from formal contexts has had many applications in the knowledge discovery field (Stumme et al. (1998), Poelmans et al. (2010b)). Concept discovery is an emerging discipline in which FCA based methods are used to gain insight into the underlying concepts of the data. In contrast to standard black-box data mining techniques, concept discovery allows analyzing and refining these underlying concepts and strongly engages the human expert in the data discovery exercise. The main goal is to make previously inaccessible information available for practitioners easy to interpret visual display. In particular, the visualization capabilities are of interest to the domain expert who wants to explore the information available, but at the same time has not much experience in mathematics or computer science. The details of FCA theory and how we used it for KDD can be found in (Poelmans et al. 2009). Traditional FCA is mainly using data attributes for concept analysis. We also used process activities (events) as attributes (Poelmans et al. 2010c). Typically, coherent data attributes were clustered to reduce the computational complexity of FCA.

2.2 Temporal Concept Analysis

Temporal Concept Analysis (TCA) is an extension of traditional FCA that was introduced in scientific literature about nine years ago (Wolff 2005). TCA addresses the problem of conceptually representing time and is particularly suited for the visual representation of discrete temporal phenomena. The pivotal notion of TCA theory is that of a conceptual time system. In the visualization of the data, we express the "natural temporal ordering" of the observations using a time relation R on the set G of time granules of a conceptual time system. We also use the notions of transitions and life tracks. The basic idea of a transition is a "step from one point to another" and a life track is a sequence of transitions (Elzinga et al. 2010).

2.3 Emergent Self Organising Maps

Emergent Self Organizing Maps (ESOM) (Ultsch 2003) are a special class of topographic maps. ESOM is argued to be especially useful for visualizing sparse, high-dimensional datasets, yielding an intuitive overview of its structure. Topographic maps perform a non-linear mapping of the high-dimensional data space to a low-dimensional one, usually a two-dimensional space, which enables the visualization and exploration of the data. ESOM is a more recent type of topographic map. According to Ultsch, "emergence is the ability of a system to produce a phenomenon on a new, higher level". In order to achieve emergence, the existence and cooperation of a large number of elementary processes is necessary. An emergent SOM differs from a traditional SOM in that a very large number of neurons (at least a few thousands) are used (Ultsch et al. 2005). In the traditional SOM, the number of nodes is too small to show emergence.

2.4 Hidden Markov Models

A Hidden Markov Model (HMM) is a statistical technique that can be used to classify and generate time series. A HMM (Rabiner 1989) can be described as a quintuplet I = (A, B, T, N, M), where N is the number of hidden states and A defines the probabilities of making a transition from one hidden state to another. M is the number of observation symbols, which in our case are the activities that have been performed to the patients. B defines a probability distribution over all observation symbols for each state. T is the initial state distribution accounting for the probability of being in one state at time t = 0. For process discovery purposes, HMMs can be used with one observation symbol per state. Since the same symbol may appear in several states, the Markov model is indeed "hidden". We visualize HMMs by using a graph, where nodes represent the hidden states and the edges represent the transition probabilities. The nodes are labelled according to the observation symbol probability.

3. Data sources

In this research three main data sources have been used. The first data source was the police database "Basis Voorziening Handhaving" (BVH) of the Amsterdam-Amstelland police. Multiple datasets were extracted from this data source, including the domestic violence, human trafficking and terrorism dataset. The second data source was the World Wide Web, from which we collected over 700 scientific

articles. The third dataset consist of 148 breast cancer patients that were hospitalized during the period from January 2008 till June 2008.

3.1 Data source BVH

The database system BVH is used by all police forces of the Netherlands and the military police, the Royal Marechaussee. This database system contains both structured and unstructured textual information. The contents of the database are subdivided in two categories: incidents and activities. Incident reports describe events that took place that are in violation with the law. These include violence, environmental and financial crimes. During our research we analyzed the incident reports describing violent incidents and we aimed at automatically recognizing the domestic violence cases.

Activities are often performed after certain incidents occurred and include interrogations, arrestment, etc., but activities can also be performed independent of any incident, such as motor vehicle inspections, an observation made by a police officer of a suspicious situation, etc. Each of these activities performed are described in a textual report by the responsible officer. We used the observations made by police officers to find indications for human trafficking and radicalizing behavior.

In the year 2005, Intelligence Led Policing was introduced at the police of Amsterdam, resulting in a sharp increase in the number of filed activity reports describing observations made by police officers, i.e. from 34817 in 2005 to 67584 in 2009. These observational reports contain a short textual description of what has been observed and may be of great importance for finding new criminals. The involved persons and vehicles are stored in structured data fields in a separate database table and are linked to the unstructured report in a separate database table using relational tables. The content of all these database tables is then used by the police officer to create a document containing all the information. We however did not use these generated documents because it is possible that the information in the database tables is modified afterwards without updating the generated documents.

Therefore, we wrote an export program that automatically composes documents based on the most recent available information in the databases. These documents are stored in XML format and can be read by the CORDIET toolset.

Before our research, no automated analyses were performed on the observational reports written by officers. The reason was an absence of good instruments to detect the observations containing interesting information and to analyze the texts they contain. Only on the structured information stored in police databases, analyses were performed. These include the creation of management summaries using Cognos information cubes, geographical analysis of incidents with Polstat and data mining with Datadetective.

3.2 Data source scientific articles

For the survey of FCA research articles, we used the CORDIET toolset. Over 700 pdf files containing articles about FCA research were downloaded from the WWW and automatically analyzed. The structure of the majority of these papers was as follows:

1. Title of the paper

2. Author names, addresses, emails
3. Abstract and keywords
4. The contents of the article
5. The references

During our research we used parts 1, 2 and 3. Parts 2 and 3 to detect the research topics covered in the papers. Part 1 was used for doing a social analysis on the authors of the papers i.e. which research groups are working on which topics, etc.

During the analysis, these pdf-files were converted to ordinary text and the abstract, title and keywords were extracted. The open source tool Lucene was used to index the extracted parts of the papers using the thesaurus. The result was a cross table describing the relationships between the papers and the term clusters or research topics from the thesaurus. This cross table was used as a basis to generate the lattices.

We only used abstract, title and keywords because the full text of the paper may mention a number of concepts that are irrelevant to the paper. For example, if the author who wrote an article on information retrieval gives an overview of related work mentioning papers on fuzzy FCA, rough FCA, etc., these concepts may be irrelevant however they are detected in the paper. If they are relevant to the entire paper we found they were typically also mentioned in title, abstract or keywords.

One of the central components of our text analysis environment is the thesaurus containing the collection of terms describing the different research topics. The initial thesaurus was constructed based on expert prior knowledge and was incrementally improved by analyzing the concept gaps and anomalies in the resulting lattices. The thesaurus is a layered thesaurus containing multiple abstraction levels. The first and finest level of granularity contains the search terms of which most are grouped together based on their semantic meaning to form the term clusters at the second level of granularity.

The term cluster “Knowledge discovery” contains search terms “data mining”, “KDD”, “data exploration”, etc. which can be used to automatically detect the presence or absence of the “Knowledge discovery” concept in the papers. Each of these search terms were thoroughly analyzed for being sufficiently specific. For example, we first had the search term “exploration” for referring to the “Knowledge Discovery” concept, however when we used this term we found that it also referred to concepts such as “attribute exploration” etc. Therefore we only used the specific variant such as “data exploration”, which always refers to the “Knowledge Discovery” concept. We aimed at composing term clusters that are complete, i.e. we searched for all terms typically referring to for example the “information retrieval” concept. Both specificity and completeness of search terms and term clusters was analyzed and validated with FCA lattices on our dataset.

3.3 Data source clinical pathways

The third dataset consist of 148 breast cancer patients that were hospitalized during the period from January 2008 till June 2008. They all followed the care trajectory determined by the clinical pathway Primary Operable Breast Cancer (POBC), which structures one of the most complex care processes in the hospital. Before the patient is hospitalized, she ambulatory receives a number of pre-operative investigative tests. During the surgery support phase she is prepared for the surgery she will receive,

while being in the hospital. After surgery she remains hospitalized for a couple of days until she can safely go home. The post-operative activities are also performed in an ambulatory fashion. Every activity or treatment step performed to a patient is logged in a database and in the dataset we included all the activities performed during the surgery support phase to each of these patients. Each activity has a unique identifier and we have 469 identifiers in total for the clinical path POBC. Using the timestamps assigned to the performed activities, we turned the data for each patient into a sequence of events. These sequences of events were used as input for the process discovery methods. We also clustered activities with a similar semantic meaning to reduce the complexity of the lattices and process models. The resulting dataset is a collection of XML files where each XML corresponds with exactly one activity.

4. CORDIET system architecture and business use case diagram

4.1 Business use case diagram

In Poelmans et al. (2010) we instantiated the C- K design theory with FCA and ESOM and showed it was an ideal framework to structure the KDD process on a conceptual level as multiple iterations through a design square. The C-K theory is also at the core of CORDIET. For each C-K phase there are use cases that describe the functionality of the phases. The results of the use cases of a previous phase serve as input for the use cases of the next phases. The business use case model in Figure 1clearly shows this C-K inspired architecture of CORDIET.

The first C-K space, “start investigation”, aims at transforming existing knowledge and information into objects, attributes, ontology elements etc. (conceptualization). The second C-K phase, "compose artifact”, will create artifacts from the data that visualize its underlying concepts and conceptual relationships (concept expansion). The third C-K phase, “analyze artifact”, is about distilling new knowledge from these concept representations. The fourth and last C-K phase is about summarizing this newly gained knowledge and feeding it back to the domain experts who can incorporate it in their way of working. After this final step, a new C-K iteration can start based on the original information and/or newly added knowledge. Iterating though the design square will stop when no new knowledge can be found anymore.

Fig. 1. Business use case diagram of CORDIET.

4.2 The software lifecycles of CORDIET

The architecture of the CORDIET software underwent some serious changes during the development of this research. During the first stage of this research we were working on the domestic violence data (Poelmans et al. 2010) and CORDIET consisted of an FCA, ESOM component and a commercial text mining tool was used to index the documents. Our own programming took care of the documents extraction from the database and the conversion of the data to be used as input for the artefact creation components. This first version had its limitations and was seriously modified for the terrorism and human trafficking research. Amongst others, indexing of documents was done with Lucene.

A separate RDBMS database was used for the maintenance of the ontology with an ERD model. The latest version used a topic map for maintaining the ontology and the open source topic map editor "ontopoly". This latest version will be described in detail in this chapter.

4.3 The development of an operational version of CORDIET

KULeuven and Moscow Higher School of Economics decided to jointly develop an operational software system based on the latest version of CORDIET toolbox. This system will be a user friendly application making visualizations such as FCA, ESOM and HMM available to its users. This version of the toolbox will be based on a distributed web service architecture. Web services are a well standardized, easy to access and flexible piece of technology that can be adapted for different languages and environments.

As a consequence, all input/output activities are represented as XML. Figure 2 shows the general architecture of the new version of CORDIET.

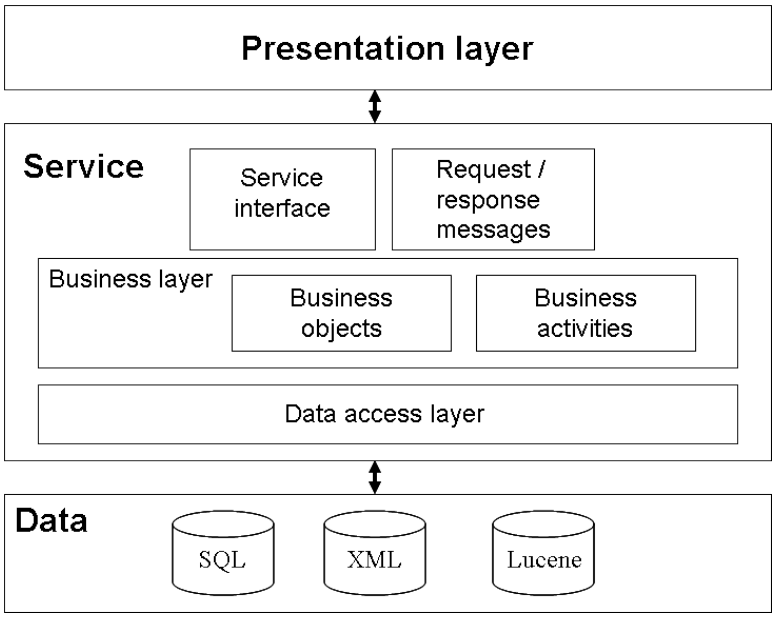


Fig. 2. A representation of the CORDIET web service oriented architecture

4.3.1 Presentation layer

The presentation layer is the graphical user interface where the interactions of the user with the system are handled. This presentation layer is being developed using Silverlight. CORDIET will use two types of main visualizations. The master mode will mainly be used by domain experts who have limited knowledge of data analysis. The user will be able to load a profile for each of the four C-K transition steps, this profile contains all information the tool needs to automatically complete the step in the data analysis. This profile has been prepared by a data analyst. The user can go to the advanced mode. In the advanced mode, he can fully edit an existing or create a new profile. In the advanced mode, a graph-like display will be used to create, modify and compose different attributes.

4.3.2 Service layer

The service layer will be the core of CORDIET. The service interface takes care of the I/O activities with the presentation layer and accessing of the data through the data access layer.

4.3.3 Business layer

The business layer is divided into two sections, the business objects and the business activities which refer to the different activities within the C/K cycle. This layer includes functionality for creating HMMs, ESOMs, Concept lattices, etc.

4.3.4 Data access layer

The data access layer is used to access the data sections: the relational database, the XML data and the Lucene indexes. The data sets consist of a relational database (PostgresSQL), a dataset with XML files and a Lucene index. The data-indexer component reads the XML files from a selected dataset, parses the XML into the SQL database and generates the Lucene index.

4.3.5 Language module

Different languages including English, Dutch and Russian should be supported. The user must be able to choose between these languages. The version of Lucene indexer of documents used has a large variety of analyzers like Russian Analyzer, Dutch Analyzer, German Analyzer etc. The default Analyzer is English.

5. Conclusions

In this paper we briefly described the toolbox, Concept Relation Discovery and Innovation Enabling Technology (CORDIET), for gaining new knowledge from unstructured text data. This toolbox has been embedded within the C-K theory, which captures the essential elements of innovation. The tool uses Formal Concept Analysis (FCA), Emergent Self Organizing Maps (ESOM) and Hidden Markov Models (HMM) as main artifacts in the analysis process.

At the core of the CORDIET toolbox is the business use case where the C-K transitions are mapped on the functionalities of the toolbox. The C-K functionalities are described in detail and demonstrated with real life cases. CORDIET in its current version has become a very powerful toolbox for mining all general reports of the BVH database of the past 5 years. KULeuven and Moscow Higher School of Economics decided to jointly develop an operational software system based on the latest version of CORDIET toolbox. This system will be a user friendly application making visualizations such as FCA, ESOM and HMM more uniformly available to its users where as the current toolbox makes use of integrated open source packages with different user interfaces.

Acknowledgements
Jonas Poelmans is aspirant of the “Fonds voor Wetenschappelijk Onderzoek – Vlaanderen” or “Research Foundation Flanders”.

References

[1] Collier, P.M. (2006) Policing and the intelligent application of knowledge. Public money & management. Vol. 26, No. 2, pp. 109-116.

[2] Elzinga, P., Poelmans, J., Viaene, S., Dedene, G., Morsing, S. (2010) Terrorist threat assessment with Formal Concept Analysis. Proc. IEEE International Conference on Intelligence and Security Informatics. May 23-26, 2010 Vancouver, Canada. ISBN 978-1-42446460-9/10, 77-82.

[3] Poelmans, J., Dedene, G., Verheyden, G., Van der Mussele, H., Viaene, S., Peters, E. (2010c). Combining business process and data discovery techniques for analyzing and improving integrated care pathways. Lecture Notes in Computer Science, Advances in Data Mining. Applications and Theoretical Aspects, 10th Industrial Conference (ICDM), Leipzig, Germany, July 12-14, 2010. Springer

[4] Poelmans, J., Elzinga, P., Viaene, S., Dedene, G. (2009). A case of using formal concept analysis in combination with emergent self organizing maps for detecting domestic violence. In : Lecture Notes in Artificial Intelligence, Vol. 5633(XI), (Perner, P. (Eds.)). Industrial conference on data mining ICDM 2009. Leipzig (Germany), 20-22 July 2009 (pp. 402 p.).

[5] Poelmans, J., Elzinga, P., Viaene, S., Dedene, G. (2010b), Formal Concept Analysis in knowledge discovery: a survey. Lecture Notes in Computer Science, 6208, 139-153, 18th international conference on conceptual structures (ICCS 2010): from information to intelligence. 26 - 30 July, Kuching, Sarawak, Malaysia. Springer.

[6] Rabiner, L.R. (1989) A tutorial on Hidden Markov Models and selected applications in speech recognition. Proceedings IEEE 77 (2): 257-286.

[7] Stumme, G., Wille, R., Wille, U. (1998) Conceptual knowledge discovery in databases using Formal Concept Analysis Methods. PKDD, 450-458.

[8] Ultsch, A. (2003) Maps for visualization of high-dimensional Data Spaces. In proc. WSOM'03, Kyushu, Japan, pp. 225-230.

[9] Ultsch, A., Hermann, L. (2005) Architecture of emergent self-organizing maps to reduce projection errors. In Proc. ESANN 2005, pp. 1-6.

[10] Wille, R. (1982). Restructuring lattice theory: an approach based on hierarchies of concepts. I. Rival (Ed.): Ordered sets, 445-470. Reidel. Dordrecht-Boston.

[11] Wolff, K.E. (2005) States, transitions and life tracks in Temporal Concept Analysis. In: B. Ganter et al. (Eds.): Formal Concept Analysis, LNAI 3626, pp. 127-148. Springer, Heidelberg.

Concept Lattice Implementation in Semantic Structuring of Adjectives

Potemkin S.

Philological faculty, Moscow State University, Russia
potemkin@philol.msu.ru

Abstract. Methods of the formal concepts analysis (FCA) in application to construction of ontological relations in a class of Russian adjectives characterizing appearance of a person with use of WordNet are discussed. Analysis of their semantic paradigm on the basis of the formal context constructed with application of the bilingual dictionary is made.

Keywords. adjectives, concept lattice, hierarchy, dictionary, human appearance

1 Introduction

In the recent years creation of the computer thesaurus of Russian similar in structure and functionality to WordNet thesaurus [16] attracts large interest [1, 7, 8]. Such thesauri give ample opportunities for investigating semantic relations between the meanings of the words of some Natural Language. Unfortunately, the lexical covering by such thesauri for the languages other than English is limited, despite considerable efforts on sinset expansion and their interrelations (sinset is the basic semantic unit of WordNet; a set of English words which code some semantic value). So a necessity of the automated revealing of lexical-semantic relations from the existing sources, such as test corpora or explanatory dictionaries exists. For the decision of this problem methods of the formal concept analysis (FCA) [11, 13, 14] are involved.

We develop methods using bilingual (English-Russian) dictionaries as a source of the formal context and the further construction of a conceptual network for representation of ontological relations in the class of Russian adjectives.

2 Lexical sources

At revealing the structure of semantic paradigm of certain group of words it is necessary to lean against as full lexical sources as possible. We use:

- The common and special English-Russian dictionaries – the lexical database (LDB) [5].

