Semantic Transformation of Search Requests for Improving the Results of Web Search

Vladimir A. Fomichov and Anton V. Kirillov

Department of Innovations and Business

in the Sphere of Informational Technologies
Faculty of Business Informatics
National Research University "Higher School of Economics"
Kirpichnaya str. 33, 105679 Moscow, Russia

vfomichov@hse.ru and antonv.krillov@gmail.com

Abstract

The paper describes a new method of constructing semantic expansions of search requests for improving the results of Web search. This method is based on the theory of K-representations - a new theory of designing semantic-syntactic analyzers of natural language texts with the broad use of formal means for representing input, intermediary, and output data. The current version of the theory is set forth in a monograph published by Springer in 2010. The stated approach is implemented with the help of the Web programming language Java: an experimental search system AOS (Aspect Oriented Search) has been developed.

Keywords

Semantic transformation of search request; semantic representation; theory of K-representations; SK-languages; algorithm of semantic-syntactic analysis

Introduction

Every day the amount of information stored on the Internet is considerably increased. The format of presented information is heterogeneous, and its is unstructured; most often, the information is expressed by means of natural language (NL) – English, Russian, etc. Though there are known many approaches to the search for Web-based information (Kirillov, 2009; Halpin and Lavrenko, 2009; Fomichov, 2010), finding a solution to the following fundamental problem would be very important for the design of Web search systems – the calculation of the indicator of relevancy of the found document to the search request. In the course of studying this problem, a number of different approaches for recognizing a syntactic correspondence of a document to a search request: VSM (vector-space model), the functions BM25 and BM25F (taking into account the various weight factors of the words from a document), the functions Okapi, Ponte, the algorithm LCA and other. These approaches solve the problem of syntactic search, but a semantic correspondence of the found documents to the search request is not considered.

In order to solve this problem, several formats of meta-data describing semantic components of the documents have been developed, first of all, RDF, RDFS, OWL. Semantic description of a document provides the possibility to more exactly recognize its content and respectively the relevancy as concerns a search request. However, the documents very seldom include the meta-data of the kind, therefore the meta-data can be considered as a standard in the course of developing a Web-page. Since meta-information most often is inaccessible, the focus of the methods of finding the document relevance has shifted to the analysis of information stored in a natural language form.

During last years, many systems based on semantic analysis of the contents of requests and documents have been developed, in particular, SemSearch (Lei, Uren, and Motta, 2006), AquaLog (Bernstein, Kaufmann et all, 2005), Semantic Crystal (Bhagdev, Chapman et al, 2008).

Though there are numerous approaches to the search for information on the Internet, one observes the lack of the solutions combining the following possibilities:

- semantic-syntactic analysis of natural language search requests;
- typization of the requests;
- recognizing the objects of interest of a search request;
- the search for semantic equivalents of the objects of interest of a search request;
- finding the facts reflecting achieving a certain goal by an intelligent system;
- finding the evidence of the dynamics of certain sets (Management Boards of the firms, etc.).

The selection of just this collection of the possibilities is motivated by the following factors:

- a natural language interface allows for formulating the questions being of direct interest for the user but not forces the user to select a special combination of the words for successful syntactic search;
- the possibility to obtain the most complete collection of relevant information describing various aspects of the system's behavior, its state and achievements (or failures) of an intelligent system (including the organizations).

Central Ideas of the Proposed Solution

This paper proposes a solution optimizing the work of traditional search systems by means of semantic analysis and expanding the natural language input requests. Taking into account the calculating power of the biggest existing systems fulfilling the key words based search, it is proposed to shift the focus from the detailed semantic analysis and indexation of the content of electronic documents to the analysis of the inputted search requests and generation of a set of semantically expanded (adapted) requests that will be transmitted to a syntactic search system. The results of the search corresponding to each request from this semantically expanded set will be analyzed and compared with the aim of increasing semantic relevancy of the search results.

Example. Suppose that a user-businessman would like to get a certain information about the company X in order to consider the possibility of starting a collaboration with this company. In this connection, the questions about the achievements of the company during last year would be quite natural. For instance, the user may ask the questions “What achievements did the company X have last year?” or “What failures did the company X have last year?”.

Both questions belong to the class of questions about the result of achieving a goal. Imagine that, as a result of the search, the user has received the information about the launch by the firm X of a new product or service Y. Correspondingly, the user would like to get the information about some characteristics of the product or service Y and, besides, about some distinguishing features of Y. The examples of the questions may be as follows: “What are the characteristics of the product Y?” and “What are the peculiarities of the product Y?”. The questions of this class will be called below *aspect-oriented questions*, and their processing will be considered in more detail.

Finally, having received the mentioned information, the user-businessman wants to get to know about the stability of the Board of Directors of the company X. For instance, he/she may formulate the question “What were the changes in the Board of Directors of the company X during last year?”.

