Business Intelligence Processing on the Base of Unstructured Information Analysis from Different Sources Including Mass Media and Internet

O. Zolotarev[1], M. Charnine[2], A. Matskevich[2], K. Kuznetsov[2]
[1]Russian New University, Moscow, Russia
[2]Institute of Informatics Problems, Federal Research Center,
Russian Academy of Sciences, Moscow, Russia

Abstract - This article describes approaches to the business intelligent processing on the base of unstructured information analysis. The problem of reputation management of companies is considered. The review of linguistic, statistic and semantic instruments of text analysis is presented. Several examples of natural language texts analysis are discussed. The information is extracted from the natural language texts from different sources. Entities, relations and processes are built in the form of semantic network. The questions of model constructing of a subject domain are discussed. The models are presented as fragments of semantic network built on the base of the declarative instrument DECL.

Keywords: Business Intelligence Processing, Business process, semantic networks, fragments of knowledge, objects, thesaurus, Big Data

1 Introduction

The functioning of a modern enterprise is connected with challenges and risks, both external and internal. Being in a tough competition, companies are forced to use a variety of analytical and intellectual methods to protect their reputation, to save the confidential information, to identify different kinds of leaks of documents, to minimize the potential damage from unauthorized activity on the part of their employees and the representatives of competing firms. For this purpose data is obtained from a variety of sources, including the Internet. Significant processes, activities and events are tracked; the links between individuals are checked. Related companies with different kinds of information resources are determined by the degree of participation of their individuals in different actions that may harm the enterprise.

Ongoing studies are considered from the point of view of tactical and strategic goals. The modern approach to the analysis of company's activity is based not only on obtaining direct information about its work, enterprise processes, documentation, monitoring the work of the enterprise, but also it is based on a serious study of open-source information. Every year Internet has more and more influence on this process. The most important things are information analysis, the formation of conclusions based on the results of the analysis and the information filtering due to the huge volumes of data in the public domain. Of particular importance is the information posted on the Internet, because in this case, we can process it automatically that couldn't be possible until the information resources were transferred into electronic format.

To reduce the space of search for relevant information the subject area associative portraits (SAAP), subject-specific dictionaries (SSD) and thesauri are generated [7]. As a result the knowledge base is formed. This allows us to identify certain objects and processes of the subject area, to significantly increase the usefulness retrieved from the Internet information. For this purpose, texts retrieved from the Internet are processed with the linguistic processor [1].

As a result of information processing there are a large amount of data for SAAP, thesauri and subject dictionaries that are constantly replenished with new data improving the quality of filtration of texts from the Internet to improve the reliability of analytical results.

The analytical research of the company background can be done by a group of analysts who explore the results of automatic processing of information from the following sources [2]:

- information from the Internet;
- materials from various databases;
- materials of analytical centers;
- electronic documents of the company.

There are significant issues in automatic processing of data from public sources:

- huge amounts of information including spam;
- difficulty in identifying specific objects, processes, situations on the basis of analysis and comparison of information from different sources;
- unreliability of data;
- explicit and implicit misinformation;
- incomplete data;
- constant volatility of environment, situations;
- historical variability;
- data aging, etc.

As a result of continuous improvement of the recognition mechanism of identification of objects and processes, and also because of the accumulation of knowledge that can be measured against the experts, the efficiency of the process of determining leaks and identifying the actions of individual actors, directed against the organization, can be substantially improved [3].

One of the objectives of any enterprise is the task of forming a positive image of the company. For this purpose can be used the results of the analysis of the external and internal environment of the enterprise to form purposeful influences on the organization background for the correcting its image.

The company's reputation is a kind of evaluation of a group of individuals about the company, group of people or individual on the basis of certain criteria. Moreover, there is a clear dependence of the capital of the company from its reputation. Goodwill or Reputational Capital is directly related to the financial condition of the company [14]. One of the goals of the company is minimizing the amount of negative information about the firm in Internet. Now there are a lot of companies that offer Internet services to correct negative image of the organization, for example, Online Reputation Management (ORM) companies and Search Engine Reputation Management (SERM) companies [4-6].