LDB contains English-Russian equivalents from more than 30 common and special dictionaries, including The English-Russian dictionary (ed. Apresian), Muller's dictionary, electronic dictionaries Lingvo, Poliglossum, Promt, and many others. Translation dictionaries are exposed to some kind of natural selection as they are daily used by translators for practical purposes, and the bad dictionary are rejected;

- Assessment of a person appearance (dictionary) [2];
- WordNet [16];
- Explanatory dictionaries by Ojegov, Evgenieva, Sharov's frequency dictionary [9];

In this paper we describe the semantic paradigm of the adjectives characterizing appearance of the person. The frequency of the words in this group is rather considerable: *большой (big)* - 1631 ipm (items per million), *хороший (good)* - 854 ipm, *старый (old)* - 528 ipm, *белый (white)* - 493 ipm, [9] etc. This group is chosen also in view of its importance for specification of system relations of the Russian rating lexicon, notions about types of lexical values, features of connotation, standard lexical associations [3], understanding the structure of a fiction novel [6]. It is important for lingvo-didactics, as a basis for creation of various manuals for speech developing, training in Russian for the Russians and the foreigners, and also for translation of legal, psychological, etc. documents.

Investigation of the meanings of adjectives is similar to investigation of other parts of speech. The component analysis of adjectives with attraction of explanatory dictionaries is used; corpora research is used for the compatibility analysis of syntagma of type adjective - noun which allows to cluster adjectives as the attributes of certain noun for which some classification [12] is already constructed. Methods of direct infield testing for revealing connotations, i.e. narrowing the set of possible syntagmatic partners (adjectives) of the given lexeme (noun) [4] are used. System relations in lexicon are reflected in thesaurus where the lexical meaning of an adjective is frequently the same as this of a semantically similar verb or noun.

It seems promising to use bilingual dictionaries and the existing thesauri like Roget's or the widely used WordNet for revealing of semantics of adjectives. The synonymic and antonymic relations between adjectives are developed well enough, however in this area also attraction of bilingual dictionaries essentially enriches lists of synonyms and especially - antonyms [5]. Other types of relations: hyponymy, meronymy, metonymy and so forth are much less investigated. Revealing of the specified relations between adjectives is of theoretical and practical interest, especially in application to the Automatic Text Processing and Natural Language Understanding. In this case the direct support on the WordNet structure is unproductive. Really, that the semantic organization of qualitative adjectives in WordNet completely differs from the semantic organization of nouns or verbs. Adjectives are organized in clusters linked to a "focal" adjective having an antonym, i.e. antonymic relation is the base semantic relation for coding meaning of adjectives. This approach is connected with the fact, that adjectives have attributive function and that a considerable number of attributes are bipolar. No hierarchical relations similar to the hyponymy relations between nouns or troponymy relations between verbs are revealed in WordNet for adjectives and, as a rule, the direct hypernym is not indicated, instead of it the refer-

of it the reference «Pertains to noun …» is given, that hypernym of an adjective often is a noun, for example for the adjectives designating size (*big, small, narrow, spacious*) a generic hypernym is the noun *"size"*. In this paper we expect, however, to find hierarchical, etc. relations within the class of adjectives.

3 Formal Concept Analysis (FCA)

The formal concept analysis is based on intuitive guess that concept has two parties: *an extent* which contains some objects, and intent which includes all attributes peculiar to these objects [16]. For the formal analysis of concepts it is necessary to define, first of all, *a formal context,* K: = (G, M, I), where G = set of objects; M = set of attributes; and I = the binary relation between elements of G and M, showing, what attributes m are attributed to objects g. It is easy to present a formal context in the form of a table. Table 1 contains some adjectives of Russian as objects, a set of translations of these adjectives – as attributes; the certain Russian word, e.g. *алчный* has a translation equivalent *rapacious*, crossing of the corresponding line and column is marked by cross (X). Derivation operation over the formal context is defined as follows:

$X \subseteq G$: $X \rightarrow X'$: $\{m \in M | gIm$ for all $g \in X\}$; $Y \subseteq M$: $Y \rightarrow Y'$: $\{g \in G | gIm$ for all $m \in X\}$

In our example let *X: = {ХИЩНЫЙ, прожорливый}* and let *Y: = {ravening, wolfish}*. Then *X ' = {ravening, rapacious, ravenous}, Y ' = {ХИЩНЫЙ, жадный}*, further *X "= {ХИЩНЫЙ, жадный, прожорливый }*, etc. It is possible to show that generally $X \subseteq X''$ and $X' = X'''$ and also $Y \subseteq Y''$ and $Y' = Y'''$. The *formal concept* for the given formal context is the pair (A, B) where A = B', B=A', i.e. A = set of objects, having all attributes from the set B, B = set of attributes attributed to all objects of the set A. All formal concepts for the given formal context are generated as (X'', X') or (Y', Y''), for all subsets $X \subseteq G$ or $Y \subseteq M$. A number of algorithms for the fast construction of formal concepts are developed [15]. The cells representing formal concept (A, B) are highlighted in our table; A = {*алчный, грабительский}*; B = {*rapacious, ravenous*}. Relation ≤ establishes a partial order over the formal concepts for the given formal context *B*(K): $(A_1, B_1) \leq (A_2, B_2)$ <-> $A_1 \subseteq A_2$ ($B_2 \subseteq B_1$). This relation is called as the relation *subconcept – superconcept* and ≤ defines a complete lattice $\underline{B}$(K) over *B*(K) which can be depicted in the form of the labeled oriented graph (fig. 1). The nodes this graph are the formal concepts, and the edges reflect the *subconcept – superconcept* relation.

We propose to use thesaurus WordNet and FCA methods to reveal semantic paradigm of Russian adjectives. Basic semantic unit of WordNet is a synset - a set of English words which in aggregate code some semantic meaning. An element of synset is word

meaning (WM) - the meaning of a single word (word-combination), included in a synset.

Table 1. the Formal context for a synset. The objects from the Dictionary are capitalized.

	edacious	esurient	ravening	rapacious	ravenous	voracious	wolfish
ЗВЕРИНЫЙ							X
ЗВЕРСКИЙ							X
СВИРЕПЫЙ							X
ХИЩНЫЙ			X	X	X		X
Алчный				X	X		
Грабительский				X	X		
Волчий					X		X
Голодный		X			X		
голодный как волк					X		
жадный	X	X	X	X	X	X	X
жаждущий					X		
захватнический				X			
изголодавшийся					X		
ненасытный		X		X	X	X	
относящийся к волкам							X
очень голодный					X		
падкий						X	
похожий на волка							X
прожорливый	X	X	X	X	X	X	
свинский				X			
характерный для волка							X
эгоистичный				X			

A word can participate in various synsets, that reflects polysemanticism and homonymy (homography) inherent in the given word. Synsets participate in hypo – hypernymic relations (for nouns), troponymic relations (for verbs), antonymic, meronymic relations and so forth. Synsets, containing adjectives, as a rule, are not captured by hyponymy relations, establishment of hierarchical relations between adjectives is hard both from the theoretical and practical points of view [1,12]. Nevertheless, using synsets for revealing of semantic paradigm of adjectives is obviously possible and promising. We note, first of all, that the bilingual English-Russian dictionary can effectively be applied to expansion of the list of synonyms, and also definition of semantic affinity among Russian synonyms [5]. It is possible to assume, that taking a set of the English words of a synset, $\{e_i\}$, i.e. synonyms with certain meaning, and all their translations into Russian $L_j(e_i) = r_{ij}$, intersection $\bigcap_{ij} r_{ij}$ will contain a set of the Russian words coding meaning, equivalent to the synset $\{e_i\}$ meaning. Owing to various reality partitioning in English and Russian which is the direct reflection of discrepancy of the category assignment and, hence, concept assignment of attributives,

and also propensity of English to the greater detailing of the world a nomination of various features, such intersection as a rule, is empty, or contains several words with very wide semantics. Therefore we propose to use FCA which will allow revealingthe whole structure of sets $\{r_i\}_j$ in their interrelation with synset $\{e_i\}$. Formal context K: = (G, M, I) in this case consists of a set of objects $G = \cup_j \{r_i\}_j$ of all translations of all English words from a synset; set of attributes $M = \{e_i\}$; the binary relation I is defined by attaching the Russian equivalent j to each English word e_i (Table 1).

4 Experimental results and interpretation

The experimental approbation of our technique was carried out over the Dictionary « Assessment of a person appearance» [2], (hereinafter - the Dictionary) containing more than 200 dominants and more than 1200 members of synonymic series of the adjectives attributed to appearance of a person. In particular, 603 adjectives for which more low 1040 conceptual lattices with number of attributes more than 2 have been constucted. For each adjective ar_i all English equivalents $ae_{ij}=L_j(ar_i)$ from the Dictionary containing in the lexical database (LDB) are listed. For every ae_{ij} the set of synsets $\{s_k\} = WN(ae_{ij})$ containing ae_{ij} is defined. For each synset s_k all Russian adjectives which are the translation equivalents of the synset elements are listed; doubles are rejected. Thus, the set of objects G and a set of attributes M of formal context K are received. At this stage we do not carry out the semantic division of inconsistent translation equivalents (which actually exist, e.g. *large-handed* it is translated as *жадный* and as *расточительный*). Also the adjectives concerning appearance of the person are not selected; such selection is carried out later, at an analysis stage of the constructed conceptual lattice. All pairs of equivalents are included in the Table.

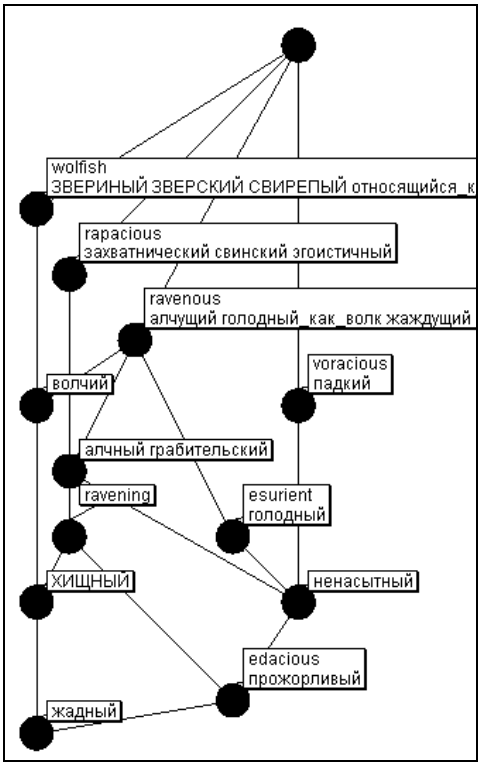


Fig. 1. Conceptual lattice over a formal context of Table 1.

Within the framework of synset №00011320 object ХИЩНЫЙ is a hypernym for objects ЗВЕРИНЫЙ, ЗВЕРСКИЙ, СВИРЕПЫЙ. Such definition of a hypernym generally is not seems to be correct (*зверь (animal)* is not necessary *хищник (predator)*, see Efremova []: *зверь1 = Wild, usually predatory animal*), but as the characteristic of the person the beasty, brutal, furious person most likely is the predatory person. The following hyponymy relations are revealed while analyzing other synsets:

мертвый (dead) ⊆ неподвижный (motionless) ⊆ вялый (languid)
апатичный (apathetic), оцепенелый (freezed) ⊆ вялый (languid)
изящный (graceful) ⊆ тонкий (delicate)
коварный (artful) ⊆ хитрый (sly)
нахальный (impudent), самоуверенный (self-confident) ⊆ дерзкий (daring) ⊆ смелый (brave)
решительный (decisive) ⊆ твердый (hard)
ястребиный (hawk) ⊆ хищный (predatory)
мерзкий (vile), отвратительный (disgusting),противный (offensive), ужасный (awful) ⊆ неприятный (unpleasant)

Some of these relations coincide with those registered in the Dictionary: *изящный (graceful) ⊆ тонкий (delicate), коварный (artful) ⊆ хитрый (sly),* the others are

newly revealed, or contradict the Dictionary, e.g. in the Dictionary adjective *ястребиный (hawk)* is a hyponym of the adjective *беличий* (*squirrel)* (?).

Using FCA it is also possible to find adjectives attributed to the human face which could enter the Dictionary: *бесчувственный* (*insensible), будничный (every day), выцветший (faded),загадочный (mysterious), заспанный (sleepy),зловещий (ominous),искаженный(deformed), легкомысленный (thoughtless), матовый (matte),незамысловатый (plain), нездоровый (unhealthy),неприметный (imperceptible),плоский (flat), полусонный (dozing), придурковатый (foolish),притворный (feigned),разбойничий (predatory), смущенный (confused),сухощавый (lean), флегматичный (phlegmatic), худой (thin)…*

Also the attributive word-combinations which are not included in the Dictionary at all are revealed: *с буйной растительностью* (*with the violent vegetation), наводящий скуку (boring), с хитрецой (sly) …* Comparison of all received hierarchical relations to the Dictionary is out of scope of this research. The proposed method has only allowed to reveal additional lexical units and to establish semantic relations which can be used both in lexicography, and for Automatic Text Processing.

5 Conclusions and research prospects

Complexity of the problem of revealing semantic structure of adjectives is confirmed by the previous researches. Application of methods of the formal concept analysis (FCA) for its decision can appear useful as addition to the corpora – based methods, the component analysis, etc. It is supposed to develop the described methods for formal revealing hierarchical relations from the concept lattice. Besides, expansion of the proposed approach on other semantic relations is possible.

References

1. Azarova, I.V., Sinopalnikova, A.A., Javorsky, M.V. Principles of construction of WordNet-thesaurus RussNet (in Russian) In: *Computer linguistics and intellectual technologies. - Proceedings of the International conference Dialogue'2004* pp.542−547.Moscow (2004)
2. Boguslavsky, V.M. Assessment of appearance of a person, Dictionary. (in Russian) Publishing house "Ast" Moscow (2004)
3. Kedrova, G. E, Potemkin, S.B. Semantic discrimination of homonyms using bilingual dictionary and dictionary of synonyms (in Russian) In: *Proceedings of II International congress "Russian: historical destiny and the present"*, Moscow. (2004)
4. Kobozeva I.M. (2000) Linguistic semantics publishing house «Editorial УРСС», M. 2000, 350 pp.
5. Potemkin, S.B. Lexical database with the imposed semantic metrics (in Russian). In: *Proceedings of II International congress "Russian: historical destiny and the present"*, Moscow (2004)

6. Potemkin, S.B. Detection of event by analysis of antonyms in N.V.Gogol and A.P.Chehov's texts. (in Russian) In: *The word and the dictionary - Proceedings of the International scientific conference «Modern problems of lexicography»*, pp.93-95, Grodno (2009)
7. http://www.cir.ru/.
8. Sukhonogov, A.M. Yablonsky, S.A. (2004) Automation of English-Russian WordNet construction. (in Russian) In: *Proceedings RDCL 2004. September, 29 - October, 1*. Pushino (2004).
9. http://www.artint.ru/projects/frqlist.asp
10. Javorsky, M. B, Azarov, I.V. Structure of attributive meanings in RussNet thesaurus. (in Russian) In: *Proceedings of the International conference Dialogue'2009* pp.542−547 Bekasovo (2009)
11. Cimiano, P, Hotho, A., Staab, S. Learning Concept Hierarchies from Text Corpora using Formal Concept Analysis. In: *Journal of Artificial Intelligence Research*. Volume 24, p.305-339 (2005)
12. Mendes, Sara Adjectives in WordNet. In: *PT//GWC 2006, Proceedings*, pp. 225-230. (2006)
13. Priss, U. Linguistic Applications of Formal Concept Analysis. In Ganter; Stumme; Wille (eds.), *Formal Concept Analysis, Foundations and Applications*. Springer Verlag. LNAI 3626, pp. 149-160. (2005)
14. Stepanova, N.A. Automatic acquisition of lexical-semantic knowledge from corpora. In: *SENSE'09 Proceeding shop* pp.91-100, Moscow (2009)
15. Wille, R. Restructuring lattice theory: an approach based on hierarchies of concepts. In: Rival, I. (ed.) *Ordered Sets.* p.445-470. Dordrecht-Boston, (1982)
16. Fellbaum, Ch. (ed.) WordNet: An Electronic Lexical Database. MIT Press. (1998)

Exploring Semantic Orientation of Adverbs

Potemkin, S.B., Kedrova, G.E.

Philological faculty, Moscow State University, Moscow, Russia
{potemkin,kedr}@philol.msu.ru

Abstract. Sentiment analysis often relies on a semantic orientation lexicon of positive and negative words. Determining the semantic orientation of words is necessary for correct estimation of the content of statements in the media, Internet, in the writings and speech. Qualitative adverbs expressing evaluation, intensity, direction of action are important as the modifiers of the main sentence predicate. In this paper we propose a method for extracting seed set of adverbs from a collection of pairs of antonym. A model based on the representation of a set of synonyms from the Russian lexicons as a graph, and determination the semantic orientation of the adverbs concerning three main dimensions of the semantic differential also demonstrated. The assessment of performance of the method in comparison with the dictionary data shows effectiveness of the method obtained.

Keywords. Sentiment, semantic differential, antonyms, seed set, graph.

1 Introduction

Nowadays, the availability of resources for Natural Language Processing (NLP) remains a hot topic, in particular for Russian especially due to the lack of comprehensive semantic resources, despite efforts made to provide a freely-available Russian WordNet [1]. Ability to establish relativity, similarity, or semantic distance between words and concepts is the basis of computational linguistics. This paper deals with measuring of distance within the syntactic category of adverbs. This set of words is crucial for some applications because adverbs modify or clarify the meaning of other words (verbs, nouns, adjectives). The adverbs are of particular interest to determine the semantic orientation of syntagma containing a main word and its modifier (adverb). Measuring the semantic distance or similarity between the English words most often is based on WordNet [2], and almost exclusively on taxonomic relationships established in this database. So such approach is applicable only to the syntactic categories of nouns and verbs.

The aim of this paper is to extract a list of semantically oriented adverbs and develop the measure of proximity based on dictionaries of synonyms. The article is structured as follows. In Section 1 the problem of extracting the seed set of semantically oriented adverbs from the lexicon of Russian antonyms is discussed. In Section

2 we describe the previously proposed measures of semantic distance between words, as well as an elementary way to map synonyms onto a graph. In Section 3 the basic characteristics of the subjective understanding of the meaning and the measures based on the distance in a graph of synonyms are discussed. Finally, Section 4 presents some results and conclusions. Additionally, we explore the use of visualization techniques to gain insight into the results obtained.

2 Extracting the seed set of adverbs

A number of approaches have been proposed for creating semantic orientation lexicons in English, most of them are computationally expensive and rely on significant manual annotation and large corpora. Particularly, the General Inquirer [3] created in the beginning of the last century is used as the gold standard for assessment the quality of new-generated lexicons. For Russian language there is no open-source and reliable lexicon with positively and negatively marked entries. We propose some approaches to generate a broad coverage semantic orientation lexicon for Russian adverbs which includes both individual words and multi-word adverbial expressions using only dictionaries of antonyms and synonyms, requiring a small amount of manual pruning and database processing.

First of all we have analyzed a list of antonyms collected from published dictionaries of antonyms [4,5]. This list contains 7300+ antonymous pairs (adjectives, nouns, verbs, adverbs and prepositions as well). The semantically oriented words were manually extracted from this list and arranged in 2 separate lists – positive (1859) and negative (2229) words. This seed lexicon could be compared with the GI lexicon which contains orientation labels for only about 3600 entries.

