# The Knowledge on the Basis of Fact Analysis in Business Intelligence

Alexander Vokhmintsev, Andrey Melnikov



# The Knowledge On the Basis of Fact Analysis in Business Intelligence.

Vokhmintsev Alexander, Melnikov Andrey

Chelyabinsk State University
vav@csu.ru

**Abstract.** This article discusses the problem of searching and analyzing scientific and technical information gathered from open sources in the field of business intelligence. Subject area is represented through ontological model extended by fuzzy links between objects. Using this ontology-based approach, we have created "Cyber-analytic", information analysis system which is capable of extracting facts from natural-language texts in Russian and English, and representing them as object relationships. We use i2 Analyst's Notebook. as a visualization tool.

**Keywords:** Fact analysis, semantic network, knowledge base, full-text databases, ontologies, text mining, data processing, applied linguistics, fact extract, factual data model, syntactic-semantic tree of dependencies.

## 1 INTRODUCTION

Modern information technology is characterized by rapid growth of data volumes of different kinds (text documents, relational databases, multimedia and biometric data, etc.). Major part of the data is produced in form of natural language texts (according to Gartner, more than 96% of data falls in this category.) This way of representing information is most natural for the human being because we use natural language constructs in our thinking processes. At the same time, this type of information requires complex models and algorithms which help to build formal representation of text document through linguistic analysis.

As of today, there are successful systems for linguistic analysis based on morphological text analysis (including those with fuzzy search algorithms), thematic analysis, building of semantic portrait of text document, and intellectual ontology based systems for specialized subject fields. As a result of morphologic and thematic analysis, usually a bundle of thousands of text documents is formed, where documents are sorted by relevance. This amount of documents is very difficult to analyze, that's why with the development of Internet (the biggest depository of text documents) new ontology-based methods of semantic analysis of text documents started to appear. The term "ontology" was used in several AI research communities, firstly in the field of knowledge engineering, natural language processing, and later in knowledge representation (Jos de Bruijn, Fensel D., Staab S., Studer R.). Later on, ontologies started

adfa, p. 1, 2011.

to be regarded as key element in Semantic Web project. The main drawback of those methods is that they can be used in practice only for information systems working with narrow subject fields (for example: industry, metallurgy, mechanical engineering, power engineering) where it's possible to precisely describe technology and business processes of a company. All attempts to use these methods for more general cases (for arbitrary subject fields) were unsuccessful.

To solve this problem in general case, it is necessary to use methods based on deep syntax-semantic text analysis, which fall into three main groups:

- syntactic methods: analyzing syntactic characteristics and lexical composition of texts;
- statistical methods: establishing of semantic links on the basis of extracted statistical relations;
- semantic methods: those methods works through exploring deep semantic connections in a sentence, a text document and beyond (for example: knowledge base, thesaurus, ontology, explanatory dictionary, etc.).

Research on deep semantic analysis of texts has been carried out for quite a long time, including works of N. Chomsky (transformational grammar, universal grammar), I.A. Mel'čuk ("meaning-text" model), G. Scragg ("Semantic nets as memory models"), R. Schenk ("Processing of conceptual information"), C. Fillmore ("Frame semantics and the nature of language", "The Case for Case") [1,2]. In these methods, semantic analysis uses knowledge bases which contain information about subject area.

In this article we'll discuss following points:

- Describing the knowledge model to store data on monitored objects of scientific and technical information;
- Developing a method for factual analysis of text documents for knowledge bases;
- Applying the method of factual analysis in "Cyber-analytic" information analysis system.

## 2 ONTOLOGY BASED MODEL OF KNOWLEDGE BASE

Considerable proportion of scientific and technical information in modern world is open, especially during the stages of idea formulation, discussion and approbation. Counter-espionage services and industry use this open information for the purposes of industrial design (e.g. pre-production models, classification of documentation) or promotion of science intensive products on the market. To solve the problem, an expert would integrate the knowledge, analyzing heterogeneous information produced by other scientists and engineers in order to get some knowledge on the object or technology in question. In this case, electronic Internet resources (scientific articles, conference proceedings, industrial and business news, professional resources), design documentation, offline publications in scientific libraries and social networks (online conferences, forums, blogs) are the main source of information.