With respect to the progress of voice interfaces and the computer means of synthesis and analysis of spoken speech, the process of looking for this information can be represented by the following dialogue:
User: “What achievements did the company X have last year?”.
System: “The company X launched the product Y, showed the benefit increase of 7%, and started a new office in Moscow”.
User: “What are the peculiarities of the product Y?”.
System: “High refusal stability and low price”.
User: “What are the distinctions of the product Y from the product Z?”.
System: “The product Y exceeds the product Z as concerns the following indicators:___”.
User: “What were the changes in the Board of Directors of this company during last year?
System: “Peter Stein entered the Board of Directors”.

Thus, if a user wants to find information about a company, its achievements and failures, the launched products, various characteristics of the products, and about stability of its Board of Directors, then the complete process of search is covered by the proposed classes of questions and corresponding methods of search requests transformation. In this way, the speed, convenience, and relevancy of search will be increased.

The Method of Searching for Information of Interest

Let’s consider a method of looking for the information being of interest for the user under the framework on the proposed approach. A generalized algorithm consists of five main steps, two o f them are unique for ach of the considered types of questions.
Step 1. The inputted search request is analyzed for finding its type. It is necessary to distinguish the primary and secondary objects of interest of the search request W. Suppose that the request “What achievements did the company Intel have in the year 2010?”. Then primary object of interest is W1 = “achievements”, and secondary object of interest is W2 = “the company Intel”. The object W1 enables us to classify the search request W as an element of the class of questions about achieving a goal.

Step 2. After finding the type of the request it is possible to go to creating a set of secondary search requests generated by the request W, that is, to forming a semantic expansion of the inputted request. The construction of the semantically expanded set of requests is being fulfilled with the help of a knowledge base containing the information needed for forming new requests.

Step 3. As soon as the expanded set of requests has been formed, it is transmitted to the traditional search system, the latter returns a set of documents which syntactically correspond to the generated requests. Dependent on the preferences of the user, i.e. dependent of the user's behavior and selection of certain results of the search, the weights of the substitutions and the order of generating the requests (during the previous step) will be calculated.

Step 4. The documents received from the search system are analyzed and filtered with the help of a knowledge base (in order to calculate the number of occurrences of the indicators of interest in the document) and with the help of the indicators of documents' syntactic relevancy (the documents having the values of these indicators below a certain border will be excluded as non-relevant). The indicators will bed understood as such natural language expressions that their occurrence in the text allows for judging about the correspondence of the document to the initial search request. First of all, the documents with the big amount of duplications will be considered. The reason is as follows: if a document more often occurs in the results of search proceeding from different requests, this document contains more indicators and, hence, contains more information corresponding to the initial request.

Step 5. The analyzed and filtered documents are then returned to the user.

Three classes of natural language questions are considered under the framework of the proposed approach, and the questions from these classes require certain speculations for constructing a semantically expanded set of search requests. These three classes of questions are (a) the questions concerning the achievement of a certain goal; (b) aspect-oriented questions, (c) the questions about the dynamics/stability of the sets (for instance, about stability of the Management Board of a certain company). Let's consider in more detail the methods of processing the search requests from the first class.

Processing of Questions about Achieving a Goal

We will say about the questions about achieving a goal in case of interrogative questionns where one asks about information reflecting the results of functioning of an object, a system. In other words, these are *the questions about the achievements and failures*.

The success of functioning (or existing) of an object or a system is determined by achieving by the considered entity of the formulated goals. By a goal of a company we'll understand the final desirable result that is set in the process of planning and is regaled by the control functions. An example of the questions about achieving a goal is as follows: "What failures experienced the company Sun in the year 2010?".

For fulfilling a detailed analysis of questions about achieving a goal, we've selected, studied, and divided into several groups the goals associated with the activity of the enterprises. The examples of such goals are as follows: "The launch of a new product", "Starting a new office by a company", "The increase of benefit", "The absorption of a company".

The data of the kind should be stored in a special knowledge base, it will be called *a goal base*. This base is used for the generation of natural language expressions showing the availability in the documents of the information about success. The goal base is formed with the help of the theory of K-representations.

It is a new theory of designing semantic-syntactic analyzers of NL-texts with the use of formal means for representing input, intermediary, and output data is proposed (Fomichov 2010). This theory can be interpreted as powerful and flexible tool of designing the NL-interfaces to applied intelligent systems. The structure of this theory is as follows.

The *first basic constituent* of the theory of K-representations is the theory of SK-languages (standard knowledge languages). The kernel of the theory of SK-languages is a mathematical model describing a system of such 10 partial operations on structured meanings (SMs) of natural language texts (NL-texts) that, using primitive conceptual items as "blocks", we are able to build SMs of arbitrary NL-texts (including articles, textbooks, etc.) and arbitrary pieces of knowledge about the world.

The analysis of the scientific literature on artificial intelligence theory, mathematical and computational linguistics shows that today the class of SK-languages opens the broadest prospects for building semantic representations (SRs) of NL-texts (i.e., for representing meanings of NL-texts in a formal way).