Next step was to extend our seed lexicon to obtain a broad coverage of different texts under consideration concerning sentiment analysis. Automatic approaches to create (English) semantic orientation lexicon and, more generally, approaches for word-level sentiment annotation can be grouped into two kinds: (1) those that rely on manually created lexical resources—most of which use WordNet; and (2) those that rely on text corpora [6]. As a lexical source we use a structured list of Russian synonyms collected from a number of published and Internet-available dictionaries such as [7] and others (11 sources). List of synonyms contains ~600000 word-pairs including ~10000 pairs of adverbs. All synonyms {s(wi)} of each seed word wi receives the same semantic orientation as wi. The number N of occurrences of a synonym s(wi) in the extended set contributed by different seed-words wi, (i=1…N) indicates the confidence of semantic orientation. After manual pruning we have got a list of positively marked (5990, including 731 adverbs) and negatively marked (6853, including 592 adverbs) words. Since the most part of Russian adverbs could be derived as the short form singular neutral or short form plural adjective (3135) the list of semantically orientated adverbs could be expanded.

3 Measures of distance

A number of distance or similarity measures exist for English based (completely or partially) on WordNet. In particular, such measure is defined as the number of edges of the path through the taxonomic relations (IS-A, Part-of, or WordNet's hyponymy relation). In [8] the concept of bond length was extended for all relations in WordNet by their clustering in the horizontal (synonyms) or vertical (hyponymy) direction and assigning a penalty for changing the direction of the path motion. Overview of five measures and evaluation of their effectiveness using the associations between the words is given in [9]. Exclusive usage of hyponymy delimits the measure of distance or similarity only to the syntactic categories of nouns and verbs, as hyponymy relations in WordNet are established only for these grammatical categories. Therefore, such measures could not be applied to adjectives and adverbs.

The semantic distance between the words could be determined in the similar way as the definition adopted in graph theory [10]. The simplest approach is just to gather all the words from the Dictionary of synonyms and to link each member of a synonymous group with its dominant word as indicated in the Dictionary. Let G(W,S) be the undirected graph, with W the set of nodes being all the words from the Dictionary with associated part-of-speech, S - the set of edges connecting each member of synonymous group with its dominant word. Every group of synonymous words could be connected to each other and form a clique in G graph. A path P is the sequence of nodes connected by edges of G and geodesic is the shortest path between two nodes. Geodesic distance, D(wi,wj) between two words wi and wj is the length (number of edges) of the shortest path between wi and wj. If there is no path between wi and wj, the distance between them is infinity. The minimal path-length defines a metric on the set of synonyms. All axioms of the metric space are fulfilled in this case. Usually synonymous groups comprises the words of the same grammatical category and entire graph G is decomposed into disjoint sub-graphs or networks for nouns, verbs, adjectives and adverbs. (Fig. 1). In each network exists a maximal connected component that contains 70-90% of all nodes of the graph constructed from the Dictionary of synonyms. Maximum component in the class of Russian adverbs contains about 8500 words. The words in this connected component could be analyzed using the metric defined by the length of geodesics.

4 Semantic orientation of adverbs

Classical work on the measurement of emotional or affective values in texts is the theory of semantic differential by Charles Osgood. Word meaning in cognitive psychology, is "a strictly psychological one: those cognitive states of human language users which are necessary antecedent conditions for selective encoding of lexical signs and necessary subsequent conditions in selective decoding of signs in messages." [11]. Semantic differential method was applied mainly to the adjectives measured in such dimensions as active/passive, good/bad, positive/negative, beautiful/ugly, etc. Each pair of bipolar adjectives is a factor or an axis in the method of semantic

differential. Application of factor analysis to extensive empirical material gave an unexpected result: most of the variance in judgment could be explained by only three major factors including the evaluative factor (e.g., positive/negative); the potency factor (e.g., strong/weak); and the activity factor (e.g., active/passive). Among these three factors, the evaluative factor has the strongest relative weight for determining the semantic orientation.

Turning to the selected Russian adverbs, we note that the vast majority of adverbs is matched with the words of other parts of speech primarily with the adjectives (cheerful – cheerfully // бодрый - бодро, brutal – brutally // жестокий - жестоко), so that the semantic differential can be naturally extended to motivated adverbs, which bear semantic meaning and, accordingly, deliver the information on their semantic orientation. All three pairs of bipolar adverbs negatively/positively (плохо/хорошо); weakly/strongly (слабо/сильно), passively/actively (пассивно/активно) are contained in the maximal component of the sub-graph of synonymous adverbs Gadv. One can assume that the distance to positively (хорошо) is a measure of positive assessment of an adverb. However, it is easy to show that this measure is in fact rather controversial.

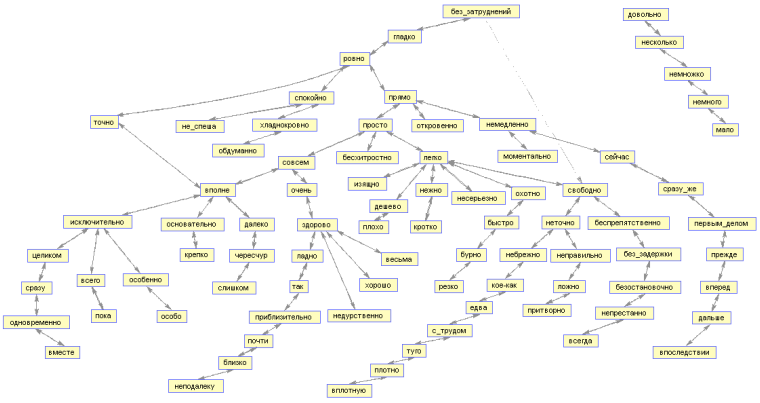


Fig. 1. A fragment of the maximum connected component subgraph of adverbs, G_{adv}.

A striking example of this is that the words *positively (хорошо)* and *negatively (плохо)* are closely related through the path of synonyms. There is a sequence of only 5 words in English (*negatively, hardly, tightly, thoroughly, comprehensively, soundly, positively*), and 6 words in Russian (*плохо, дешево, легко, просто, совсем, очень, здорово, хорошо*) – see Fig. 1 – connecting opposites, each pair of words in this sequence is certainly synonymous (at least in one of their meanings). Thus, we find that *d(positively, negatively)* = 6; *d(хорошо, плохо)* = 7. Despite the fact that the adverb *positively (хорошо)* and *negatively (плохо)* have opposite meanings, they are closely related by synonymy path. Of course, this is not due to any error in the Dictionary of synonyms. Partial explanation lies in the wide use of two Russian adverbs *хорошо*

(625 ipm), *плохо* (187 ipm) [12]. Due to the fact that both words are members of the maximum connected component of G_{adv} sub-graph, we can consider not only the shortest distance from any adverb to "*positively*", but the shortest distance to its antonym, "*negatively*". This idea is concretized [13] in the definition of EVA function, which allows to measure the relative distance from the word of two opposites, "*positively*" and "*negatively*":

$$EVA\ (w) = (d\ (w, neg)\text{-}d\ (w, pos)) / d\ (neg, pos).$$

Under the assumption that there is no word "worse than *negatively*" or "better than *positively*" the values of EVA lie in the interval [-1,1], for example, the word "honestly" is evaluated by function EVA (*honestly*) gives a value of 1 as follows EVA (*honestly*) = (d (*honestly*, *neg*) - d (*honestly, pos*)) / d (*pos, neg*) = (8-2) / 6 = 1,. The measures for other Osgood's dimensions is defined similarly. For the potency factor the function: POT (w) = (d (w, *weakly*)-d (w, *strongly*)) / d (*strongly, weakly*) is defined; for the activity factor the function: ACT (w) = (d (w, *passively*)-d (w, *actively*)) / d (*actively, passively*) is defined. This fact allows to define measures for any two words belonging to the maximal connected component of the adverbs subgraph.

An assumption on the boundary position of words *negatively/positively* is not entirely justified. Intuitively, *perfectly (превосходно)* is better than positively, *disgustingly (отвратительно)* is worse than negatively. Bearing this in mind and using the geometry of a triangle with vertices {w, *pos, neg*}, we redefine the function of EVA, namely:

$$EVA_1(w)=(d(w, neg)\text{-}d(w, pos))*(d(w, neg)+d(w, pos))/d^2(neg, pos).$$

The values of EVA_1 sometimes are beyond the interval [-1,1]. Similarly, we can redesign POT(w) and ACT(w).

For English adjectives (and motivated adverbs) there exists the source for assessing the measure constructed above in comparison with the independently obtained answers to the "General Inquirer" [11], which contains a set of words to assess three Osgood's factors. Word lists were obtained from the Stanford political dictionary, where each of the 3000 most frequent common words were assessed by three or more experts concerning each Osgood's factor. Thus 765 positive and 873 negative words for the assessment factor were obtained, 1474 strong and 647 weak word for the potency factor and 1568 active and 732 passive words for activity factor. Comparison of results obtained with the General Inquirer gave the values of 70 - 80% of matches, depending on what words were considered as neutral in terms of EVA function.

In the absence of available data for content analysis we used the Russian dictionaries of antonyms as an independent source. Antonymous pair is a pair of words (or rather, the specific meanings of words), one opposed to the other on semantic grounds, such as *hot - cold fast - slow, present - absent*. We suggest that adverbs belonging to pair of antonyms lie on the "opposite sides" of the entire set of adverbs. Methods of multidimensional scaling deliver a mapping of multidimensional space with the defined distance between individual points $d(w_i, w_j)$ onto a space of smaller dimension, namely the plane (Fig. 2). Figure 2 a, b shows that the pairs of antonyms lie near the diameters of the set of adverbs. For a more profound study of the structure of the space of adverbs we have constructed chains of synonyms connecting antonyms pairs within the sub-graph G_{adv}.

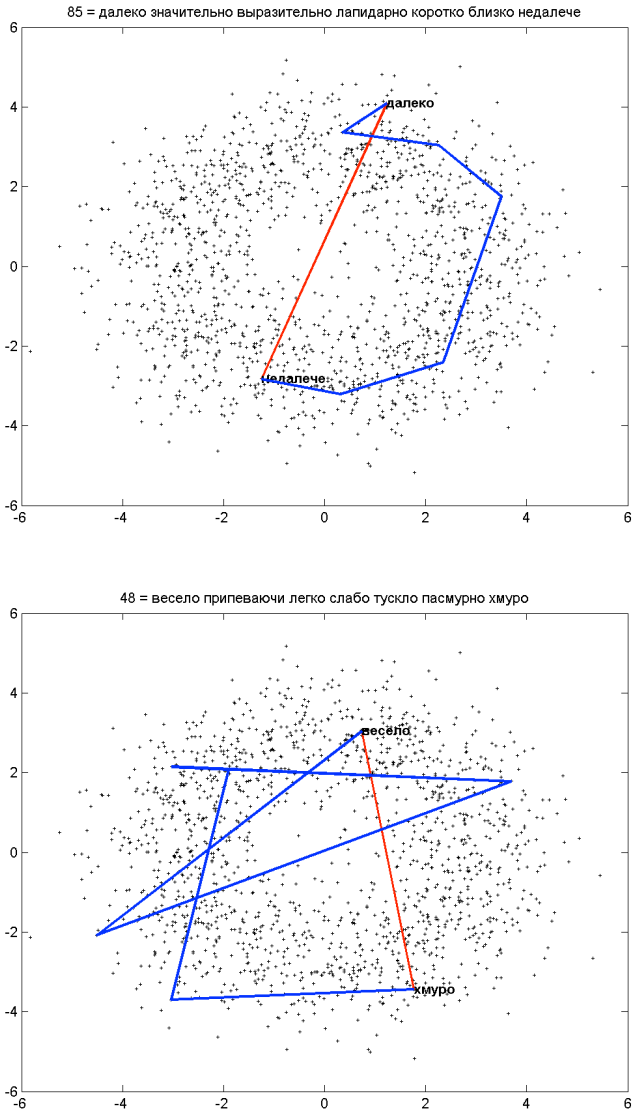


Fig. 2. Two chains of synonyms, joining antonymous pairs of adverbs.
a) up – a consistent path; b) down – an inconsistent result.

Chain in Fig. 2a is a consistent result, i.e. the chain of synonyms passes on the periphery of the set of adverbs and the distances between the synonyms do not exceed the distance between the antonyms. Unfortunately, the situation is not always as fa-

vorable. In Fig. 2b pair of antonyms is close to the diameter, but the chain of synonyms is not at the periphery of the set, but lays in the central part of the set, alternates its direction, and the distances between synonyms is often greater than the distance between antonyms. Probably it is necessary to determine more accurate distance between the words and to choose correctly the axes of the adverb space using the principal components method. These new axes should not coincide Osgood's dimensions.

5 Discussion and conclusions

In this paper we define a measure of the distance between adverbs using synonyms graph. It seems obvious that the choice of similarity measure, or distance largely depends on the type of the problem. The choice of distance measure on the grounds of synonyms is connected with the goal of determining the semantic orientation of adverbs. In contrast to Osgood's semantic differential associated with the reaction of people on the stimulus - words presented, or the possible emotional impact of words, this model is based solely on the lexical material and is intended to represent relatively objective meanings which are fixed in Dictionaries. Further studies will determine the semantic orientation of sentences or the whole text on the basis of the orientation of its constituent words. Our method allows to evaluate other classes of words such as nouns, adjectives and verbs, but this extension will require a significant increase of calculations and special methods for processing large data sets, since an algorithm for computing shortest paths requires $O(n^3)$ operations, where n is the number of words in graph G(W,S).

References

1. Azarova, I.V., Sinopalnikova, A.A., Javorsky, M.V. Principles of construction of WordNet-thesaurus RussNet (in Russian) In: *Computer linguistics and intellectual technologies. - Proceedings of the International conference Dialogue'2004* pp.542−547.Moscow (2004)
2. Fellbaum, C. (ed.), WordNet: An Electronic Lexical Database // Language, Speech, and Communication Series. The MIT Press, Cambridge MA (1998)
3. Stone, P.J. Thematic text analysis: new agendas for analyzing text content. In: C. Roberts (ed.), Text Analysis for the Social Sciences. Lawrence Erlbaum Associates, Mahwah NJ (1997).
4. Vvedenskaia, L.A. Dictionary of Russian Antonyms (in Russian) OOO «AST Publishing House », Moscow (2004)
5. Lvov, M.R. Dictionary of Russian Antonyms (in Russian) "Press-Kniga" (2006)
6. Mohammad, S et.al Generating High-Coverage Semantic Orientation Lexicons From Overtly Marked Words and a Thesaurus In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Singapore (2009)
7. Alexandrova, Z.E. Dictionary of Russian Synonyms (in Russian) "Russkii Iazyk-Media, Moscow (2005)
8. Hirst, G., St-Onge, D. Lexical Chains Representations of Context for the Detection and Correction of Malapropisms". In Fellbaum, C. (ed.) WordNet. An Electronic Lexical Database, The MIT Press, (1998).

9. Budanitsky, A., Hirst, G. Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. In: Workshop on WordNet and Other Lexical Resources. Second meeting of the NAACL, Pittsburgh, (2001)
10. Potemkin, S.B. Semantic Distance over Lexical Database and WordNet In: Proceedings of 10 International Conference «Cognitive Modeling in Linguistics », CML, Montenegro Bechichi (2008).
11. Osgood, C.E., Succi, G.J., Tannenbaum P.H., The Measurement of Meaning.. University of Illinois Press, Urbana IL (1957).
12. Sharov, S. Frequency Dictionary of Russian `http://www.artint.ru/projects/frqlist.asp` (2003)
13. Kamps, J., Marx, M., Robert, J., Mokken, M. Using WordNet to Measure Semantic Orientations of Adjectives In: Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04, Vol. IV) pp. 1115-1118. (2004)

The Third Personal Pronoun Anaphora Resolution in Texts from Narrow Subject Domains with Grammatical Errors and Mistypings

Daniel Skatov and Sergey Liverko

Dictum Ltd, Nizhny Novgorod, Russia
{ds,liverko}@dictum.ru

Abstract. The third personal pronoun anaphora resolution in texts from the Internet sources (forum comments, opinions) with a given subject domain (cars, household appliances etc) is being discussed. A concrete solution to the task is offered. High precision with acceptable recall (and vice versa) is shown by an example of opinions about mobile phones.

Keywords. Computational linguistics, natural language processing, anaphora resolution, machine learning, opinion mining.

1 Introduction

The problem of the third personal pronoun anaphora resolution discussed in this paper consists in the replacement of pronouns such as *"he", "his", "her", "it", ...* with nouns (antecedents) that these pronouns were used instead. Its solution is needed firstly in text mining applications, such as opinion mining (about goods, people) or fact extraction. Without resolved anaphoras those applications lose in recall of their results. The loss degree depends on the type of proceeded texts: e.g., in opinions about goods the density of *"it"* (masculine gender in Russian) pronoun is 1,5 times higher than in news[1].

The known methods of anaphora resolution can be divided into two groups — (1) statistical and (2) syntactical. Methods from class (1) [3] are based on the results of machine learning and are potentially applicable to texts of significantly different nature. Class (2) [1,2] exploits the sentence syntactical parsing tree (or semantic graphs as their derivatives) and as a result the applicability of such methods is limited to relatively «correct» texts (e.g., dossier texts [2]). This article describes a method combining these two approaches in a certain sense.

[1] A random sample of news from [12] (the anaphora density — 0.34 per 1 K and a sample of opinions about mobile phones from the sources such as [13] (the anaphora density — 0,53 per 1 K were used to perform measurements, each one of 1 Mb.

Texts from «real life» are full of typos and specialized slang with their grammar far from correct one:

Ive got a **whit ceise** and buttons peel **gradauly** and they becomes gray no cleaning helps or anything **likethat..!** Weak processor also made upset as well as small memory amount, it works terribly slow. (1)

The method of anaphora resolution, offered by the authors, takes mistypings and the results of syntactic parsing of text fragments (with mistypings corrected) into account. It is adapted to process texts from specific subject domains. Method can work with «correct» texts as well as informal ones (such as opinions or notes). To achieve a high processing quality for texts from a selected domain, a preliminary adjustment to the method is needed. It consists in learning on an unmarked corpus and composing the operating terminological dictionaries.

Three modes of the method have been implemented:

(A) good precision (70-80%) with high recall (90-95%),

(B) approximately equally good precision and recall (75-85%),

(C) excellent precision (up to 95%) with high acceptable recall (40-50%).

The implementation of the technology is represented by a software module called DictaScope Anaphora. It is adjusted to processing opinions about mobile phones from Internet sources. Within the bounds of the article an estimation of recall-precision ratio for processing such kind of data is carried out. The model is being used in the real application for online opinion monitoring. Modes A, B and C were obtained in the process of looking for a solution effective for this application – i.e. the one with high precision on possibly intentionally reduced input data.

2 Problem statement

Basic statement. For each pronoun pr_i, $i = 1,\ldots,N$ from text T choose the resolving pronoun (antecedent) a_i. *Remark.* In certain cases it is impossible to choose a_i, e.g.:

This mobile phone has a sensor screen. It's very inconvenient. **(screen or phone?)** (2)

Resolving of such an ambiguity (which can conditionally be called semantic) is a hard task even for a human, as both variants are of equal possibility. In the current problem statement it is offered either to choose a concrete antecedent or not to resolve the anaphora.

Advanced statement. It sometimes turns out that an acceptable precision of selecting a sole variant is unreachable. Therefore the following task specification is proposed: for each pronoun pr_i, $i = 1,\ldots,N$ form a list of possible resolving variants $\left(a_i^1,\ldots,a_i^{l_i}\right)$ sorted in accordance with their ranks (the first one is the best). Then a_i^1

can be chosen as a_i. In case a requirement of a high recall takes place (e.g., for posterior hand processing of results) it is sufficient to ensure high quality of ranking.

The variants of resolving antecedents can be supplied with real-value weights $w = w\left(a_i^k\right) \in (0,1]$, $i \in \{1,\ldots,N\}$, $k \in \{1,\ldots,l_i\}$, which correspond to each variant's confidence.