2 Linguistic, statistic and semantic instruments of text analysis

Currently the market offers a large variety of tools for analyzing texts. There are a lot of informational instruments for Internet research called textual-analytical tools or processors of data collection. There are quite a number of companies that offer tools for analyzing texts on the Internet aimed at identifying leaks of information, analysis of reputation, etc. Below is a description of some of these tools.

The InfoTracer program packet represents the toolkit for online public records searches. This firm is a member of the Private Investigator Union proposes several types of searches such as Comprehensive Background Report, Criminal Records Search, People Search, Property Search, Email Search, Company Background Search and so on. The most interesting thing about InfoTracer searching is the discovering of criminal backgrounds, past addresses, possible relatives, property ownership information, public profile information, registration information, birth, death, marriage and divorce record information, citizenship or even marital status.

Taiga is the French processor for extracting information within Internet from patent databases, news reports and publications on scientific conferences. It allows gaining valuable information about the most promising scientific developments and areas where competition is relatively small so far. The system analyzes not only text, but also drawings, diagrams and graphs. It collects any requested information as soon as it appeared on the Internet. The information about links between two countries can be discovered in a few hours.

Tropes software suggests high performance text analysis. It's designed for semantic classification, keyword extraction, linguistic and qualitative analysis.

Tropes can detect contexts, themes and principal actors, through the application. This system can extract objects and subjects, time, place and purpose from the situation. Tropes carries out a chronological analysis of a text with the help of Natural Language Ontology Manager based on Semantic Networks and Natural Language Text Analysis technologies, supplied with several ready-to-use classifications.

It's a searching engine that includes: topic summaries, categories, disambiguation, official sites. It can be used to define people, places, things, words and concepts; to provide direct links to other services; to list related topics; and to give official sites when available.

This company aims to provide intelligent systems based on Semantic Technologies and Big Data. Product Features: Automated semantic sentiment analysis for brand/products; Trend analysis and consumer profiling; Broad coverage and customization support; Semantic Search (Search by meaning not by keyword).

The urgency of the task of analytical intelligence is not in doubt. A large number of groups engaged in similar issues. Using methods of semantic analysis is practiced all over the world, which allows substantially to protect the enterprise from external and internal threats and to manage the organization's reputation in the media.

3 Texts procession, Information retrieval and related objects allocation

Now for the formalization of domain knowledge and building the structure of the business processes are used different methods of knowledge representation. This article discusses an approach for the formation of domain knowledge in the form of a semantic network, which is represented as nodes and relations between them [13]. Usually in the form of nodes are represented certain objects (subjects or entities) of the subject area (SA) [8], and in the form of relationship - operations, in which data objects involved or links between these objects. A semantic network is described in the form of fragments [1,2,3,9]. Below presented fragments of a semantic network describing the abstract objects and relationships between them:

R1(A1,A2/N1) R2(A3,A4/N2) R3(A5,A6/N3) (1)

Here R1,R2,R3 are the names of relations;

N1, N2, N3 - are the names of fragments (N1 - the name of the whole fragment R1(A1,A2); thus it can be omitted, if not required while processing);

A1 - A6 - object names.

For example, as a result of analysis of information from the Internet, there was extracted the following text fragment:

"Bondarev took part in a project to purchase equipment for the company Nogos". For convenience all the insignificant details were omitted (name and patronymic). As a result of the processing of this phrase is creating a corresponding fragment of a semantic network [10]:

Purchase of equipment (Nogos, Bondarev) (2)

Thus, it is known that Bondarev does not work at the Nogos enterprise and can't be formally involved in the project for purchasing equipment. But miss Kravtsova works at the Nogos company and she is a sister to Bondarev. In this case, the following information is placed in the system in the form of a semantic network (or the network should be built on the basis of the analysis of text information of the enterprise)[11]:

Sister(Kravtsova, Bondarev) (3)