Fulltext databaseis a set of cross-referenced documents, which can be most adequately represented as a semantic network. To store detailed information about subject area, the information needs to be supplemented by set of ontologies and metadescriptions of knowledge contained in text documents (Berners-Lee T., Hendler J., Lassila O.) [3]. Thus, we can introduce a definition of ontology for the purpose of scientific and technical monitoring:

$$O = \{O_i, A, O_{links}, O_{properties}, A_a, A_p\}$$

where

- $O_i$ - object identifier;
- $A$ - object attributes;
- $O_{links}$ - connections in semantic network, l inking the object to other objects;
- $O_{properies}$ - object characteristics;
- $A_p$ - manipulations with the object;
- $A_a$ - actions of the object;
- $C$ - contexts of the object;
- $T$ - timeline of the object (change of object attributes over time).

Proposed knowledge model has advantages over traditional ontologies, as well as over semantic networks, since the objects are viewed in context of source text document (paragraph, sentence, etc.) and chronological timeline. This approach permits to separate morphologically equal lexemes using semantic information about their attributes, contexts and object timelines. For example, persons “Angela Merkel politician” and “Angela Merkel florist” will be separate in semantic search results.

In the process of monitoring full-text document collections, analyst tries to find in texts some regular patterns, events, persons (organizations), product names (technologies), and their relationships. In modern applied linguistics these constructions are called facts [4]. The figure demonstrates graphic representation of a fact: relationship between a person and an organization in the context of employment.

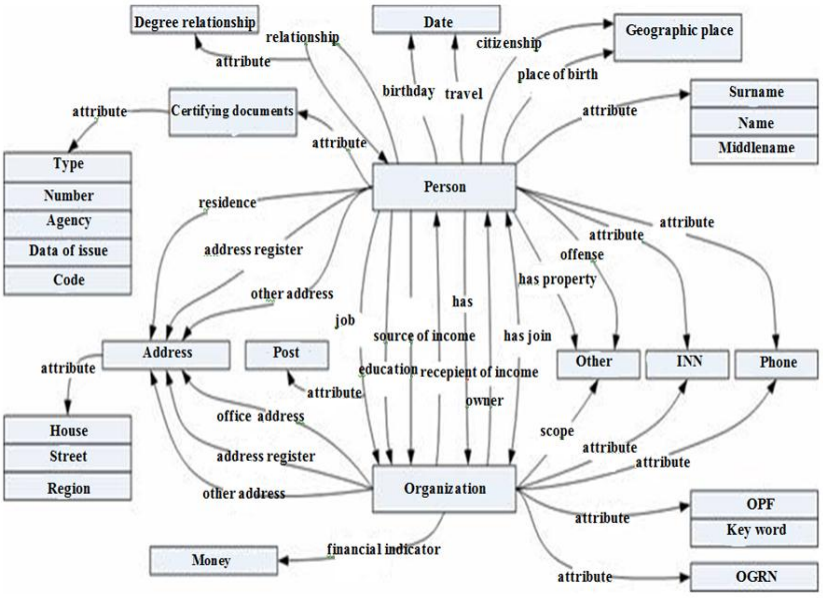


**Fig. 1. -** Graphic representation of a fact

## 3 FACTUAL ANALYSIS OF TEXTS

Knowledge bases can be built by subject area experts, but also automatically using Text Mining algorithms that extract data on objects, their attributes and connections using formal rules. Among tools for semantic analysis we could note Convera RetrievalWare and Russian Context Optimizer (for documents in Russian language) [5]. In this work we will discuss the factual analysis method of natural language texts, which allows to extract information on monitored objects using predefined semantic inference rules. Let's introduce basic definitions.

**Definition 1** **Target object** – basic text unit, represented in text as a word, word-combination or alpha-numeric construction. Let's call the tuple $O_{target} = \langle a_{name}, a_1, a_2, \ldots, a_n \rangle$ a **target object**, where $a_{name}$ – target object's name; $a_1, a_2, \ldots, a_n$, $a_i, i = \overline{1,n}$ – attributes of subject area concept, so that $\bigcap_{i=1}^{n} a_i = \varnothing$.

Examples of target objects: artificial intelligence, E.ON Ruhrgas, Jurgen Müller, one thousand years ago, DE0008469008, US78378X1072, +49 (0) 172 9587 163, $3000, 98%.

Target objects (TO) can be subdivided into two classes: significant TO and insignificant TO. Insignificant TO include all auxiliary sentence elements not having any

meaning by themselves (service parts of speech — unions, prepositions and punctuation). Significant target objects, in their turn, fall into three categories:

1. **Named objects** – this class includes following semantic types: persons, organizations, geographic objects, technology and product names, and other proper nouns.
2. **Unnamed target objects** - this class includes full words of following parts of speech: common names, adjectives, auxiliary verbs, animated and unanimated objects, object attributes, events.
3. **Special target objects** – entities encountered in some special constructions in text, consisting of alpha-numeric characters: dates, adverbial modifiers of time, money amounts, identification data of people and organizations, etc.