Traits. Let's resort to an example to make the task statement clear:

```
bought it for business, very useful because [it] {* =
0.652166, business = 0.2371, NULL = 0.168611} supports
two sim cards. Nice, big display, no dead spaces found
on [it]{display = 0.466248, * = 0.284525, NULL =
0.0777368, business = 0.0101848} (3)
```

For pronoun $pr_1 = «it»$ the list of variants is formed ($a_1^1 = «*»$, $a_1^2 = «bussiness»$, $a_1^3 = «NULL»$) with weights $w\left(a_1^1\right) \approx 0.65$, $w\left(a_1^2\right) \approx 0.237$, $w\left(a_1^3\right) \approx 0.1686$ (similarly for $pr_2 = «it»$). There are also special «*» and «*NULL*» designations:

- «*» — «the current object of discourse», so-called «implicit» antecedent. This is typical for opinions and reviews — i.e. for texts representing direct speech in writing. In the example above the word *«phone»* (as well as its concrete model reference) is not found anywhere before $pr_1 = «it»$, though the teller means exactly *«this phone»*.
- «*NULL*» — a directive «not to resolve pronoun». If «*NULL*» is at first position in the list of variants, the pronoun is left unresolved.

Thus, there are two cases in a basic problem statement in which the anaphora will not be resolved:

1. No variants for pronoun resolution is found;
2. «*NULL*» is the first in the ranged list of variants. It is easy to see that if, in case of semantic ambiguity, the probability of the correct choice of antecedent is less than ½, the precision will not fall on the average. Therefore, in this case the choice of «*NULL*» variant is justified.

In the example (3) the task in the basic statement is resolved correctly by choosing the first variant for each pronoun. A solution in a basic statement will be further estimated.

3 Review

The subject area of this paper is covered in the works of three Russian groups.

1. Ermakov A.E., RCO. In [2] empirical regularities of persons referencing are shown for texts from Russian mass media; they can be used to build a mechanism for

anaphora resolution in text sources of this class (with the help of natural language syntactic parser).
2. Tolpegin P., Vetrov D., Kropotov D. Article [3] describes an experience of this group in resolving the third personal pronoun anaphora in news by machine learning methods. The approach is typical for this type of solvers, the precision shown equals 62% on a control collection.
3. Okatiev V., Erechinskaya T., Skatov D. In the report [1] it is shown how pronoun anaphoras of different types can be resolved with the help of syntax parsing trees analysis. This approach is well applicable to the texts in which most of the sentences allow building correct syntax trees.

The specificity of this article — processing texts from narrow subject domains with mistypings and slang — is not touched upon in the works listed above.

The question discussed is more widely represented in foreign scientific works:

- from English-speaking authors patented system [11] and work [8] (which demonstrates values of basic indicators at a level about 80% while using probability model) are first to be mentioned;
- authors of [9] use maximum entropy method to resolve the third personal pronoun anaphora in Chinese, with F-measure about 70%;
- [10] describes an application of machine learning to personal pronouns anaphora resolution in Turkish with recall-precision at about 60-70%.

The overall impression of these works is the following: competent combination of analysis methods and rather full vocabulary data results in recall-precision not less that 70%.

4 Solution

4.1 Lists of variants and attributes

After tokenization (when the lists of grammar values of the tokens are supplemented taking mistypings into consideration) and dividing text into "conditional" sentences all the pronouns are looked through in the text from left to right. A concrete pronoun pr is fixed, $i = 1,\ldots,N$, and list $\mathrm{var}(pr)$ of possible antecedents is formed:

1. from all the words located within $\kappa = 2$ sentences to the left of pr, nouns in concordance with pr_i by gender and number are selected;
2. from the same words pronouns which are in concordance with pr by gender and number are selected and the list $\mathrm{var}(pr)$ is supplemented with nouns that resolve these pronouns.

Possible antecedents can also be found **to the right** of pr; however, not more than 30 examples of this were found in the corpus, with the correct variant also found to

the left of *pr* in ⅓ cases. Therefore, the possible variant location to the right is ignored by the method.

The proposed scheme has a chain character: pronouns on the left of given pr, which are close to it and already resolved, add antecedents which are located to the left of the boundary of the window $\mu = 2$ to $\mathrm{var}(pr)$. The scheme presents a certain compromise: the list can be imprecise but $\mathrm{var}(pr)$ remains quite compact. Advancing the window border κ up to 5 with the chain scheme disabled has led to a noticeable decrease in the solution precision during the experiments, so the decision was made to reject the varying left border.

For the further ranking of the lists $\mathrm{var}(pr)$ a vector of attributes $A(a)$ is calculated for each $a \in \mathrm{var}(pr)$. Let us mention the following attributes from the operational ones:

- $IsVoc \in \{0,1\}$ — the belonging of a to a terminological dictionary *TermVoc*
- $Freq \in \mathbb{N} \cup \{0\}$ — the number of mentionings of the given word (in any form) to the left of *pr* ;
- $Dist \in \mathbb{N}$ — the distance between the pronoun pr and the position of a inside the text (measured in words);
- $IsVerb \in \{0,1\}$ — the presence of direct father in a form of verb in syntax tree for a fragment containing a ;
- $NumNodes \in \mathbb{N} \cup \{0\}$ — the number of nodes in a bush subordinate to a .

The last two attributes have been introduced based on exploring correlation between numeric properties of a tree and resolving antecedents. For example, greater *NumNodes* were often correspondent to proper variants of resolution. These attributes values are set into null in case the tree was not formed.

The distance is measured in words for a number of reasons: (a) to get a valid syntactical unit (clause, noun phrase) was not possible (at that moment) due to the laboriousness of the adaptation of the syntactical parser to the special features of input texts (e.g. the absence of punctuation); (b) a paragraph is too large for being a unit of measure — the majority of opinions consist of one paragraph; (c) windows are measured in sentences and a two-sentence diapason is considered to be sufficient for the research.

IsVoc attribute implements the following idea: taking a subject domain's specificity into account allows to obtain higher quality of analysis. In fact, *IsVoc* allows to raise the priority of variants relating to subject domain of the text — they are of most interest (not always, though).

4.2 The test corpus

To evaluate the work of the methods a corpus of 3M was built from opinions about mobile phones from the sources like [13,14,15]. Due to the specificity of the application the corpus was additionally divided into three groups: positive, negative and neutral opinions, each of 0.8–1.2 M. As a next step it was marked up with the resolved anaphoras according to the following scheme:

- if the correct antecedent could be chosen directly from the text, its occurrence which was closest to the left of the pronoun being resolved was marked in a special way;
- in case of semantic ambiguity the pronoun was marked with «*NULL*» variant;
- the resolving word was written next to the pronoun in the corresponding case.

The statistical characteristics of the corpus were estimated.

- The whole number of 8.3 thousand opinions formed of 37 thousand unique word forms (including mistypings).
- The most frequent opinion length varying from 15 to 35 words; average opinion length — 54 words; the bulk of the opinions containing 10 to 90 words; opinions of more than 100 words are rare. The length scatter — from 2 to 340 words (Fig.1).
- Opinions consisting of one sentence are the most frequent; average opinion length — 4 sentences. The majority of opinions include 1 to 16 sentences; lengths more than 24 sentences are very rare (Fig.2).
- The corpus contains about 6.2 thousand third personal pronouns, including 4.5 thousand ones of masculine gender, 0.8 thousand of feminine gender, 0.7 thousand of plurals. The reason for a great number of masculine pronouns is the subject of the opinions (mobile phones).
- Less than 50% of the opinions do not contain any of the pronouns under research. 35% contain only one pronoun, about 10% — two of them. The maximum is 9 pronouns per opinion (Fig.3).

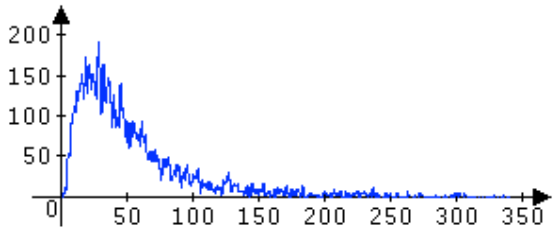


Fig. 1. Distribution of opinions lengths in words

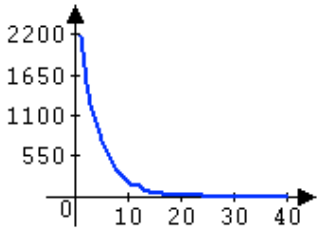


Fig. 2. Distribution of opinion lengths in sentences

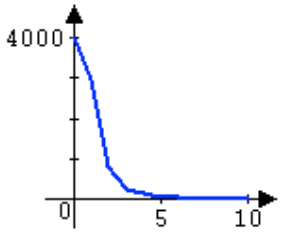


Fig. 3. Opinions distribution by a number of pronouns

4.3 Lexicographical analysis method

At the initial stage of studying a heuristic method for the options ranking was implemented:

- a system of priorities is formed on the set of attributes, which were listed in subparagraph 4.1;
- attribute values for each option are sorted according to the priorities;
- options are sorted lexicographically according to their sets of attributes.

The method resolves all the anaphoras for which it has found variants to the left with precision rate not more than 60%. The experiments in introducing new attributes and varying their priorities were not efficient. This has led the authors to the idea of filtration of the input data in order to achieve higher precision rate.

4.4 SVM-method based on machine learning

Let there be a general set of objects Y, divided into previously unknown classes, and a sample set $O \subset \Omega$, for each element of which its class is known. The task of classification is to answer the question: *"which class does each object* υ *from* Y *belong to"*, knowing only the sample set O (or the probabilities of belonging).

Let us fix a list $\mathrm{var}(pr_i)$ for one specific pronoun pr_i. In this case $O_i = \{A(a) \mid a \in \mathrm{var}(pr_i)\}$, $i = 1,\ldots,N$, and two classes are of interest — "are antecedents" and the inverse to it. Then the first class distance can be taken as $w(a)$.

Now we need to generalize the approach for N pronouns. Each set O_i represents an independent group, each of which consists of two classes — *"is the antecedent for pr_i"* and the inverse one, $2N$ classes for the whole training set. It is impossible to use this classification in practice with a different number $Q \neq N$ of *other* pronouns. In order to get exactly two classes for any number of pronouns, it is necessary to construct an acceptable combination of these groups. For this purpose, the authors propose adding attributes characterizing the group to each set $\omega_i \in O_i$. Thus within the same group all its members are additionally provided with the same set of numbers describing the group. The centroid can be taken as these numbers.

After expanding of the group members a sample set $\overline{O} = \bigcup_{i=1}^{N} O_i$ with the corresponding universe $\overline{\mathrm{Y}}$ and a fuzzy classifier $K(\omega) \in (0,1]$ which determines a distance between υ and the class *"are antecedents"* are constructed.

$K(\upsilon)$ is constructed in a form of so-called probabilistic decision function as described in [5,6] based on a classical C-SVM with a nonlinear kernel [7]. Selection of the core and the constants for the SVM was performed by minimizing the overtraining on the parameters grid while verifying the recall-precision ratio on the training and control samples. In the end, the kernel was chosen to be a polynomial one with a small degree.

Centroids raised the precision of the SVM-method from 70% to 80% (mode A).

4.5 Recall-precision regulator

To reach the precision rate of 90% linear discriminative analysis [4] was used: its aim is to find a line between classes, in the projection on which they are most discernible. With the help of discriminant, pronouns which may be not resolved (for the purpose of rising the precision rate) were identified. The combination of this filtration and SVM-method allowed to reach the desired result (mode C). Along the way, it was managed to derive mode B in which basic rates are balanced in the region of 75-85%.

5 Analysis of the results

5.1 Quality requirements and evaluation

Processing of the input set containing L third personal pronoun anaphoras is carried out in 2 steps.

1. **Filtration of anaphoras.** From the total number of L objects those for which the algorithm: (1) failed to form the set of variants, (2) put «»*NULL* in the first place in the list of variants or (3) eliminated from the examination due to regulator work are deleted. As a result, N anaphoras are left, for each of them the algorithm can choose an antecedent (not necessarily the correct one). If the whole of L anaphoras resolved *correctly* are considered as relevant, the recall rate of this step is N/L while the precision is equal to 1, as all chosen objects (N) are included in the relevant (L).
2. **Resolution of the left anaphoras.** In this step the whole of N anaphoras resolved correctly are considered as relevant. The algorithm attempts to resolve them, succeeding in K cases. Due to the coincidence between the volumes of relevant objects and those being resolved, the precision and recall rates are both equal to K/N.

Two out of four rates mentioned above (precision and recall for each step) are informative:

- recall is a portion of pronouns for which the algorithm succeeded in finding an antecedent;
- precision is equal to a percent of this portion containing correctly identified antecedents.

To the writers' opinion, this approach to evaluation conforms to the quality requirements. In addition, the estimations do not depend on the mechanism of anaphora resolution (including the size of variant lists).

5.2 The quality of SVM-method and sensitivity to the sample volume

Opinions containing at least one of the pronouns under research (4 thousand altogether) were selected from the corpus. To evaluate the SVM-method sensitivity to the sample volume this set of opinions underwent the procedure of q-fold cross validation.

Verification was carried out for $q = 1,\ldots,300$, i.e. $q = 1$ means verification of the model for the whole 4 thousand opinions, $q = 300$ — for a sample of 13 opinions. For each q the mean of recall and precision was calculated for each iteration as well as their minimum and maximum for the diagrams reflecting the dependency between quality and the volume of input data.

Measuring was done for modes A, B and C (Fig.4, abscissa corresponds to q).

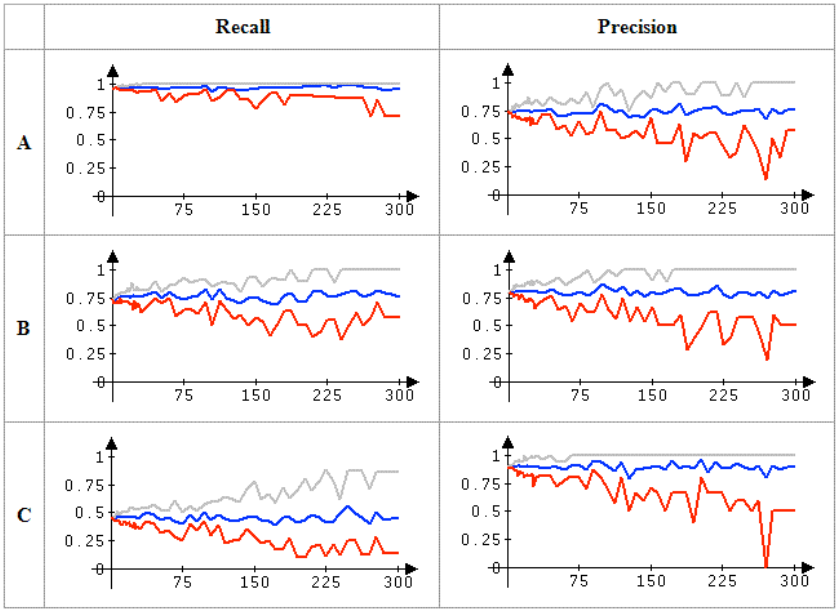


Fig. 4. Results for SVM-method cross-validation in A,B,C modes

It can be seen that all the means are stable even for small-sized samples.

Table 1. Averaged quality measures for SVM-method

	Recall	Precision
A	97.3%	74.2%
B	75.4%	80.7%
C	45.6%	90.3%

5.3 The results of ROC-analysis of SVM-method

Fig.5 illustrates ROC-curves for SVM-method in A, B and C modes.

The area under **A** curve is 0.74, under **B** one — 0.76, which is considered as "good" according to the expert scale. The area under **C** curve is 0.81 with this mode considered as "very good".

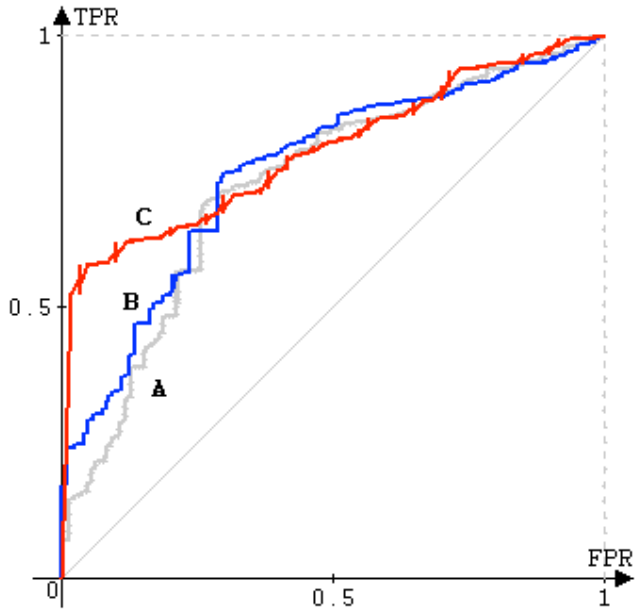


Fig. 5. ROC-curves for SVM-method in A, B, C modes

5.4 The SVM-method independence of the sentiment of the corpus

It was additionally verified in empirical way that the SVM-method is independent of the sentiment of the texts processed, since it cannot be forgotten that anaphoras in negative opinions might be different from those in positive opinions.

The "negative" corpus was used as a training set, the "positive" one as a control set.

Table 2. Check for SVM-method independency from sentiment

(RECALL %, PRECISION %)	**(A)**	**(B)**	**(C)**
Negative *(training)*	(95.1, 80.2)	(77.8, 86.7)	(43.1, 93.2)
Positive *(control)*	(96.3, 78.7)	(79.1, 83.4)	(56.2, 89.9)

5.5 Significance of the factors

Discriminative analysis provides an estimation of contribution of the attributes to the common decision — the judgment can be made based on the coefficients for the corresponding attributes in the linear discriminant and the range of attribute values. It

is also possible to estimate how much influence components of the centroid bring to the solution.

According to the Table 3, the frequency is two times more important than the distance, the presence of a father-verb is more important than the number of nodes in the bush (even if correcting this by a wide range of *NumNodes* — sometimes up to 10-15 knots). Picture according to the centroid is consistent on a whole, except for *IsImp* and *IsVoc* , so their contribution can be estimated to be approximately equal.

Table 3. Valuing the attributes significance according to the results of discriminant analysis

Attribute	*Coefficient in linear discriminant*	*Corresponding coefficient near the component of the centroid*
$IsImp \in \{0,1\}$	– 2.9	18.8
$IsVoc \in \{0,1\}$	9.3	1.1
$HasVerb \in \{0,1\}$	– 7	35.8
$NumNodes \in \mathbb{N} \cup \{0\}$	– 0.5	18.9
$Freq \in \mathbb{N} \cup \{0\}$	– 21.5	-1.6
$Dist \in \mathbb{N}$	– 10.6	0.1

Compiling vocabularies for *IsVoc* is rather laborious. The authors have discovered that the main coefficients in modes A and C (recall and precision respectively) reduce from about 90 to 70% when this attribute is not used; in mode B both coefficients reduce by ~10%. It can be stated that it is precisely *IsVoc* attribute that allows to achieve the precision rate of 90% and higher.

5.6 Evaluation of lexicographical method

The advantage of this method is that no marked-up corpus is needed for its initialization. The practical use of the SVM-method has shown that a trained classifier copes with texts from domains different from that of the training set with the rates declining by several percents (with the exception of *IsVoc* attribute — new vocabularies are needed).