Worker(Nogos, Kravtsova) (4)

A given set of predicates corresponds to the following fragment of a semantic network (Fig. 1):

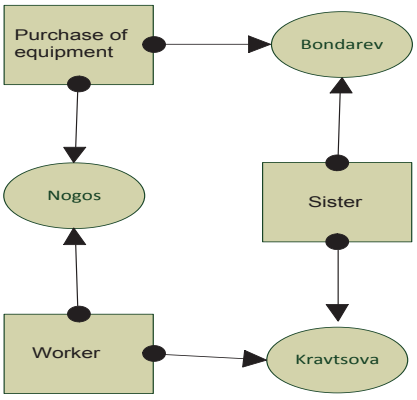

Fig. 1. A fragment of semantic network (1) - (4)

So we can conclude that the company's employee Kravtsova used his family ties to promote a project, either for personal gain from the transaction of funds or as a contractor of the project of purchasing the equipment.

In this case, we can talk about the using of neighborhood approach to the analysis of knowledge. Every object has its own surrounding with other objects (entities) or nodes in the framework of a semantic network. The neighborhood of the object is defined by the presence of "close" semantic relations. There are several levels of proximity within the vicinity of the object - the first, the second, etc. For example, for the object "Kravtsova" mr. Bondarev (her brother) is adjudged to be in the neighborhood of the first order, i.e. they are connected directly through one of the elements of a semantic network. In this case it was done by means of fragment Sister(Kravtsova, Bondarev). These fragments correlate with those objects built in the system ontology, which contain information on the relationships (sister, brother), classes of objects (classes of people). In addition, if the ontology has already the described above information (about Bondarev, Kravtsova, their family relationships), then identification of objects is done, for example, there are already a lot of objects from Internet included in the ontology, which substantially simplifies the situation recognition [12].

The described above case can be considered as ideal, in other words, in this situation, there is no difficulty with doing the conclusions. But there are much more complicated situations, when the analysis of the semantic proximity is done not only by family ties and indirect ties but with other aspects too. Consider the following example, making changes in the situation:

"Mt. Petrov took part in a project to purchase equipment for the company Nogos" and we have no direct information about the relationships of Mr. Petrov. If there is information on the Internet, in social networks, that Petrov is a friend of Bondarev, then will be built the following semantic network fragments:

Sister(Kravtsova, Bondarev) (3)

Worker(Nogos, Kravtsova) (4)

Purchase of equipment (Nogos, Petrov) (5)

Friend(Petrov, Bondarev) (6)

As a result, we have the following. The information leaked from the enterprise, perhaps in this case, the company incurred additional losses in the form of payments to the person concerned, or higher prices for the equipment. The situation is already described by the second order neighborhood where there are no direct links between objects. As a proxy fragment is the predicate "Friend".
A given set of predicates corresponds to the following fragment of a semantic network (Fig. 2):

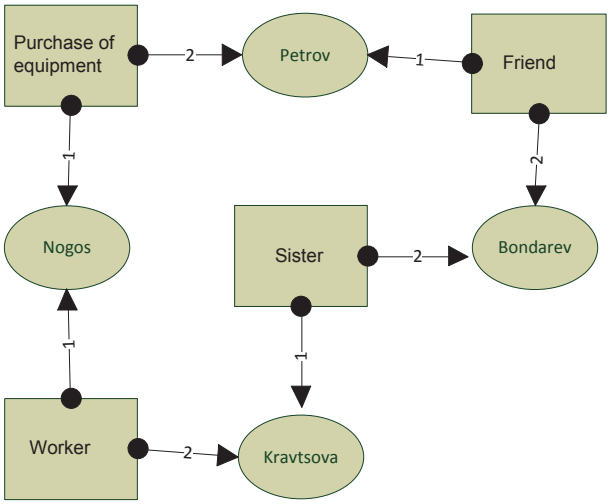

Fig. 2. A fragment of semantic network (1) - (6)

Neighborhoods of vertices can be based on different grounds when there are not direct relationships between objects. You can determine the presence or absence of communication between objects in a number of ways, for example:

- by finding objects in the same place (several times);
- by finding objects in the same place at the same time (several times);
- by training in the same school;
- by occurrence in the same organization (club, car);
- by the similarity of interests;
- by localization of the place of residence or work (in a broad sense).