**Definition 2 Target object attributes**

$A_{target} = \{A_i, A_{class}, A_{morph}, A_{syntactic}, A_{semantic}, A_{structure}, A_{ref}, A_c\}$

Every target object has the following set of attributes:

- $A_i$ -Target object identifier is represented by a normalized string or known object identifier (person ID or organization ID);
- $A_{class}$ - target object's category. It is also possible to define custom categories;
- $A_{morph}$, $A_{syntactic}$ - grammatical attributes (part of speech, genger, number, case, person, etc.);
- $A_{semantic}$ - semantic attributes;
- $A_{structure}$ - structural characteristics of TO include its distinctive features, for example: first and last names for a person;
- $A_{ref}$ - set of links to all text sentences that mention the TO;
- $A_c$ - communicative characteristics — set of attributes, determining communicative position of the TO in the sentence: part of the sentence and type of syntactic construction (clause).

**Definition 3 Semantic network**

Nodes of semantic network are the TO, mentioned in the text, links are all the different types of syntactic and semantic relations between the objects. The simplest semantic network is a result of syntactic analysis and additional semantic transformations of the syntactic dependency tree in a given sentence. Let's call tuple $\tilde{H} = \langle O_{target}, \tilde{E} \rangle$ semantic network of a document, where $e = \{\mu_{R(\tilde{H})^{\Sigma-\Delta}}(x_\alpha, x_\beta)/(x_\alpha, x_\beta), e_{type}, e_{role}, e_{case}, e_{key}, e_{force}, e_R\}$ Connections in semantic network have the following set of attributes:

- $\mu_{R(\tilde{H})^{\Sigma-\Delta}}(x_\alpha, x_\beta)/(x_\alpha, x_\beta)$ - degree of contiguity of nodes in the semantic network (grade of membership);

- $e_{type}$ - type of the semantic and syntactic relation between nodes: ablative, agent, addressee, argument, derivative, destinative, causative, characteristic, membership, etc.;
- $e_{role}$ - semantic role defined for connections between predicate and argument, usually taken from the list of government models, for example subject, object, instrument;
- $e_{case}$ - semantic case and connector (preposition, union) between nodes in dependency tree;
- $e_{key}$ - connection identifier;
- $e_{force}$ - strength of connection between nodes;
- $e_{R}$ - references to sentences in text document.

Semantic network of a sentence is an example of fact in its simple form. More complex facts could be defined through anaphoric or other references in the paragraph, text block, text document or collection of text documents.

Factual analysis can be broken down into following main stages [6]:

1. Pre-syntactical analysis

- morphological text analysis;
- processing of word forms not present in the dictionary;
- identification of stop words;
- preliminary analysis of typical structures in text document.

The result of 1st stage of processing is a set of sentences, where each sentence contains ordered list of words with variants of homonymous lexemes.

2. Syntactical analysis

- extraction of standard sentence constructions using the morphology data, building of word-combinations;
- identification of syntactic and semantic constructions in the text;
- construction of syntactic and semantic dependency tree.

After the 2nd stage, the set of syntactic and semantic dependency trees is formed. The most probable parsing variant is then chosen using heuristic algorithms.

3. Post-syntactical analysis

Analysis of word forms not included in final variant of syntactic and semantic dependency tree.

4. Semantic analysis

- extraction of target objects from the text;
- building the logical scheme of the situation;

- classification of semantic networks;
- context analysis;
- chronological analysis, elimination of collisions.

The result of 4th stage of processing is a set of target objects with semantic connections. To extract target objects, it is necessary to build the logical scheme of the situation. Resulting set is then classified into categories.

Rule of extraction of target objects (Person: Full Name):

```
macro: strict_full_name (
(name_feminine (middlename_feminine)?
(nastname_feminine):morph_info ) |
(name_masculine (middlename_masculine)? (mast-
name_masculine):morph_info ) |
((lastname_feminine):morph_info name_feminine (niddle-
name_feminine)?) |
((nastname_masculine):morph_info name_masculine middle-
name_masculine)? ) |
((name_feminine):morph_info middlename_feminine) |
((name_masculine):morph_info middlename_masculine )
):whole_strict_full_name
rule: strictfullnamerule strict_full_name
--> :whole_strict_full_name.token = { semantictype =
"fullname:fullname", type = "word" },
:whole_strict_full_name.morph = { :morph_info.morph },
:whole_strict_full_name.rule = { rule = "strictfullnam-
erule" }
```

Factual analysis of the text can find descriptions of situations corresponding to certain templates, for example invention of address bus or stock purchase. Fact search is performed within semantic network of text document. Logical scheme of the situation identifies class of possible situations and contains slots that unambiguously separates it from other situations. Semantic analysis module fills the slots with values, some slots can be left empty. Further, semantic module relates the set of sentence semantic networks of text document with corresponding template than defines a fact, using modified decision tree algorithm C4.5. according to the situation templates in the knowledge base. Factual analysis also employs statistic methods to calculate frequency/weight of target object or its connections in the document. Statistical methods are based on the calculation of TF*IDF.