Table 4. Estimation of the lexicographical method quality

	With IsVoc	**Without IsVoc**
(RECALL %, PRECISION %)	(93.7, 51.9)	(93.7, 42.4)

The main error f the method is an excessively strong influence of an attribute with the highest priority. E.g. using *IsVoc* attribute often results in an incorrect choosing a vocabulary word while not using it — in choosing the word closest to the left.

6 Conclusion

This paper offers a solution to the problem of the third personal pronoun anaphora resolution. The software complex called DictaScope Anaphora was implemented based on the models and methods discussed in this paper. It has the following characteristics:

- there are three modes, which allow to achieve both recall and precision rates of 80% or to give preference to one of them and achieve the result of 95%;
- it is possible to take mistypings and grammatical errors into account, which is important for processing texts from online sources (such as reviews);
- in this case an adjustment of the parameters for a specific subject area is needed.

The features of the internal structure of the system and the mathematical foundation are described; the detailed evaluation of the test data and the quality of its processing is carried out.

Among the shortcomings it is a drop in accuracy on the masculine pronouns that should be noted. It is caused by the choice of the subject of opinions (a mobile phone). It is mentioned very often (including implicit mentioning) and the main part of malfunctions consists in choosing an implicit antecedent «*». In authors' opinion, the problem can be solved by taking new attributes connected with the result of syntactical parsing into consideration.

The development plans include the application of the system to other domains and improving the recall-precision ratio by introducing new attributes and refining the adjustment of the coefficients.

7 References

1. Okatev V.V., Gergel V.P., Alexeev V.E., Talanov V.A., Barkalov K.A., Skatov D.S., Erekhinskaya T.N., Kotov A.E., Titova A.S. Report on research implementation on the topic: "Development of a pilot version of syntactical analyzer for the Russian Language", VNTIC Inventory Number 02200803750 // VNTIC, Moscow (2008)
2. Ermakov A.E. Referencing the designations of persons and organizations in Russian media texts: empirical laws for computer analysis. In: Proceedings of the International Conference "Dialog'2005", Computational Linguistics and Intelligent Technologies (2005)
3. Tolpegin P.V., Wind D.P., Kropotov D.A. Algorithm for automated third-person pronouns resolution on the basis of machine learning methods. In: Proceedings of International Conference "Dialog'2006", pp. 504-507. Izd RGGU, Moscow (2006)
4. Oldenderfer M.S., Blashfield R.K. Factor, discriminant and cluster analysis. Under. Ed. Igor Enyukova. Finance and Statistics, Moscow (1989)

5. Platt John C. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. In Advances in Large Margin Classifiers, Alexander J. Smola, Peter Bartlett, Bernhard Sch olkopf, Dale Schuurmans, eds., MIT Press, (1999)
6. Hsuan-Tien Lin, Chih-Jen Lin, Ruby C. Weng, A note on Platt's probabilistic outputs for support vector machines. In: Machine Learning, v.68 n.3, p.267-276 (October 2007)
7. Vapnik V. Statistical Learning Theory. Wiley (1998)
8. Niyu G., Hale J., Charniak E. A statistical approach to anaphora resolution // In; Proceedings of the Sixth Workshop on Very Large Corpora. COLING-ACL'98. Montreal, Canada (1998)
9. Ning Pang, Jun-feng Shi. The third personal pronoun anaphora resolution in the paroxysmal text of the Chinese web. In. Coll. of Appl. Sci., Taiyuan Sci. & Technol. Univ., Taiyuan, China
10. Yıldırım S., Kılıçaslan Y. A machine learning approach to personal pronoun resolution in Turkish. In Proceedings of 20th International FLAIRS Conference, FLAIRS-20. Key West, Florida (2007)
11. Michael P., Kazuhide Y., Eiichiro S. Anaphora analyzing apparatus provided with antecedent candidate rejecting means using candidate rejecting decision tree. Patent US6343266 (2002)
12. Novoteka — news of the day: http://www.novoteka.ru.
13. Yandex.Market — search, selection and purchase of goods: http://market.yandex.ru.
14. CNews Internet-portal: http://zoom.cnews.ru.
15. All for Nokia phones: http://www.allnokia.ru.

An FCA-Based Approach to the Study of Socialization Definitions[1]

Sergey Vinkov

National Research University Higher School of Economics, Social and Economic Systems and Social Policy, Pokrovsky Boulevard, 11, Building "}l{", Room 628, 109028 Moscow, Russia
svinkov@hse.ru

Abstract. The most typical definitions of socialization used in Russian academic and educational literature are considered in this article. To make a typological analysis of definitions and construct their taxonomy, a mathematical method of formal concepts analysis is applied.

Key words: formal concept analysis, typological analysis, socialization, type

1. Problem statement

Any scientific discipline has a special conceptual and categorical apparatus. Social sciences and humanities, as well as natural sciences, have their own narrow lexical units. Unlike branches of the exact sciences, social and humanitarian sciences frequently have some ambiguity, or lack a uniform definition for one or another concept, despite its wide and frequent use[2].

«Socialization» is a base category in a number of disciplinary areas: philosophy, sociology, psychology, pedagogics. It is obvious that there is a general understanding of this process, but a conventional definition is not avaliable.

The retrospective analysis of defining "socialization" shows that initially this concept was used in the common sense in the middle of the previous century by J. Dollard in the context of researching «social learning» and formation of an individual [1, p.14]. The etymology of the concept “socialization” leads to the German language when two words «Sozialisierung» and «Vergesellschaftung» were borrowed by the Anglo-Saxon language system for the description of absolutely new social phenomena and processes [1, p.12-13]:

1) «Sozialisierung» – transition of private property to public one (or state one);

[1] This paper is a translation of the Russian version published in Journal of Orel University

[2] For example, cf. Discussion about social issues. Seminars in the State Educational Establishment – High School of Economics/Polit.ru, URL: http://www.polit.ru/dossie/2010/03/23/social_2.html (reference date: 12.03.2011).

2) «Vergesellschaftung» – as «cooperation of persons in a mental unity of group life» and as «the central process in social evolution»[3].

V.G.Nikolaev underlines that «we have transition of the borrowed from German"tradition"(particularly from Zimmel) concept of "nationalization" as a dynamic process of making a society, formation by a society, transformation of independent individuals into a society by means of interaction, to the concept of "socialization" focused on« developmental aspects of individual behavior, revealed in interaction — or directly integrated into the interaction — with one or more persons, i.e. in a social context »[1, p.15].

The "socialization" polysemy is supplemented by other contexts of scientific literature, different from the ones mentioned above, pertaining, for example, to «socialization of needs» [2, p.24], gender socialization, emotional competence, cognition etc. [3, pp.561-691]. This usage of the word "socialization" emphasizes its historical nature, shows the formation trajectory, time perspective of the subject. But along with that, the subjects/objects are always connected and correlate with a specific person, whether a single representative of the mankind or of a social group or of a society as a whole.

This research is further devoted to considering the definitions of socialization in view of certain subject-object and subject-subject interactions where, on the one hand, a human (an individual, a person) will act as an object of social environment or public group efforts, and on the other hand – as an independent creator of the course of life who forms the social reality.

The problem of concept formation in social and humanitarian sciences, and its subsequent ways of application, pointed out the purpose of the research which is to make a typology of some definitions of "socialization", to detect similarities and differences in various descriptions of this social phenomenon.

In this work, logical outlines of the concept "socialization" are found out and proved by means of formal concept analysis. Certainly, it is also necessary to take into consideration the bases of the social phenomenon that are not logical. [4, p.14].

2. Methodology

The typological analysis is a meta-technique of making socially significant, internally homogeneous, qualitatively different from each other groups of empirical objects [5, p.75].

The research practice used in this work implies analyzing semistructured data made into a set of definitions for the concept "socialization". The general strategy of such analysis is ascending and is based on constructing a pyramid out of all generalized answers to the question: «What is socialization?»

With the support on this meta-technique, the research potential of a quite young scientific technique and method is used. This method is called formal concept analysis

[3] Ref. [1, p.13].

and is widely developed and actively used in Russia and abroad[4]. Preconditions to the appearance of formal concept analysis date back to the middle of the XX century to the works on the theory of lattices; the French works on Galois lattices became a major landmark in its formation. Ultimately, formal concept analysis became a scientific method due to the works of the German scientist and mathematician Rudolf Wille. On the whole, it is possible to tell that formal concept analysis is a mathematical method of researching construction of subject domain taxonomies [11]. It opens wide possibilities for studying family, professional interactions, structures of values, social networks and so on [12].

This research includes three stages:

1st stage. Primary gathering of text matter[5]. For the purpose of forming the matter it was necessary to select one sentence in Russian as a definition of socialization published in hard-copy form. Both national and foreign (translated into Russian) literature was included into the research: textbooks and manuals, monographs, the dictionaries which help fully understand the essence and the content of the concept "socialization". Sampling was random because of the difficulty to establish the whole scope of all possible definitions in Russian and the limited availability of all the editions that contain such definitions. Selecting definitions from books defining "socialization" continued till there was a significant difference between one definition and the one considered before. Only one was selected from the wordings having the same number of words and phrases.

A table with three columns is made up to record the definitions: the first is the number of a definition, the second has the definition itself, the third contains a full bibliographic description of a source, from where the definition was taken. The text matter has been formed by means of Excel Microsoft Office 2003.

2nd stage. Creation of a formal context. A binary matrix presented in a tabular format (Table 1)[6] is generated from the text matter received at the preliminary stage.

The rows of the table contain some objects – the set of definitions, and the columns contain meanings of the characteristics derived from each object. The presence of a characteristic in a matrix is designated as 1 and its absence as 0. For the subsequent analytical work, the final element of a column is an indicating figure which is the total number of characteristics in an object, and the final element of a row is the frequency of a characteristic occurrence, calculated as follows: the number of objects with the characteristic divided by the total number of objects.

---

[4] Theoretical basis and ways of application of the method are to be found in the following works [6, 7, 8, 9, 10].

[5] All the materials serving as the basis for this research are available on the author's personal webpage: http://www.hse.ru/org/persons/12435171 (in Russian).

[6] Sources: 1 – Kulikov L.M. Basic sociology and politology: manual. – Finances and statistics, 2008. – p. 336 – p.117.; 2 – Lapin N.I. General Sociology: manual for universities / Lapin N.I. – 2-nd rev., enlarged. – M.:High School., 2009. – p. 452: pic. – p.59.; 3 – Masionis J. Sociology. – 9-th ed. – St.P.: Piter, 2004. – p. 752: pic. – p.170; 4 – Zborovsky G.E., Orlov G.P. Sociology. Textbook for humanitarian establishments of higher education. Inter-praks. 1995p. – 320p. – p.309, 289.; 5 – Dobrenkov V.I., Kravchenko A.I. Sociology: textbook. – M.: INFRA-M, 2009. – 624p. – p.575.; 6 – General sociology: Manual/ edited by prof. A.G. Efendiev. – M.: INFRA-M, 2005. – 654p. – p.492.

In « the area of social methodology one of the fundamental scientific problems is polysemy of interpretations and contextual conditionality of articulated theoretical propositions» [13, p.3] which complicates deriving characteristics from a formal context, therefore it is necessary to explain this procedure.

The basic principle of deriving characteristics from definitions was dividing definitions into 2 groups according to the content.

The first group designated the object of a subject's activity, and the second group described a type of a subject's activity, directed to an object or procedurally describing the condition of the subject. The first group of characteristics answers the question: «What is the ultimate goal of the activity, the continuous action contained in a definition? What is the subject of the activity?» And the second group answers the question: «What does the individual do or what happens to one?»

As a rule, characteristics were derived as soon as mentioned by the author and were not made excessive. Only some of the characteristics were united into one: I have included the following phrases into the column "Behaviour": «examples of behaviour», «behaviour characteristics», «behaviour patterns», because the author made frequent references to the behavioural aspects of an individual. When the author used the phrase «behaviour rules» when defining "socialization", it was included in the column "Rules".

The problem of accurate demarcation of values and value preferences caused these words to be ascribed to one characteristic of "Values".

Synonymic words "training", "studying", «learning» were also grouped in one column called "Education".

"Adaptation" and "adjustment", "reproduction" and "propagation", "mastering" and "assimilation" are united due to their explicit homogeneous semantics.

The logic of selection can be clarified by the example of the definition number 3 from table 1.

1. Fixing an activity type. It is necessary to remark that the definition is worded in an extremely non-standard way, and certain subjectivity in this procedure is inevitable. However, asking a question: «What happens to an individual?» and answering it, it is easy to deduce "development" and "assimilation".

2. Definition of the subject of the activity. And again a question: what are "development" and "assimilation" directed to? The answer is: «human potential» and "culture".

Appreciable similarity can be found between objects № 3 and 6. In both rows there is some indirectness, "instrument" (comes from the phrase "due to"), but the latter, when expanding on the content, lacks the "activity-object" link, therefore the second part of the definition was not considered. In most of the objects (definitions) this link is obvious and quite frequent.

The formal context, in form of a binary matrix, is mathematically described by three variables: K: = (G, M, I), where G is a set of objects (some total number of definitions); M — a set of characteristics, and the incidence relation $I \subseteq G \times M$ shows which objects possess which characteristics. For the arbitrary, $A \subseteq G$ and $B \subseteq M$ the Galois operators are deduced:

$$A' = \{m \in M \mid \forall\, g \in A\ (g\ I\ m)\};$$
$$B' = \{g \in G \mid \forall\, m \in B\ (g\ I\ m)\}. \quad (\mathbf{1})$$

Table 1. An example of a formal context

		Object of activity										**Activity type**							
	Object , G / Characteristic , M	Values and values preferences	Norms	Culture	Knowledge, system of knowledge	Skills	Human potential	Social qualities, features	Roles	Human potential	Guidelines	Behaviour	Introduction	Training, learning, studying	Mastering, assimilation	Formulation	Internalization	Development	**Total number of characteristics**
1.	Socialization is a process of how an individual is introduced into a society.	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1
2.	Socialization is training and assimilation, internalization by a person of values, norms, guidelines, rules of behavior characteristic of a society, a social group or a community.	1	1	0	0	0	0	0	0	0	1	1	0	1	1	0	1	0	7
3.	Socialization is understood as the social experience gained throughout life due to which individuals develop their human potential and assimilate culture.	0	0	1	0	0	1	0	0	1	0	0	0	0	1	0	0	1	5
4.	Socialization is a process where a person assimilates examples of behavior in a society and a group, their values, norms, guidelines.	1	1	0	0	0	0	0	0	0	1	1	0	0	1	0	0	0	5

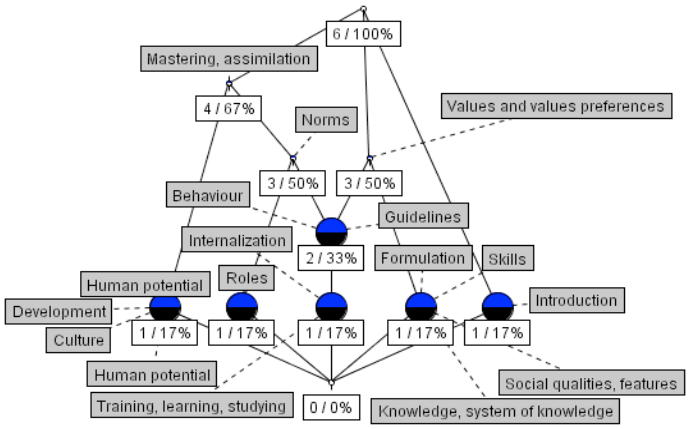
5.	Socialization is a lifelong process (from infancy till the old age) of assimilating cultural norms and mastering social roles.	0	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	3
6.	Socialization is formation of social qualities, properties, values, knowledge and abilities due to which a person becomes a capable participant in social communications, institutions and communities.	1	0	0	1	1	0	1	0	0	0	0	0	0	0	1	0	0	5
	Characteristic occurrence frequency	0,059	0,059	0,020	0,020	0,020	0,020	0,020	0,020	0,020	0,020	0,039	0,039	0,020	0,020	0,078	0,020	0,020	0,020

In that case, the pair (A, B), obeying conditions: $A \subseteq G$, $B \subseteq M$, $A' = G$, $B' = A$, is called a formal concept of the context K = (G, M, I). The set of objects A constitutes the concept scope, and the set of all characteristics B which the objects possess constitutes the concept content. Each object $a \in A$ possesses all characteristics from the subset B. Consequently the formal concept corresponds to a set of objects from a specific area, which possesses all the characteristics from some subset. The set of all formal concepts of a context are partially ordered by the relation «more than …». A poset of concepts form a lattice. The concept lattice is used to visualize formal concepts.

3rd stage. Construction of a concept lattice. So-called Hasse diagrams are used to visualize a lattice. In the diagrams two adjacent concepts (i.e. there are no other concepts in-between them) are connected by a link, and a more general concept always lies above a less general one. In this research the visualization is executed by means of freely available software Concept Explorer 1.3[7].

[7] http://conexp.sourceforge.net/.

Picture 1. Diagram of a concept lattice for 6 definitions



A formal concepts lattice (Picture 1) allows to structurally classify definitions of socialization and to find the most similar ones by means of the derived characteristics.

Let us start reading the diagram of a concept lattice from the top. In the diagram based on Table 1, the top formal concept makes it clear that there is a lack of at least one characteristic referring to the whole set (not a single object includes all characteristics). Lower there is a formal concept uniting 4 objects that have "Assimilation" in them. One of these objects differs from the other three because apart from "Assimilation" it has "Development", "Culture" and “Human potential” and thus forms a separate formal concept at another level of the diagram. It is obvious from the diagram that there is a formal concept, different from all the rest except the top and the bottom ones; it is the definition (object) containing the characteristic of "Introduction". The formal concept lattice allows deducing the greatest number of objects constituting the concept as well as the total set of characteristics.

In a formal context the connection A → B is implied, if all objects with the set of characteristics A also have the set of characteristics B. Thus in our case it was concluded that:

1. All definitions with the characteristic of "Norms”, have the characteristic of "Assimilation";

2. The characteristics of "Value" and "Assimilation" are included in definitions only when there are "Norms" in them.

The advantage of concept lattices is optimization (simplification) of research procedures on deducing connections i.e. essential interrelations of objects and characteristics, whereas the direct search of such connections and interrelations is quite inconvenient even with rather small, as in the investigated case, data array.

3. Results

1. The first investigation stage resulted in a tabular data array of 51 definitions of "socialization" taken from 50 sources. Chronologically, the literature which got into the research field, ranges from 1994 to 2010. The sampling did not exceed the number of 51 definitions, because on the one hand, new definitions did not have any new characteristics, and on the other hand, the available scope allowed searching for the most typical definitions.

The array included the definitions of the socialization developing the phenomenon through the following:

- The subject-object approach where an individual has the role of an object in the course of interaction of the society with the social environment, which is close to G.Tarde's imitation theories, W.James' socialization theory, interactional socialization theories, M.Weber's theory of socialization and the ones by P.Sorokin, R.Merton;
- The subject-subject approach represented in the theories of «symbolical interactionism», «looking-glass self», works by W.Thomas and F.Znanetsky who considered an individual not just as a passive agent, but also as an active participant of his own formation;
- the inter-subject approach describing socialization solely as introduction or interaction of an individual with the social environment (a social group or a society); that explains the objects (definitions) with one characteristic ("Interaction", "Introduction", "Transmission", "Influence") at the further research stage.

2. Transition from semistructured data to the structured ones opens additional possibilities to search for the regularities, typical definitions, not on the basis of the verbal content of a definition, but on its characteristic description in the formal context represented in a binary matrix[8].

As a result of deducing characteristics the object-characteristic table (a formal context) contains 51 definition (object) for each of which there is one or more characteristic out of 61, frequency of a characteristic occurrence and the total number of characteristics for each object are calculated.