This situation can be augmented with additional information found on the Internet, if two objects are in the same place at the same time, and can be described by fragment type:

Place(PlaceName, Object, Month, Year) (7)

For example, if there was extracted the text like this "Petrov had a vacation in June 2015 in Sochi" and at the same time there was extracted the following text "Bondarev usually has a vacation in Sochi in August, then we can add to fragments already constructed two more fragments (excluding the fragment Friend(Petrov, Bondarev)):

Place(Sochi, Petrov, June, _) (8)

Place(Sochi, Bondarev, June, 2015) (9),

This will provide some connection between these two individuals in case if Petrov takes part in the procurement project tj buy the equipment (see above). As a result will be built the following semantic network (in this fragment presents a synthesis of information from the ontology of the subject area) (Fig. 3):

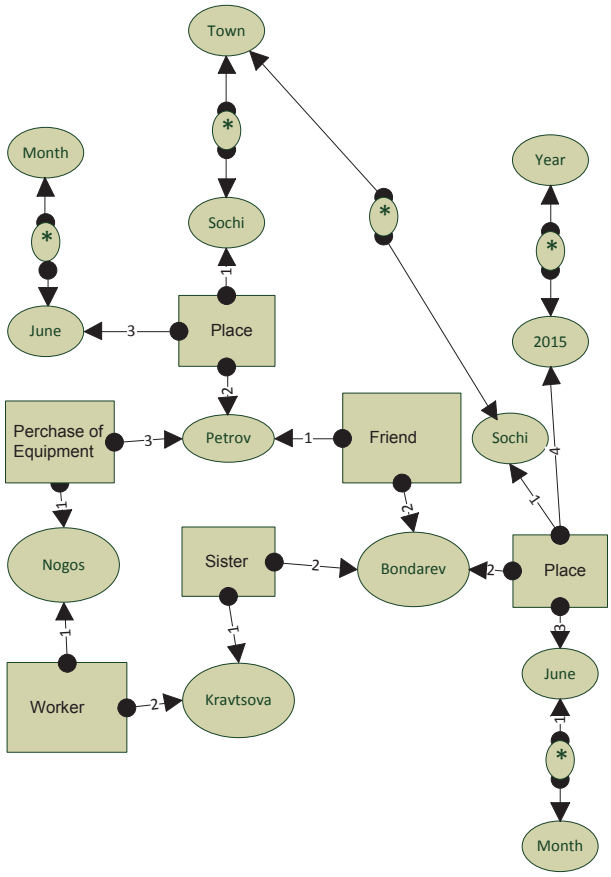

Fig. 3. A fragment of semantic network (1) - (9)

The following fragment presents a semantic network fragment with the information from ontology in the form of a series of generalizing concepts (categories), namely, "month", "year", "city", which was described above in verbal form. Such categories forms polyhierarchy and it is extraordinarily useful for performing inference in the process of natural language processing. A particularly interesting special case of polyhierarchy categories, which is allocated in the main tree contains all vertices of polyhierarchy. This polyhierarchy with a dedicated tree is called Keywen-hierarchy. Keywen hierarchy allows to find higher categories for each concept, which increases the flexibility of the representation of the categories system. At the same time Keywen hierarchy allows us to construct an unambiguous highlighted path to the root from all the nodes of the hierarchy, for example, Russia > city>Sochi. The pathway reminds address that's easy to navigate and to control the correctness of the structure of categories. Methods for constructing Keywen-category hierarchy described in [7,8].

4 Conclusion

The discussed in this article approach allows not only to make the organization more secure, but also to determine the most relevant and promising directions of its development. We are not only interested in people's actions that can harm the business, but with the opportunities as a result of changes in

the environment. New interesting directions are revealed. New firms and stakeholders are extracted from specific subject areas, with analysis of the connection between them and areas for development. Priority activities are detected. The analysis highlighted the strategic goal of enterprise development [15].