## 4 «CYBER-ANALYTIC» — INFORMATION ANALYSIS SYSTEM

Applied software lab of Chelyabinsk State University has been developing text mining analytical systems since 1998. “Cyber-analytic” has indexed more than 2400 Rus-

sian and English document sources, including major media such as Vesti, New York Times, BBC, Interfax, etc.

Fact search in "Cyber-analytic" information system uses ROLAP approach because it consistently shows good performance in software solutions for financial planning, data warehousing, and Business Intelligence. OLAP cube is created from table join using star scheme, with fact table at the center of the star. Multiple tables with different indicators are joined to fact table. The system has large number of predefined analytical queries, corresponding to subject field ontology.

For the present problem — monitoring of scientific and technical information — the following indicator tables have been created:

- Organizations (detailed data);
- Persons (detailed data);
- Information sources (detailed data);
- List of priority product types;
- Key events;
- Events (detailed data);
- Product types (detailed data);
- List or priority technologies;
- Scientific and technical journals;
- Qualitative measures (frequency, impact factor, citation index, etc.);
- Topic index of the document;
- Document contexts.

«Cyber-analytic» system supports following types of analytical queries:

1. **Object query**: allows to get detailed information on queried object (organization, person, product or technology) and its connections with other objects.
   **Example**:
   Get information about Johannes Teyssen and his relations with Russian energy companies in years from 2010 to 2012.

2. **Transitive query**: allows to get information about all existing paths between two given target objects.
   **Example**:
   Retrieve all existing connection paths between companies OGK-4 and E. ON in years from 1999 to 2011. Maximum path length greater or equal 2, path should consist of objects of type person or its derivaties. Return following fact types: stock purchase, energy production, investment, energy sell.

3. **Event query**: allows to get information about events involving the object. This type of query can build a chronological or event-driven diagram of object's behavior.
   **Example**:
   get information about all events involving Wulf Bernotat in 2006 related to management of E.ON Energie A.

4. **Trigger query**: starts the monitoring of object or group of objects. Users in notification list of the trigger query get notified when new data on monitored object, its attributes and connections arrives.
   **Example**:
   Notify of any new data about connections of Johannes Teyssen with target objects of types “bank” or “insurance company” between the dates 15.03.2013 and 20.06.2013.

Analytical queries are created with the help of query creation tool, and it's important to note the the analyst can use natural language when formulating the context dependencies. Query results are displayed to the analyst in the form of analytical diagrams. “Cyber-analytic” employs i2 Analyst's Notebook software as its visualization tool. There are three types of diagrams:

1. Link Analysis charts - shows connections between persons, bank accounts, organizations, phone numbers, property and other things. We can then use the filtration of the score relation. But the frequency factor should be considered carefully, as the only fact may appear the most important in the analytical research.