Frequency of a characteristic occurrence led to 38 less informative characteristics which were then removed from the formal context, and thus only 23 most informative characteristics have been left. Less informative characteristics were considered to be those with less than 0.1 and more than 0.9 occurrence frequency ratio. Such method is often called entropic, as entropy H (measure of uncertainty) of a random object with frequencies $p_1, ..., p_M$ equals to

$$H = p_1 \log(1/p_1) + ... + p_M \log(1/p_M) \quad (2)$$

[8] The binary matrix as well as Hasse diagram of concept lattices, constructed on basis of the matrix, can be found in xls-format on the author's personal webpage: http://www.hse.ru/org/persons/12435171 (in Russian).

and hence the rejection of «margin» outcomes (i.e. either hardly occurring, or occurring almost always) only insignificantly changes the amount of the information constituting an object [14, p.236].

All objects have been grouped according to their "weight" from 0 to 8, i.e. how often the informative characteristics were used, and within each weight lexicographic ordering is used.

An object with a zero weight corresponds to the following definition:

«Socialization is an inseparable element of constant everyday process of interaction of an individual with the surrounding social environment.» [15, p.49]. All the objects with a zero weight, having no similarities to the other objects, have been excluded from the further work.

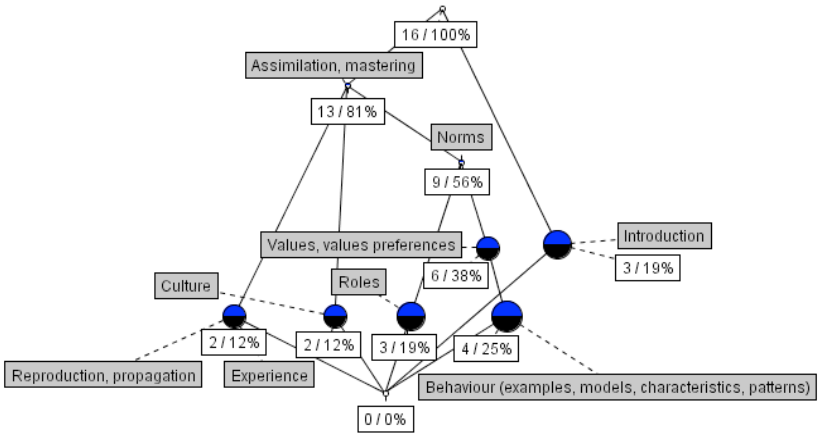
As a result of «matching», i.e. coincidence of two or more most informative characteristics in definitions, 6 groups of 16 objects have been formed with the weight ranging from 1 to 4, different in wording, but identical in the chosen informative characteristics (Table 2). It is natural to consider such definitions as the most typical.

Table 2. Characteristics of typical definitions

Group	Number of objects	Weight	Characteristic
I.	3	1	«Introduction»
II.	2	2	«Assimilation», «Culture»
III.	2	3	«Assimilation», «Norms», «Values»
IV.	3	3	«Assimilation», «Roles», «Norms»
V.	2	3	«Reproduction», «Assimilation», «Experience»
VI.	4	4	«Assimilation», «Behaviour», «Norms», «Values»

Thus, the informative quality of the characteristics "Culture", "Experience", "Reproduction" has gone down, which obviously indicates of the existence of a certain kernel in socialization definitions which includes: "Introduction", "Assimilation", "Behaviour", "Norms", "Values".

Picture 2. Diagram of concept lattice for 16 definitions



It follows from the concept lattice diagram that there is only one formal concept uniting 3 objects and independent of any group of definitions characterized by "Introduction". For similar definitions, the interconnection "who-where"- «process of an individual's introduction into a society» is typical. [16, p.117]. All the other formal concepts include all the definitions united by the characteristic "Assimilation". The further distinctions between socialization definitions are in a set of certain characteristics. The greatest number of characteristics was formed in the formal concept with the characteristic "Behaviour"; one of them defined socialization as «a process of assimilation by an individual of examples of behavior, social norms and values necessary for a successful functioning in the society» [17, p.102].

4. Conclusion

This work deals with the problem of creating a typology for various definitions of "socialization" in order to find similarities and differences in ways to describe the concept. The applied method of formal concept analysis combined with the subsequent content analysis has allowed deriving the most typical definitions of "socialization" and the most informative characteristics used in the analyzed definitions of "socialization". Undoubtedly, it is interesting how the methods and results of this research will influence the definitions of "socialization" which exist in the English literature.

References

1. Nikolaev V.G. Sociological concepts in social and cultural contexts: experience of reflective knowledge sociology / Full texts of selected publications of teachers, department of the general sociology of the Research Institute, High School of Economics, URL: http://soc.hse.ru/gsoc/full_text. (Reference date: 3/12/2011). (in Russian)
2. Zdravomyslov A.G.Needs. Interests. Values/A.G.Zdravomyslov M: Politizdat, 1986 – 221p.
3. Joan E. Grusec and Paul D. Hastings (eds), Handbook of Socialization: Theory and Research. New York/London: The Guilford Press, 2007. 737p. (in Russian)
4. Kachanov Y.L. Polyparadigm approach, logic and sociological concepts//Sociological researches. 2010. No. 8. p. 12-19. (in Russian)
5. Tatarova G.G.Basis of the typological analysis in sociological researches: the manual / G.G.Tatarova. – M: publishing house «Novy Uchebnik», 2004. – 206p. (in Russian)
6. Kuznetsov S.O., About some questions of concept analysis, STI. Ser.2. 1999. No. 1-2, pp. 57-61. (in Russian)
7. Kedrov S.A., Kuznetsov S.O., Research of groups of Internet users by methods of formal concept analysis and Data Mining//Business Informatics. 2007. No.1, pp. 45-51. (in Russian)
8. Emelyanov G. M, Mihaylov D.V. Semantic clustering of textual and subject languages (morphology and syntax)//Computer optics. 2009. Vol.33, No. 4. pp. 473-480. (in Russian)
9. du Boucher-Ryan, P., Bridge, D., Collaborative Recommending using Formal Concept Analysis//Knowledge-Based Systems. 2006. Vol. 19, No.5. pp. 309-315.
10. Cimiano, P., Hotho, A., Staab, S., Learning concept hierarchies from text corpora using formal concept analysis//Journal of Artificial Intelligence Research, 2005. Vol. 24, No.1. pp. 305-339.
11. Wille, Rudolf, Concept lattices and conceptual knowledge systems//Computers and Mathematics with Applications, March 1992, Vol.23, Issue 6-9, pp.493-515.
12. Martin J.L. Jointness and Duality in Algebraic Approaches to Dichotomous Data//Sociological Methods and Research 2006 Vol. 35, No.2. pp. 159 – 192.
13. Kanygin G.B. Contextually-focused analysis of qualitative data//author's abstract. Doctor of sociology, St.P., 2001. 32p. (in Russian)
14. Yaglom A.M., Yaglom I.M. Probability and information. vol. 5, stereotypic. – M: KomKniga, 2007. 512p. (in Russian)
15. Houston M. Introduction to social psychology. The European approach: Textbook for university students / M. Houston, V.Schtroebe; translation from English edited by prof. T.J.Bazarov; [translated by G.J.Ljubimova]. – M.: Yuniti-Dana, 2004. – 622 p. (in Russian)
16. Kulikov L.M. Basic sociology and political science: manual. – Finance and statistics, 2008. – 336 p. (in Russian)
17. Radugin A.A. Sociology. M.Tsentr, 1996. – 206 p. (in Russian)

Temporal Concept Analysis Explained by Examples

Karl Erich Wolff

Mathematics and Science Faculty
Darmstadt University of Applied Sciences
Holzhofallee 38, D-64295 Darmstadt, Germany
karl.erich.wolff@t-online.de

Abstract. The purpose of this paper is to show by examples the advantages of Temporal Concept Analysis (TCA) - the theory of temporal phenomena described with tools of Formal Concept Analysis (FCA). TCA is developed in three main branches: first the branch of Conceptual Time Systems with actual Objects and a Time relation (CTSOTs) where each temporal object is at each time granule at exactly one place; that gives rise to a first notion of states, transitions and life tracks of an object. The second branch of TCA is centered around the notion of a Temporal Conceptual Semantic System (TCSS) which allows to introduce the notion of a distributed object which may occupy at each time granule a certain volume, called its trace. That leads to a clear mathematical distinction of the notions of particles and waves in physics. The third branch of TCA is based on the notion of a Temporal Relational Semantic System (TRSS); it uses the developments in Temporal Conceptual Semantic Systems for combining the conceptual graphs by J. Sowa and the concept graphs by R. Wille with conceptual scaling.

1 Introduction to Temporal Concept Analysis

Temporal Concept Analysis (TCA) is the theory of temporal phenomena described with tools of Formal Concept Analysis (FCA). While FCA was introduced by R. Wille [9] in 1982, TCA was introduced by the author [11] in 2000 and further developed since then. In the following we assume that the reader is familiar with the basic notions in FCA as explained in [2].

One of the leading ideas in TCA was to represent the notion of a *state* of an object at a certain time in a temporal system. For that purpose a suitable notion of a temporal system including a formal representation of time has to be chosen in such a way that the notion of an object and the notion of a state of an object at a certain time granule (like 'a minute' or 'a day') can be introduced in a natural way.

Looking for a suitable notion of a temporal system it was clear from the beginning of the development of TCA that a temporal system should be described as a data table (interpreted as a result of an observation) as opposed to a system description based on rules. To use the powerful knowledge representation

by means of (nested) line diagrams of concept lattices the observations should finally be represented by a formal context. For that purpose the transformation of a data table with arbitrary values (mathematically described as a *many-valued context* (G, M, W, I)) into a *formal context* (G, N, J) is necessary; this transformation is called *conceptual scaling* and can be done in a meaningful way by representing the values of a *many-valued attribute* $m \in M$ as formal concepts of a suitable formal context $S_m := (G_m, M_m, I_m)$ which is called a *conceptual scale* of m, if G_m contains all values of m. The derived context $\mathbb{K} = (G, N, J)$ of a many-valued context (G, M, W, I) with respect to a family $S_m := (G_m, M_m, I_m)$ $(m \in M)$ is defined by $N := \{(m, n) \mid m \in M, n \in M_m\}$ and $gJ(m, n) \iff m(g)I_m n$ for $g \in G$ and $(m, n) \in N$. Any subset $Q \subseteq N$ is called a *view*. The subcontext $\mathbb{K}_Q := (G, Q, J \cap (G \times Q))$ is called the $Q-$part of the derived context. In the following, the concept lattice of a suitable $Q-$part will play the role of a map into which relevant structures like traces of objects, transitions, and life tracks will be embedded (see Fig.1,2,3,6).

The three main branches of TCA are described in the following three sections just by their leading ideas and some examples. Hints to the mathematical definitions in TCA will be given in these sections.

2 Conceptual Time Systems

The first intuitive idea about the notion of a *state*, which was the beginning of TCA, came suddenly to my mind when I was standing alone under the bright sun of Crete on the ruins of the ancient palace of Minos in Knossos, and I vocalized:

The states are just the object concepts of suitable contexts.

That happened in May 1993 just after a conference in Chania (Crete) where I had many fruitful discussions with R.E. Kalman [3] on general systems. In my first paper [11] on temporal conceptual systems the notion of a Conceptual Time System (CTS) and the notion of a state of a CTS was introduced. A simple example of such a temporal system is given in the following subsection.

2.1 A chemical process in a distillation column

In cooperation with a chemical engineer we investigated the temporal behavior of a distillation column. At each of 20 days 13 variables had been measured once. For 4 of these 13 variables the corresponding data table is indicated in Tab.1 by the measurement values at the first and the last day. It is clear that there is a simple notion of a state of this distillation column at some day, namely the tuple of all the measurement values observed at this day, for example the state of the distillation column at day 1 is the quadruple $(129, 616, 616, 119)$ (according to Table 1). As usual, the experts wish to talk a little bit coarser, for example about low, middle and high values of some variables. In close cooperation with the chemical engineers we developed conceptual scales for all of the variables. For reflux and energy1 the resulting state space together with the transitions of

the distallation column is shown in Fig.1, which is called a *transition diagram* for this distillation process with respect to the chosen granularity.

Table 1: Data table of a Distillation Column

time granule	day	reflux	energy1	input	pressure
1	1	129	616	616	119
...	...	...	...	...	...
20	20	127	556	664	120

It should be intuitively clear what Fig.1 represents. We do not repeat here the mathematical definitions of a Conceptual Time System as introduced in [11]; we just describe it here in a data table language: the values in the first column are denoted as time granules and interpreted as granules of time as for example a minute or a day; the set of the other columns is divided in two parts, the *time part*, which has in Table 1 just a single column, namely the column of *day*, while the other columns form the *event part*, which consists of the 4 columns of *reflux, energy1, input, pressure*. In this example, the time granules just denote the days, in general a time granule is described by its values in the time part, which might have several many-valued attributes, for example *day*, *month*, *year*.

We now focus on the *reflux-energy1-part* of the data table and show the derived context of this part in Tab.2. For example, at time granule 1 the distillation column has all three reflux-attributes in Tab.2, while it has only the first two energy1-attributes, but not the last, since the energy1-value 616 is not $<= 570$. The two attributes at the top-concept of Fig.1 have been introduced to tell the reader of the diagram the range of observed values for reflux and energy1.

Table 2: The derived context for reflux and energy1 each scaled with an ordinal scale of a 3-chain

time granule	reflux 126-183	reflux<= 140	reflux<= 133	energy1 514-693	energy1<= 660	energy1<= 570
1	×	×	×	×	×	
...	...	...	...	...	...	...
20	×	×	×	×	×	×

Now we discuss the role of the object concepts in a CTS. If g is a formal object (=time granule), then the intent of the object concept $\gamma(g)$ is the set of all attributes which g has in the derived context - and this intent of $\gamma(g)$ is a very nice description of the state of the CTS at the time granule g with respect to the chosen granularity. It is also good to know all time granules which have the attributes of g; they form the extent of $\gamma(g)$. Therefore, the *state* of a CTS

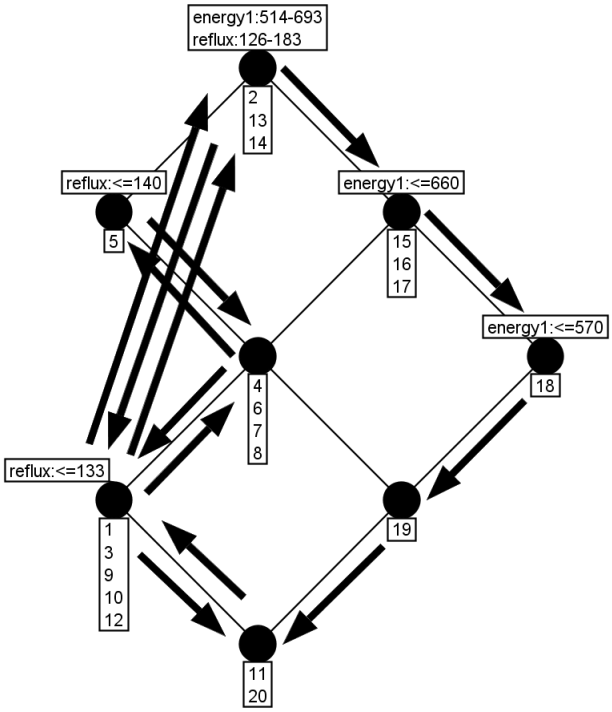


Fig. 1: A chemical process represented in a transition diagram

at time granule g is defined as the object concept $\gamma(g)$ in the derived context of the CTS (with respect to the chosen scales). That approves my intuitive state idea which I had at Knossos.

For example, state $\gamma(1)$ has the intent consisting of all attributes in Table 2 but the last, namely "energy1 <= 570". The extent of $\gamma(1)$ is the set $\{1, 3, 9, 10, 11, 12, 20\}$. The extent of an object concept $\gamma(g)$ should clearly be distinguished from its *contingent* which is defined as the set of all objects h such that $\gamma(h) = \gamma(g)$.The contingent of an object concept $\gamma(g)$ is shown in Fig.1 just below the concept node of $\gamma(g)$. For example, the contingent of $\gamma(1)$ is the set $\{1, 3, 9, 10, 12\}$.

We now focus on the arrows in Fig.1; they represent *transitions*. We have drawn an arrow from $\gamma(g)$ to $\gamma(g+1)$ to denote the transition from day g to day $g + 1$ for $1 \leq g \leq 19$. But we have to explain how the transitions are defined in general. That was introduced by the author in [13]. It is clear from the example in Fig.1 that transitions should not be defined as pairs of states (as it is done in Automata Theory), since the two arrows from state $\gamma(1)$ to state $\gamma(2)$ denote different transitions, one happens during the time step from the first day to the second day, for short: $1 \rightarrow 2$, the other one during $12 \rightarrow 13$. Therefore, we describe the first transition by the pair of pairs $((1, 2), (\gamma(1), \gamma(2)))$ and in

the same way the others. For the general definition of a transition we extend the structure of a CTS by a given binary relation R on the set G of formal objects, since we interpreted the formal objects as time granules. (Later on, in Temporal Conceptual Semantic Systems we will use the concepts of the time scale as time granules.) The relation R is called the *time relation*. In nearly all practical applications with discrete time the time relation is the set of pairs (t, t+1) for $t \in \{0, ..., n-1\}$. Since we do not like to assume any linearity of time, we just assume that R is a binary relation on the set G of time granules and interpret R as the set of observed time steps. Hence for a CTS with a time relation R any transition is of the form $((g,h),(\gamma(g),\gamma(h)))$ where $(g,h) \in R$. For more details the reader is referred to [13].

In Fig.1 we can follow the *life track* of the CTS with respect to the chosen view, if we start in state $\gamma(1)$, go to state $\gamma(2)$, and so on until we reach $\gamma(20)$.

Fig.2 shows a *nested transition diagram* where the outer diagram represents the two variables *reflux* and *energy1* (as in Fig.1), while the inner diagram represents the two variables *input* and *pressure* in a 5x4-grid.

From such diagrams the process can be understood quite well. A short and coarse description of that process might be:

During the first two weeks the distillation column was mainly in states with low input and middle pressure - with the exceptions of day 10 (high pressure) and day 11 (low pressure) - while it moved during the last week to high input, small energy1, small pressure, and during the last three days also to small reflux.

Typically, in such applications the experts suggest first a coarse granularity by a few "cuts" like "reflux$\leq$140". After having studied the concept lattice with a coarse granularity it is usually refined, depending on the data and on the interest of the experts. That leads in a few steps to valuable visualizations of multidimensional processes.

The following example of the family of an anorectic young woman served as a motivation for me to develop temporal conceptual systems in which many objects may move, each with its own life track. Hence we need a more general notion of the *state of an object of a system* as opposed to the *state of a system* as in the previous example.