Constant monitoring of open source allows us to gain advantages in the following aspects:

- analysis of possible risks and opportunities for development;
- development of a proactive plan of action to win competitors;
- identifying new areas of development,
- detecting the emergence of new competitors;
- analysis of the reputation of the company in the community;
- definition of channels of information leakage;
- formation of positive image of the company;
- identifying allies and competitors.

5 Acknowledgements

This work is supported by the Russian Science Foundation, grant #15-11-30040, and by Russian Foundation for Basic Research, grant #13-07-00272 "The methods for automatic creation of associative portraits of subject domains on the basis of big natural language texts for knowledge extraction systems".

6 References

[1] Kuznetsov I.P. Semantic Representations // Moscow: "Nauka", 1986. 290p.

[2] Kuznetsov I.P., Matskevich A.G. The System for Extracting Semantic Information from Natural Language Texts // Proceedings of the Dialog International Workshop "Computational Linguistics and its Applications", Vol.2, Moscow: Nauka, 2002.

[3] Kuznetsov I.P. Natural Language Texts Processing Employing the Knowledge Base Technology // Sistemy i Sredstva Informatiki, Vol.13, Moscow: Nauka, 2003, pp. 241-250.

[4] Somin N.V., Solovyova N.S., Charnine M.M The System for Morphological Analysis: the Experience of Employment and Modification // Sistemy i Sredstva Informatiki, Vol. 15 Moscow: Nauka, 2005, pp. 20-30.

[5] Web site "Knowledge extraction for Analytical Systems": http://Iiranlogos.com/english/

[6] Kuznetsov, I.P., Kozerenko E.B. Semantic Approach to Explicit and Implicit Knowledge Extraction // Proceedings of ICAI'11, WORLDCOMP'11, July 18-21, 2011, Las Vegas, Nevada, USA. - CRSEA Press,
USA, 2011.

[7] Charnine, Michael Vladimir Charnine. Keywen Category Structure.// Wordclay, USA, 2008, pp.1-60.

[8] Charnine, M. M., "Keywen: Automated Writing Tools", Booktango, USA, 2013, ISBN 978-1-46892-205-9.

[9] Charnine, M.M., I. P. Kuznetsov, E. B. Kozerenko,. Technological peculiarity of knowledge extraction for logical-analytical systems. WORLDCOMP'12 July 16-19, 2012. Las Vegas, USA.// CSREA Press, pp. 49-55, 2012.

[10] Zolotarev, O., M. Charnine, A. Matskevich. Conceptual Business Process Structuring by Extracting Knowledge from Natural Language Texts. Proceedings of the 2014 International Conference on Artificial Intelligence (ICAI 2014), vol.I, WORLDCOMP'14, July 21-24, 2014. Las Vegas Nevada, USA. CSREA Press, pp.82-87.

[11] Zolotarev, O., Kozerenko E. B., M. Sharnin. Principles of construction of models of business processes of the subject area based on the processing of natural language texts // Bulletin of ROSNOU. - M.: RosNOU, 2014. No. 4. P. 82-88.

[12] Zolotarev, O. Process approach to the management of projects in the implementation of corporate information systems. // Bulletin of ROSNOU. - M.: RosNOU, 2014. No. 4. P. 89-92.

[13] Zolotarev, O. Methods of extraction processes, objects, relations from natural language texts. // Security problems of the Russian society. - Smolensk: Svitok. 2014.

[14] Zolotarev, O. Control in implementation projects of distributed Enterprise Resource Planning Systems. Bulletin of the Russian New University, Moscow, RosNOU, 2012. No. 4. P. 78-80.

[15] Zolotarev, O. Innovative solutions in the formation of functional structure of the subject domain. Bulletin of the Russian New University, Moscow, RosNOU, 2013. No. 4. P. 82-84.