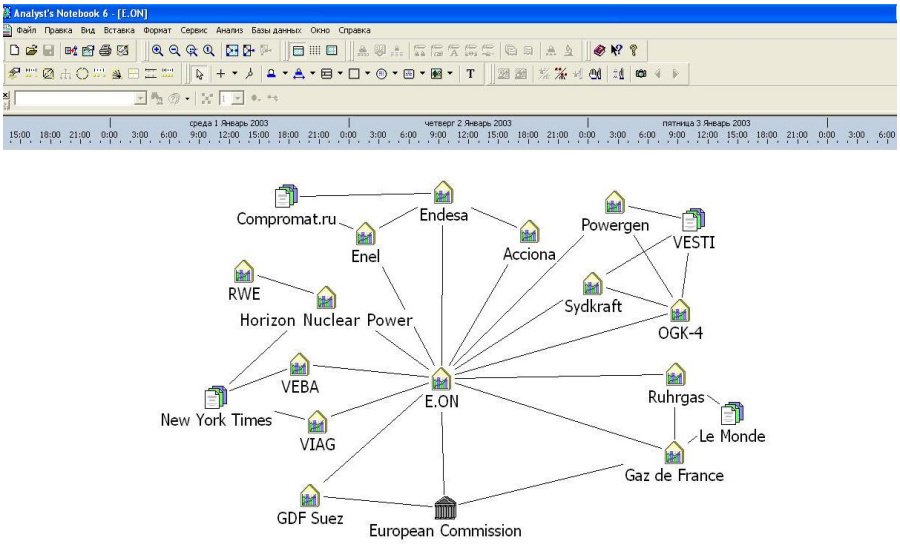


**Fig. 2.** - Link Analysis charts

2. Sequence of Events charts – diagram of this type shows the sequence of and connection between events involving target objects in chronological order, during a certain time period. Ontology allows to combine heterogeneous data into one database. Thus it makes it possible to fulfill the search in several databases. For example the results of the text search can be complemented by the results of search in database of telephone calls.

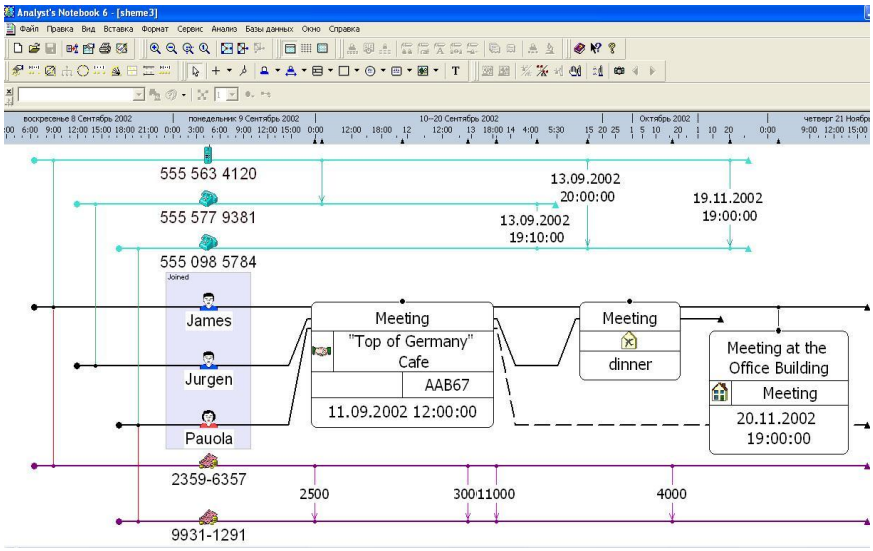


**Fig. 3. -** Sequence of Events charts

3. Transaction Pattern Analysis charts – diagram of this type help to recognize and analyze recurrent patterns of events, for example electronic mail messages, sms, phone.

The figures show the request results in the form of analytical scheme in the system of visualization i2 Analyst's Notebook. If the analyst click on the link between the objects, he receives a list of documents with the abstract where the facts about the pair of objects are revealed. In addition it is possible to get information about the attributes of the objects. After studying the structure of relationships between semantic objects analytic turns to natural-language texts containing the relevant information.

## 5 CONCLUSIONS

In this article we discussed application of ontology knowledge models to the task of monitoring objects in the context of scientific, technical, and business ingelligence. We proposed a method of factual analysis of full-text databases that allows to automatically extract situation descriptions that correspond to pre-set templates. Resulting situation descriptions are used by experts to form knowledge base in the subject field. Ontology knowledge model and method of factual analysis are implemented in "Cyber-analytic" information analysis system.

# 6 REFERENCES


1. Mel'čuk A. Actants in Semantics and Syntax. I,II, Linguistics, 2004, 42:1, 1-66; 42:2, 247—291.
2. Chomsky A. Rules and Representations. New York: Columbia University Press and Oxford: Basil Blackwell Publisher, 1980. (Excerpted in The Behavioral and Brain Sciences 3 (1980): 1-61, 1980.)
3. Manning, C., Schutze, H. Foundations of Statistical Natural Language Processing. Cambridge, MA: MIT Press., 1999.
4. Feldman, R., Anger, J. The Text Mining Handbook. New York: Cambridge University Press, 2006.
5. Bayer O., Höhfeld S., Josbächer F. Evaluation of an Ontology-based Knowledge-Management-System. A Case Study of Convera RetrievalWare 8.0 Information Services & Use 25 (2005) 181–195 181 IOS Press.
6. Vokhmintsev A., Melnikov A. Integration of proffessional knowledge bases with the help of factual analysis and ontological models // Proc. of the 13th International Workshop on Computer science and information technologies (CSIT'2011).– Vol. 1. – Garmisch-Partenkirchen, Germany, 2011. – P. 61-66.


.