2.2 The development of the family of an anorectic young woman

The following example in Fig.3 describes the development of an anorectic young woman (SELF), her father, mother, and her self ideal (IDEAL) during a period of about two years. The underlying formal context was constructed by the psychoanalyst N. Spangenberg [6–8] on the basis of four repertory grid questionings taken about each half year from the beginning (time granule 1) until the end (time granule 4) of the psychoanalytic treatment of his patient. SELF1, the self at the beginning of the treatment, has the attributes "distrust", "reduced spontaneity", "pessimistic" and "self-accusation", SELF2 has only the attributes "pessimistic" and "self-accusation", SELF3 is in the same state as SELF1, and SELF4 reaches the state of IDEAL2,3,4 having none of the (negative) attributes. Indeed, the patient was healthy again at this time. It is remarkable that the life

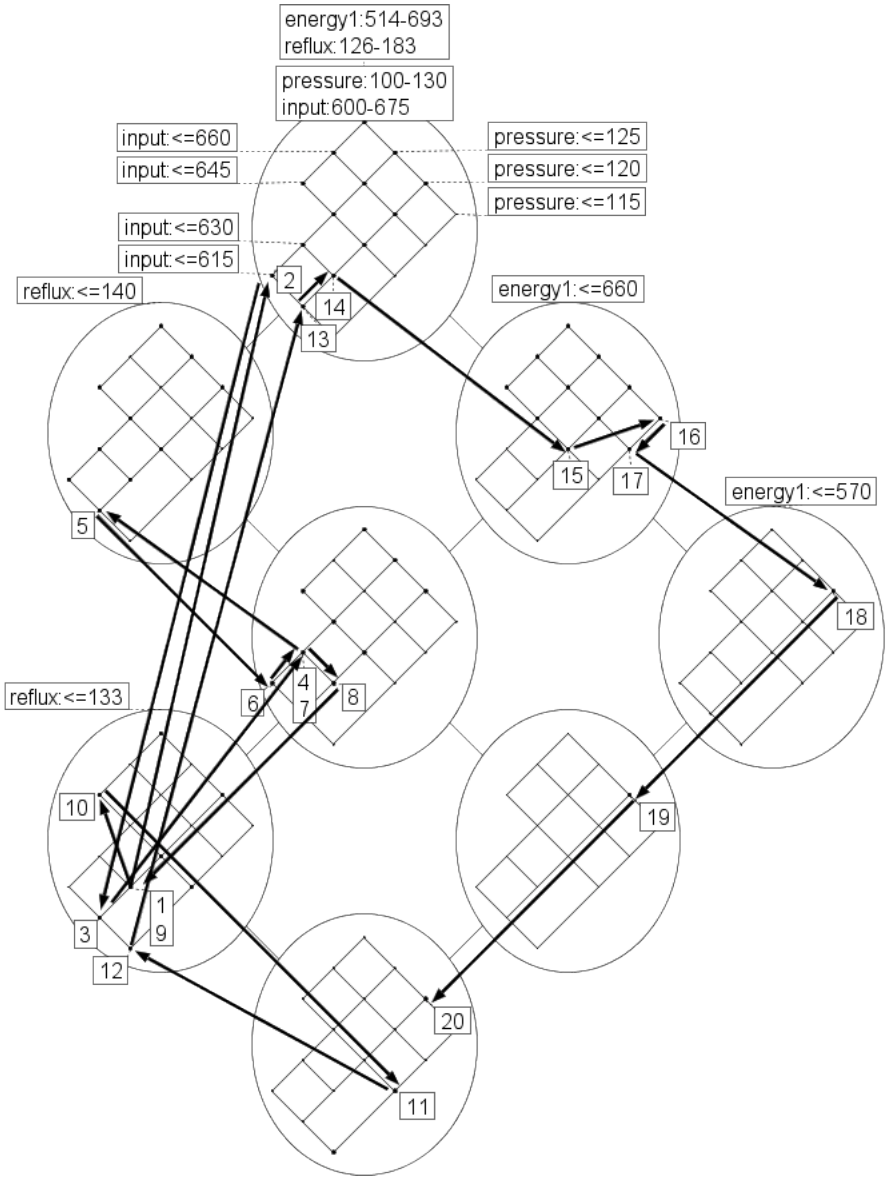


Fig. 2: A chemical process represented in a nested transition diagram

tracks of FATHER and MOTHER start from quite different states and end in similar states, the FATHER having all (negative) attributes of this context. For further information the reader is referred to [6–8].

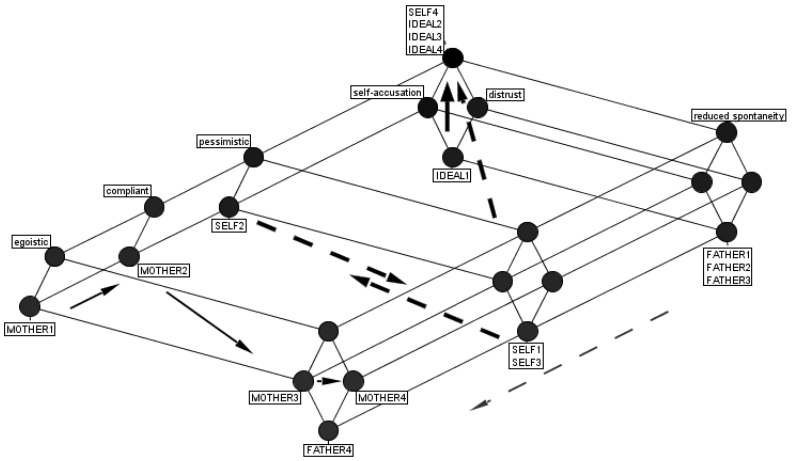


Fig. 3: The development of an anorectic young woman and her family over about two years

2.3 Conceptual Time Systems with actual Objects and a Time relation (CTSOTs)

To have a simple mathematical notation for temporal systems in which many objects (for example persons) are moving the author has introduced in [14] the notion of a Conceptual Time System with actual Objects and a Time relation (CTSOT). The main idea is a straightforward generalization of a CTS by introducing a set P of *persons* or *particles* and a set G of *time granules*, and taking a subset of $P \times G$ as the set of formal objects. For example, the pair (SELF,1) is taken as a formal object denoting the person SELF at time granule 1. Then the notion of a state of a person p at a time granule g can be introduced as the object concept of the formal object (p, g). Transitions and life tracks can be defined in the same way as explained previously [13–16].

While the CTSOTs cover the wide range from particle systems in physics to discrete systems in Computer Science they are not general enough to cover

also waves and other distributed objects and their states as for example the distributed state of an electron as represented by some cloud around the center of the electron. Such distributed objects also occur very often in workaday life, for example as a moving high pressure zone on a weather map. Such distributed objects, including waves, can be represented in Temporal Conceptual Semantic Systems which will be explained in the next section.

3 Temporal Conceptual Semantic Systems (TCSSs)

As opposed to CTSOTs where each object p is at each time granule g at exactly one place, namely at the object concept $\gamma(p, g)$ (where γ is the object concept mapping of the chosen view) we now wish to represent also objects which are 'distributed'. For that purpose, we should use neither the time granules (as in CTSs) nor the actual objects (p, g) (as in CTSOTs) as formal objects of the many-valued context of the temporal system. Instead, we take the row numbers of the data table as the formal objects; they can be interpreted as names of the *statements* represented by the information given in that row.

3.1 Basic Notions in TCSSs

Conceptual Semantic Systems have been introduced by the author in [15] for the purpose to represent distributed objects, like waves, in a natural way. The main idea was to represent the concepts in an application domain as formal concepts of formal contexts, called *semantic scales*. Since statements are often shortly described by a tuple of concepts of some application domain we represent each such statement as a tuple of formal concepts. These formal concepts (usually represented by their names) are then chosen as values in a many-valued context in such a way that each row, labeled by g, represents a statement which is denoted by the tuple $(m(g))_{m\in M}$ where $m(g)$ is a formal concept of the semantic scale of m. Since the formal objects of the many-valued context of a CSS are interpreted now as names of statements and not as time granules as in CTSs we need a new representation of time granules in TCSSs. As formal representations of time granules we take the formal concepts of the time scale of a specified many-valued time attribute. The time relation is defined as a binary relation on the set of formal concepts of the time scale. Objects in a temporal system are mathematically represented in TCSSs as tuples of concepts of the semantic scales, as for example the tuple $(\mathbf{High}, \mathbf{Monday})$ denoting a high pressure zone at Monday in the next example. Such an object (which is not a formal object) may be *distributed* in the sense, that it has been observed at more than one places (=object concepts in the concept lattice of the $Q-$part $\mathbb{K}_Q$ of the chosen view Q). This set of object concepts (of the formal objects denoting row labels or statements) where an object has been observed is mathematically defined as the *trace* of an object in $\mathbb{K}_Q$. The *state* of an object $\mathbf{o}$ at a time granule $\mathbf{t}$ is then defined as the trace of the tuple $(\mathbf{o}, \mathbf{t})$. *Transitions* of an object $\mathbf{o}$ from time granule $\mathbf{s}$ to time granule $\mathbf{t}$ are defined similarly as the transitions in CTSs. The

technique of embedding many traces into the concept lattice of some suitable context $\mathbb{K}_Q$ is the conceptual generalization of drawing a usual geographic map (see Fig.6). For the mathematical definitions the reader is referred to [15, 19, 24].

Remark: In the definition of a Conceptual Semantic System we do not explicitly represent the relational aspect of a statement as it is done in Conceptual Graphs [4, 5], in Power Context Families and Concept Graphs [10]. These relational structures have been combined with temporal CSSs by the author in Temporal Relational Semantic Systems [21–23] which will be discussed in section 4.

In the following section we use a moving high pressure zone as a typical example of a *distributed* temporal object in a TCSS. We explain this special TCSS starting with a data table and interpreting the values of the data table as formal concepts of suitable formal contexts.

3.2 A Moving High Pressure Zone

As a small example we construct a TCSS which yields a weather map with a moving high pressure zone over Germany. To keep the data table small we construct a map of Germany using a coarse grid of longitude and latitude coordinates into which we embed the (one-element-) traces of 15 towns. To represent a moving high pressure zone we assume that the pressure has been measured in some weather stations (WS) at two consecutive days, say Monday and Tuesday. The data are (partially) shown in Table 3 where the rows 1,...,15 show the latitude and longitude values of 15 German towns, the rows 16,...,25 show the pressure values (in hectopascal (hPa)) measured at certain days at weather stations located in some of the previously mentioned towns.

Table 3: Data table of a Moving High Pressure Zone

instance	place	latitude	longitude	time	pressure
1	Berlin	52.5	13.4	/	/
...	...	...	...	/	/
15	Wilhelmshaven	53.5	8.1	/	/
16	WS Dortmund	51.5	7.5	Monday	1020
...	...	...	...	...	...
25	WS München	48.1	11.6	Tuesday	980

In Table 3 the row labels are called instances, and instance 1 denotes the statement that the *place Berlin* has *latitude* 52.5 and *longitude* 13.4; no time and no pressure is recorded in this line, shown by the sign "/" in the column for *time* and *pressure*. *Instance* 16 tells that *Weather Station Dortmund* has *latitude* 51.5 and *longitude* 7.5 and has reported at *Monday* a *pressure* of 1020 hectopascal.

The semantic scale for *time* is shown in Fig. 4 by a line diagram of its concept lattice. The values "Monday", "Tuesday", and "/" in Tab. 3 are interpreted as the object concepts of the corresponding formal objects. The single arrow in Fig. 4 represents the time relation as a binary relation on the set of all concepts of the specified time attribute.

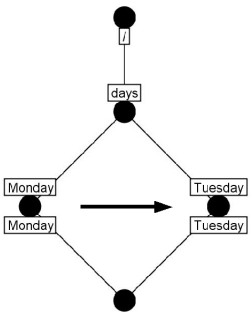


Fig. 4: The semantic scale for *time* with time relation

The semantic scale $\mathbb{S}_p$ for *pressure* is defined as a modified interordinal scale on the multiples of 10 in the interval $[970, 1030]$. Its concept lattice is shown in Fig.5.

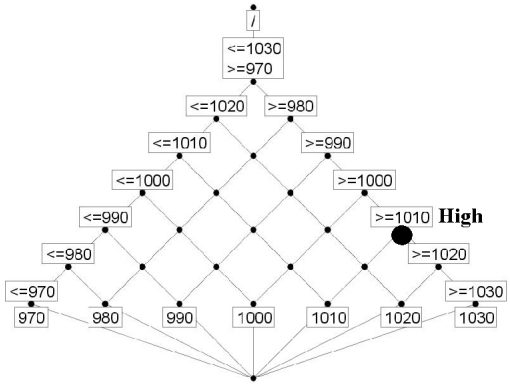


Fig. 5: The semantic scale for *pressure*

In this interordinal scale for pressure we choose the attribute concept $\mu(\geq 1010)$ for the representation of the notion of "high pressure", and call this formal concept **High**. Combined with the formal concept **Monday** in the time scale

we like to form the tuple (**High, Monday**) as our formal representation of such an abstract "object" like "**the High at Monday**".

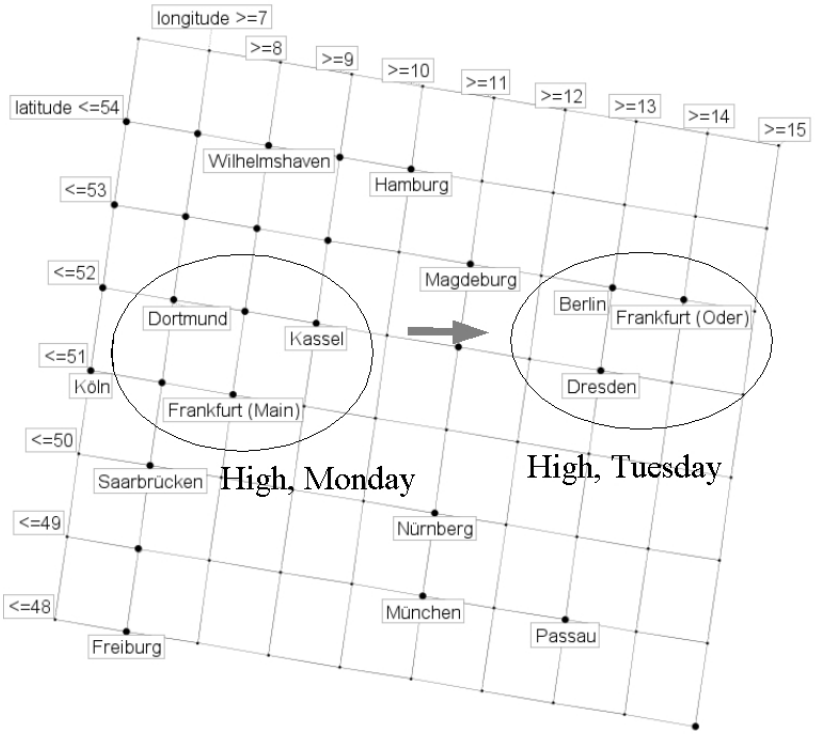


Fig. 6: A Weather Map with a Moving High Pressure Zone

The movement of the selected high pressure zone is visualized in the weather map in Fig.6. Concerning the construction of this weather map we just mention the main steps: 1. Construction of a grid of latitude-longitude-values as a semi-product of two ordinal scales for latitude and longitude. 2. Embedding the (one-element) traces of towns using their latitude-longitude-values. 3. Embedding of the traces of the tuples (**High**, **Monday**) and (**High**, **Tuesday**) as visualized by ellipses in Fig.6. 4. Visualizing a (short) life track from the trace of (**High**, **Monday**) to the trace of (**High**, **Tuesday**).

The details of the construction of the weather map in Fig.6 can be seen in [19]. For another useful application of TCSSs in a biomedical study of disease processes in arthritic patients the reader is referred to [25].

4 Temporal Relational Semantic Systems

A Temporal Relational Semantic System (TRSS) [21–23] is a relational extension of a Temporal Conceptual Semantic System (TCSS). The investigations of TRSSs are based on the theory of Conceptual Graphs as developed by J. Sowa [4, 5] and its FCA-version of Concept Graphs of a Power Context Family as introduced by R. Wille [10]. The main idea for the definition of a TRSS is to write each relational statement (with a single relation term) into a line of a many-valued context of a CSS and protocol the sequence of speaking this statement. A TRSS has a specified set of time attributes, a specified set of temporal objects, and for each temporal object its time relation [23]. Then states, transitions, and traces can be introduced as in a TCSS.

4.1 A tabular representation of a Temporal Relational Semantic System

In the rows of Table 4 we represent some temporal and non-temporal statements: 'from 2008 to 2009 Bob lived in England', 'in May 2008 Bob lived in London', 'in spring 2009 Alice lived in Berlin', 'in spring 2009 Bob met Alice in Paris', and 'Paris is the native town of Alice'.

Table 4: A data table for temporal relational information

statement	r^*	$TIME_1$	$TIME_2$	$PERSON_1$	$PERSON_2$	LOCATION
1	from.to..lived in.	2008	2009	BOB		ENGLAND
2	in...lived in.	May	2008	BOB		LONDON
3	in...lived in.	spring	2009	ALICE		BERLIN
4	in...met. in.	spring	2009	BOB	ALICE	PARIS
5	.is the native town of.			ALICE		PARIS

The main information for reading the statements in the intended sequence is represented in Table 5, called the *position table*. For example, the first position

of *.is the native town of.* is the attribute LOCATION, its second position is the attribute PERSON_1.

Table 5: The position table

r^*	TIME_1	TIME_2	PERSON_1	PERSON_2	LOCATION
from.to..lived in.	1	2	3		4
in...lived in.	1	2	3		4
in...met. in.	1	2	3	4	5
.is the native town of.			2		1

For RSSs the notion of a trace of an object (as a tuple of concepts of the semantic scales) can be used to visualize information concerning several many-valued attributes in the concept lattice of a suitable view. That is shown by some examples of *relational trace diagrams* in [22]. For TRSSs each relation, as for example the relation *in...lived in.* is formally represented as a formal concept of the scale for the relations; therefore the *state* of such a relation at a certain time granule can be defined. That is clearly a powerful tool for the representation of temporal relational knowledge.

5 Conclusions

This paper has shown the leading ideas and some typical examples of several temporal conceptual systems: first the CTSs, second the CTSOTs, third the TCCSs, and fourth the TRSSs. Future research should improve the actual computer programs [1] for the generation of powerful visualizations of temporal relational structures.

References

1. Becker, P., Hereth Correia, J.: The ToscanaJ Suite for Implementing Conceptual Information Systems. In: Ganter, B., Stumme, G., Wille, R. (eds.): Formal Concept Analysis - Foundations and Applications. ICFCA 2003. LNAI 3626, pp. 324–348. Springer, Heidelberg (2005)
2. Ganter, B., Wille, R.: Formal Concept Analysis: mathematical foundations. Springer, Heidelberg (1999); German version: Springer, Heidelberg (1996)
3. Kalman, R.E., Falb, P.L., Arbib, M.A.: Topics in Mathematical System Theory. McGraw-Hill Book Company, New York (1969)
4. Sowa, J.F.: Conceptual structures: information processing in mind and machine. Adison-Wesley, Reading (1984)
5. Sowa, J.F.: Knowledge representation: logical, philosophical, and computational foundations. Brooks Cole Publ. Comp., Pacific Grove, CA (2000)

6. Spangenberg, N.: *Familienkonflikte eßgestörter Patientinnen: Eine empirische Untersuchung mit Hilfe der Repertory Grid-Technik.* Habilitationsschrift am FB Humanmedizin der Justus-Liebig-Universität Gießen, 1990.
7. Spangenberg, N., K.E. Wolff: Comparison of Biplot Analysis and Formal Concept Analysis in the case of a Repertory Grid. In: *Classification, Data Analysis, and Knowledge Organization* (eds.: H.H. Bock, P. Ihm), Springer, Heidelberg 1991, 104-112.
8. Spangenberg, N., K.E. Wolff: Datenreduktion durch die Formale Begriffsanalyse von Repertory Grids. In: *Einführung in die Repertory Grid-Technik,* Band 2, Klinische Forschung und Praxis. (eds.: J.W. Scheer, A. Catina), Verlag Hans Huber, 1993, 38-54.
9. Wille, R.: Restructuring lattice theory: an approach based on hierarchies of concepts. In: Rival, I. (ed.): Ordered sets. pp. 445–470, Reidel, Dordrecht-Boston (1982). Reprinted in: Ferré, S., Rudolph, S. (eds.): Formal Concept Analysis. ICFCA 2009. LNAI 5548, pp. 314–339. Springer, Heidelberg (2009)
10. Wille, R.: Conceptual Graphs and Formal Concept Analysis. In: Lukose, D., Delugach, H., Keeler, M., Searle, L., Sowa, J.F. (eds.): Conceptual Structures: Fulfilling Peirce's Dream. ICCS 1997. LNAI 1257, pp. 290–303. Springer, Heidelberg (1997)
11. Wolff, K.E.: Concepts, States, and Systems. In: Dubois, D.M. (ed.): Computing Anticipatory Systems. American Institute of Physics, Conference Proceedings 517, pp. 83–97 (2000)
12. Wolff, K.E.: Temporal Concept Analysis. In: Mephu Nguifo E. et al. (eds.): ICCS-2001 International Workshop on Concept Lattices-Based Theory, Methods and Tools for Knowledge Discovery in Databases, pp. 91–107. Stanford University, Palo Alto, CA (2001)
13. Wolff, K.E.: Transitions in Conceptual Time Systems. In: Dubois, D.M. (ed.): International Journal of Computing Anticipatory Systems, vol. 11, pp. 398–412. CHAOS (2002)
14. Wolff, K.E.: Interpretation of Automata in Temporal Concept Analysis. In: Priss, U., Corbett, D., Angelova, G. (eds.): Integration and Interfaces. ICCS 2002. LNAI 2393, pp. 341–353, Springer, Heidelberg (2002)
15. Wolff, K.E.: 'Particles' and 'Waves' as Understood by Temporal Concept Analysis. In: Wolff, K.E., Pfeiffer, H.D., Delugach, H.S. (eds.): Conceptual Structures at Work. ICCS 2004. LNAI 3127, pp. 126–141, Springer, Heidelberg (2004)
16. Wolff, K.E.: States, Transitions, and Life Tracks in Temporal Concept Analysis. In: Ganter, B., Stumme, G., Wille, R. (eds.): Formal Concept Analysis - Foundations and Applications. ICFCA 2003. LNAI 3626, pp. 127–148, Springer, Heidelberg (2005)
17. Wolff, K.E.: States of Distributed Objects in Conceptual Semantic Systems. In: Dau, F., Mugnier, M.-L., Stumme, G. (eds.): Conceptual Structures: Common Semantics for Sharing Knowledge. ICCS 2005. LNAI 3596, pp. 250–266, Springer, Heidelberg (2005)
18. Wolff, K.E.: Conceptual Semantic Systems - Theory and Applications. In: Goncharov, S., Downey, R., Ono, H. (eds.): Mathematical Logic in Asia. pp. 287–300, World Scientific, New Jersey (2006)
19. Wolff, K.E.: Basic Notions in Temporal Conceptual Semantic Systems. In: Gély, A., Kuznetsov, S.O., Nourine, L., Schmidt, S.E. (eds.): Contributions to ICFCA 2007, pp. 97–120. Clermont-Ferrand, France (2007)
20. Wolff, K.E.: Applications of Temporal Conceptual Semantic Systems. In: Zagoruiko, N.G., Palchunov, D.E. (eds.): Knowledge - Ontology - Theory. Vol.2,

pp. 3–16. Russian Academy of Sciences. Sobolev Institute for Mathematics. Novosibirsk (2007)
21. Wolff, K.E.: Relational Semantic Systems, Power Context Families, and Concept Graphs. In: Wolff, K.E., Rudolph, S., Ferré, S. (eds.): Contributions to ICFCA 2009, pp. 63–78. Verlag Allgemeine Wissenschaft, Darmstadt (2009)
22. Wolff, K.E.: Relational Scaling in Relational Semantic Systems. In: Rudolph, S., Dau, F., Kuznetsov, S.O. (eds.): Conceptual Structures: Leveraging Semantic Technologies. ICCS 2009. LNAI 5662, pp. 307–320. Springer-Verlag, Heidelberg (2009)
23. Wolff, K.E.: Temporal Relational Semantic Systems. In: Croitoru, M., Ferré, S., Lukose, D. (eds.): Conceptual Structures: From Information to Intelligence. ICCS 2010. LNAI 6208, pp. 165–180. Springer-Verlag, Heidelberg (2010)
24. Wolff, K.E.: Applications of Temporal Conceptual Semantic Systems. In: Wolff, K.E. et al: Knowledge Processing and Data Analysis. LNAI 6581, pp. 60–76. Springer-Verlag, Heidelberg (2011)
25. Wollbold, J., Huber, R., Kinne, R., Wolff, K.E.: Conceptual Representation of Gene Expression Processes. In: Wolff, K.E. et al: Knowledge Processing and Data Analysis. LNAI 6581, pp. 77–99. Springer-Verlag, Heidelberg (2011)

Research Challenges of Dynamic Socio-Semantic Networks

Rostislav Yavorskiy

Witology, Models and Algorithms
Kapranova str. 3, Moscow, RUSSIA
Rostislav.Yavorskiy@witology.com
http://www.witology.com

Abstract. A general model of a socio-semantic network is presented in terms of state-transition systems. We provide some examples and indicate research directions, which seem to us the most important from the application point of view.

Keywords: social network, semantic network, socio-semantic network

1 Model of a socio-semantic network

1.1 Social network

A social network is usually modeled as a weighted multi-graph

$$G = \{V, E_1, \ldots, E_k; \pi, \delta_1, \ldots, \delta_k\},$$

where

- V represents members of the network,
- $E_1, \ldots, E_k \subset V \times V$ denote different relations between the members, e.g. being a friend, follower, relative, co-worker etc.
- $\pi : V \to \Pi$ is a *user profile* function, which stores personal information about the network members.
- $\delta_i : E_i \to \Delta_i$ $(i \in \{1, \ldots, k\})$ keeps parameters and details of the corresponding relation.

1.2 Content

The model of the content has a very similar definition. It is a multi-graph

$$C = \{T, R_1, \ldots, R_m; \theta, \gamma_1, \ldots, \gamma_m\},$$

where

- T stands for the set of all elements of the generated content, e.g. posts, comments, evaluations, tags etc.
- $R_1, \ldots, R_m \subset T \times T$ denote different relations on the content, e.g. being a reply on, have the same subject, etc.
- $\theta : T \to \Theta$ stores parameters of the content;
- $\gamma_i : R_i \to \Gamma_i$ $(i \in \{1, \ldots, k\})$, similarly, keeps parameters and details of the corresponding relation.

1.3 Authorship and other relations between the users and the content

The basic connections between the social graph and the content are defined by the authorship relation A,

$$A \subset V \times T.$$

One can also consider other kinds of connections of this kind, but usually all of them could be modeled via introducing a new type of content. For example, the relation *John is interested in post "Announcement"* could be modeled by introducing a new content node *interest evidence*, which points to "Announcement" (use the corresponding relation R_i here) and is authored by John.

1.4 The context

Before we turn to description of the socio-semantic network dynamics there is one more important parameter not to be missed. It is external context, Ω, which may include different parameters like project or campaign phase, flag for a bank holiday, or a maintenance status of the network.

2 The socio-semantic network dynamics

Now, when all the components of the network have been defined, the list of possible system updates, which determine the network evolution, is rather evident:

- addition of new members to V;
- changes in user profiles π;
- updates of social relations $E_1, \ldots, E_k$ and their parameters $\delta_1, \ldots, \delta_k$;
- creation and update of content nodes in T, (also affects the authorship relation A);
- changes in properties of and relations between the content nodes $\theta, \gamma_1, \ldots, \gamma_m$;
- changes in context Ω.

3 Examples

3.1 School or a training center

A training center usually has a standardized set of reading materials, textbooks, tasks, assignments and exam tests. At the same time, the students' network evolves permanently. In terms of the definition above one can say that the content part of this socio-semantic network is rather stable while the social network is very dynamic.

3.2 Research or analytic team

This example resides on the opposite side of the spectra. The team (the social network part) is rather stable while the content is actively processed and generated.

3.3 Fixed term project

A targeted crowdsourcing project or a collective intelligence venture provide an example of a dynamic socio-semantic network, which is created from the scratch and is aimed at solving a particular task or a problem. New content is generated and new members join the network at all stages of its lifecycle.

4 Research challenges

Assume that we have all necessary data about the network dynamics available. In all the examples mentioned above one can identify two principal tasks for analysis:

- Given all the data about the users activities and the content discover the right people (knowing, capable, skillfull etc.)
- Given all the data about the social network dynamics and the content evolution discover the right texts (interesting, influential, prominent etc.)

Many promising approaches and useful algorithms have already been developed during the last decades [1–4], several new ideas are implemented in the Witology platform [5]. Some have proved to be quite efficient, although most of them are based on fairly simple mathematical tools. Still, the field is rather in its rudimentary phase. We believe that the next breakthrough lies in interdisciplinary research covering sociology, psychology, linguistics and other related fields.

References

1. Sergey Brin, Lawrence Page. *The anatomy of a large-scale hypertextual Web search engine.* Computer Networks and ISDN Systems 30: 107117, (1998).
2. Damon Horowitz, Sepandar D. Kamvar. *The Anatomy of a LargeScale Social Search Engine.* Proceedings of WWW'2010, April 26–30, 2010, Raleigh, North Carolina.
3. Camille Roth, Jean-Philippe Cointet. *Social and Semantic Coevolution in Knowledge Networks,* Social Networks, 32(1):16-29 (2009)
4. Jean-Philippe Cointet, Camille Roth. *Socio-semantic Dynamics in a Blog Network,* IEEE SocialCom International Conference on Social Computing, Vancouver, Canada, August 2009.
5. Witology, "Search of good ideas through search of people, and search of right people through search of ideas", http://www.witology.com

Recommender System Based on Algorithm of Bicluster Analysis RecBi

Dmitry Ignatov[2], Jonas Poelmans[1], Vasily Zaharchuk[1,2]

[1]K.U.Leuven, Faculty of Business and Economics, Naamsestraat 69,
3000 Leuven, Belgium
[5]National Research University Higher School of Economics (HSE), Pokrovskiy boulvard 11
101000 Moscow, Russia
Dignatov@hse.ru
Jonas.Poelmans@econ.kuleuven.be

Abstract. In this paper we propose two new algorithms based on biclustering analysis, which can be used at the basis of a recommender system for educational orientation of Russian School graduates. The first algorithm was designed to help students make a choice between different university faculties when some of their preferences are known. The second algorithm was developed for the special situation when nothing is known about their preferences. The final version of this recommender system will be used by Higher School of Economics.

Keywords: biclustering, recommender system, educational orientation

1 Introduction

Since the introduction of the so called Common State Exam in high schools of the Russian Federation, graduates received permission to apply to enter multiple universities or faculties of the same university whereas in the past they were only allowed to apply to one institution. Students are confronted with an ever increasing complexity of the educational landscape and for this purpose we developed a recommender system to guide them in their search. Students can indicate one or more faculties where they would like to study and our recommender system will make suggestions on alternative institutions in which they might also be interested. The recommender system will also use the browsing and searching history of the candidate student to efficiently suggest relevant universities, faculties, and educational directions.

Currently on the Internet many websites make use of recommender systems, for example, Amazon recommends books in which the client might be interested based on previous items which were viewed by the user. Other examples include the websites http://facebook.com/ and http://twitter.com/ and for Russian companies, the websites http://imhonet.ru/ and http://www.ozon.ru/.

A lot of techniques have been developed for recommender systems and the main principles of these algorithms are described in [5]. We can distinguish between item-based and user-based recommender systems. In item-based recommendation relevant items are presented based on their similarity to items previously accessed or bought by the user. In user-based recommendation users with a similar profile to the current user are gathered and based on the items they accessed or bought relevant suggestions are made. These systems use different kinds of similarity measures, such as Pearson's correlation, Euclidean distance, Jacquard coefficient, and Manhattan distance.

One of the most recent innovations in recommender system research is applying methods based on biclustering. In [1-5] a wide range of biclustering applications has been described including market research, near-duplicate web-document detection, bioinformatics etc. Biclustering is an unsupervised learning method similar to Formal Concept Analysis (FCA) [7, 8, 9].

Comparing to traditional clustering methods biclustering is not a blackbox technique. Comprehensibility is one of its main advantages, i.e. it is possible to understand why objects ended up in the same cluster. For example you might ask why a cucumber and a pair of boots are assigned to the same cluster. With biclustering it can easily be revealed that they are similar because they have the same color and skin surface.

This lack of comprehensibility of traditional clustering techniques may cause serious problems in large data mining projects. To cope with these issues researchers are increasingly focusing on human-centered techniques including direct clustering (John Hartigan's work [6]) and FCA [10]. In this paper we chose to use biclustering instead of the more famous technique FCA because of the scalability issues encountered with FCA. The large number of extracted concepts quickly results in an unreadable lattice.

2 Algorithms

We will use the general biclustering definition, which was given in [1, 3]. Let A be a matrix of size $(n \times m)$, where m represents the dimensionality of the set of objects and n represents the dimensionality of the set of attributes. Then $X = \{x_1, x_2, \ldots, x_n\}$ is a set of objects and $Y = \{y_1, y_2, \ldots, y_m\}$ is a set of attributes. If $I \subseteq X$ and $J \subseteq Y$, then A_{IJ} is a submatrix of matrix A. $A_{IY} = (I, Y)$ is a cluster of objects of matrix A and $A_{XJ} = (X, J)$ is a cluster of attributes of matrix A. $A_{IJ} = (I, J)$ is a bicluster of matrix A. Its objects share similar attributes and its attributes give a description of the objects in the cluster.

There are different formal definitions of a bicluster available for several specific cases, however these are not considered in this paper.

2.1 Algorithm variant RecBi1

RecBi1 is based on the algorithm of Ignatov D.I., which was described in [4]. This algorithm takes as input two contexts and produces a list of recommendations as output.

Context 1: Formal context $K = (S, A, I)$, where S is a list of all faculties of Russian universities, A is a list of faculty attributes, I is a binary relation, that shows that faculty s from S has an attribute a from A.

Context 2: Multi-valued context containing the history of the usage of the system $K_w = (U, A, W, J)$, where U is a list of users, A is a list of faculty attributes, W is a list of weights, that shows how many times u from U has looked at and considered a from A as an interesting item, J is a ternary relation between u, a, and w.

RecBi1

Input: Formal context of faculties $K = (S, A, I)$,
Multi-valued context containing history of usage $K_w = (U, A, W, J)$,
Visits vector $V=(V_1,\ldots,V_{|U|})$,
U_0 is a target user, N is a number of recommendations.

Output: Rec is a list of recommendations

1. (U_0, s_i) // initial couple
2. For s_i in $S \neq \emptyset$ //s_i' are attributes of s_i

   $\text{CandidateS}(cs) \bigcup (s_i, s_j' \bigcap s_i')$
3. For cs_i in cs

   $kc[cs] =$
4. $kc = \text{dec\_sort}(kc)$
5. $Rec = \text{Top}(N, kc)$

Return *Rec*

2.2 Algorithm variant RecBi2

The second algorithm variant consists of two parts: Recbi2.1 and RecBi2.2. RecBi2.1 is used with so called cold start, which means that there is no previous usage history available. RecBi2.2 is used when the user is using this system not for the first time.

RecBi2.1 takes the same contexts as input as RecBi1 and outputs a list of recommendations. But in the middle of the algorithm it forms a new formal context.

Context 3: Formal context of user preferences $K_{lp} = (U, S, Z)$, where U is a list of users, S is a list of faculties, and Z is a binary preference relation between u from U and s from S.

RecBi2.1

Input: Formal context of faculties $K = (S, A, I)$,
Multi-valued context containing history of usage $K_w = (U, A, W, J)$,
Visits vector $V=(V_1,\ldots,V_{|U|})$,
U_0 is a target user, N is a number of recommended items, l_{min} is a minimal index of interestingness.

Output: *Rec* is a non-ranked list

1. $K_1 = K_w \cdot K^T = (U, S, L, Y)$ //multiplication of multi-valued contexts as a matrix
2. $Z = \emptyset$; $K_{pl} = (U, S, Z \subseteq U \times S)$
3. For u in U

   For s in S

   If $scu \geq l_{min}$ then $Z=(s, u) \cup Z$ //reduction to formal context
4. (U_0, s) //s' is a set of all visitors of s
5. For u in s'

   $S_c = \cup u'$
6. For s_c in S_c

   $V(s_c)=|s_c' \cap S|$
7. v = dec_sort(v)
8. *Rec* = Top(N,v)

Return *Rec*

RecBi2.2 works in a similar way as RecBi2.1, but its output is a ranked list of search results based on user feedback.

3 Conclusion

In this paper we presented two biclustering-based algorithms which will be used at the basis of recommender system which will guide high school students in their search for an educational institution. The main contribution of this paper is a successful application of biclustering methods to the educational domain. In the future we will operationalise this system at the Higher School of Economics.

Acknowledgements
Jonas Poelmans is aspirant of the “Fonds voor Wetenschappelijk Onderzoek – Vlaanderen” or “Research Foundation Flanders”. The contribution of Dmitry Ignatov

is partially supported by the Russian Foundation for Basic Research, project No. 08-07-92497 – NTSNIL_a.

References

1. Ignatov, D.I., Kuznetsov, S.O.: Biclustering of Object-Attribute Data Based on Closed Sets Lattices. In proceeding of 12th Russian Conference in Artificial Intelligence, Vol. 1., pp.175-182. Fizmatlit, Moscow (2010) (In Russian)
2. Ignatov, D.I., Kaminskaya, A.Yu., Magizov, R.A. A Cross-Validation Technique for Recommender Systems Evaluation. In proceeding of 12th Russian Conference in Artificial Intelligence, Vol. 1., pp.183-191. Fizmatlit, Moscow (2010) (In Russian)
3. Ignatov, D.I., Kaminskaya, A.Yu., Magizov, R.A. A Concept-Based Biclustering Algorithm. In proceedings of International conference "Intelligent Information Processing" IIP-8, Cyprus, Paphos, October 17–24, 2010, pp. 140 – 143. MAKS Press, Moscow (2010) (In Russian)
4. Ignatov D.I. Models, Algorithms and Software Tools for Biclustering Based on Closed Sets. PhD Thesis (Thesis for Candidate of Technical Sciences degree), NRU HSE, Moscow (2010) (In Russian)
5. Toby Segaran, Programming Collective intelligence. O'Reilly (2007)
6. Hartigan, J.A. "Direct clustering of a data matrix". Journal of the American Statistical Association 67 (337): 123-9 (1972)
7. Poelmans J, Elzinga P, Viaene S, Dedene G. Formally analysing the concepts of domestic violence, Expert Systems with Applications, vol. 38, no. 4, pp. 3116 – 3130 (2011)
8. Poelmans, J., Elzinga, P., Viaene, S., Dedene, G., Curbing domestic violence: instantiating C-K theory with formal concept analysis and emergent self organizing maps, International Journal of Intelligent Systems in Accountancy, Finance and Management , vol. 17, pp. 167 – 191 (2010)
9. Poelmans J, Elzinga P, Viaene S, Van Hulle M, Dedene G. Gaining insight in domestic violence with emergent self organizing maps, Expert systems with applications, vol. 36, no. 9, pp. 11864 – 11874 (2009)
10. Wille, R. (1982). Restructuring lattice theory: an approach based on hierarchies of concepts. I. Rival (Ed.): Ordered sets, 445-470. Reidel. Dordrecht-Boston.

Author Index