The expressions of SK-languages will be called the K-strings. If T is an expression in natural language (NL) and a K-string *E* can be interpreted as a semantic representation T, then *E* will be called a K-representation (KR) of the expression T.

The *second basic constituent* of the theory of K-representations is a broadly applicable mathematical model of a linguistic database. The model describes the frames expressing the necessary conditions of the existence of semantic relations, in particular, in the word combinations of the following kinds: "Verbal form (verb, participle, gerund) + Preposition + Noun", "Verbal form + Noun", "Noun1 + Preposition + Noun2", "Noun1+ Noun2", "Number designation + Noun", "Attribute + Noun", "Interrogative word + Verb".

The *third basic constituent* of the theory of K-representations is a complex, strongly structured algorithm carrying out semantic-syntactic analysis of texts from some practically interesting sublanguages of NL. The algorithm *SemSynt1* transforms a NL-text in its semantic representation being a K-representation (Fomichov 2010). The input texts can be from the English, German, and Russian languages. That is why the algorithm *SemSynt1* is multilingual.

An important feature of this algorithm is that it doesn't construct any syntactic representation of the inputted NL-text but directly finds semantic relations between text units. The other distinguished feature is that a complicated algorithm is completely described with the help of formal means, that is why it is problem independent and doesn't depend on a programming system. The algorithm is implemented in the programming language PYTHON.

The formation of a goal base is semi-automated. The first step consists in processing a special representation of a goal with the help of the algorithm *SemSynt1* described in (Fomichov 2010). For instance, the knowledge engineer inputs the sentence S1 = "#The company X# absorbs the company Y". Here the marker # is used for distinguishing such entity that its collection of goals includes the goal described in the considered sentence.

As a result of semantic interpretation of the sentence S1, the following K-representation *Semrepr1* of S1 will be constructed:

$$\mathit{(Situation(e1,\ absorption1 * (Agent2,\ certn\ company1 *(Name1,\ X) : z1)}$$
$$\mathit{(Dependent\text{-}org,\ certn\ company1 *(Name1,\ Y) : z2) \wedge (z1 \equiv Ob\text{-}intr)),}$$

where the variable *Ob-intr* is interpreted as the designation of an object of interest in the future search request.

Then the knowledge engineer constructs an expanded expression

$$\mathit{<(Situation(e1,\ absorption1 * (Agent2,\ certn\ company1 *(Name1,\ X) : z1)}$$
$$\mathit{(Dependent\text{-}org,\ certn\ company1 *(Name1,\ Y) : z2)) \wedge (z1 \equiv Ob\text{-}intr)),\ +1>,}$$

where the symbol +1 indicates that the truth of the sentence S1 reflects the achievement of a goal of the company X.

The K-string *Semrepr1* is used for constructing the pattern

{org} [absorption1] (verb) {org} .

The success of comparing this pattern with a document will be achieved in case when this document includes a distributed word combination *A B C*, where *A* and *C* are the lexical units associated with arbitrary concretizations of the semantic unit *org* (organization), B is a lexical unit associated with the semantic item *absorption1.*

The descriptions of the achievements and failures (let's call them the facts) are stored in the goal base and are used for the generation of word combinations being the indicators of the document fragments mentioning these achievements or failures. Consider in more detail a method of transforming a fact into a word combination – indicator.

The construction is being fulfilled with the help of the transformation rules being unique for each fact. A transformation rule indicates the order of the words in the word combination and the forms of combinations. These combinations will enable a traditional search system realizing the search on key words to find all documents mentioning the relevant facts. The collection of the documents returned by a search system will be analyzed from the standpoint of calculating the quantity of occurrences of various combinations – indicators, that is, the indicators of a reference in the document to a fact. The relevance of a document will be determined, firstly, by the quantity of occurrences of various facts and secondly – on the rating of a document calculated in accordance with the algorithm PageRank. The documents sorted with respect to its relevance to the initial search request will be transmitted to the user.

The stated approach is implemented with the help of the Web programming language Java: an experimental search system AOS (Aspect Oriented Search) has been developed. Now the system AOS is being tested.

References

[1] Bernstein, A., Kaufmann, E., A. Gohring, A. and C. Kiefer (2005); Querying Ontologies: A Controlled English Interface for End-users. In 4th International Semantic Web Conference (ISWC 2005), pages 112– 126, November 2005 (pp. 112-126)

[2] Bhagdev, R., Chapman, S., Ciravegna, F., Lanfranchi, V. and D. Petrelli (2008); Hybrid search: Effectively combining keywords and semantic searches. In The Semantic Web: Research and Applications; Springer, Berlin / Heidelberg (pp. 554–568)

[3] Fomichov, V.A. (2010); Semantics-Oriented Natural Language Processing: Mathematical Models and Algorithms; Springer, New York, Dordrecht, Heidelberg, London (354 pp)

[4] Halpin, H. and Lavrenko, V. (2009); Relevance Feedback Between Hypertext and Semantic Search; Proc. International Conference WWW2009 (April 20-24, 2009, Madrid, Spain).

[5] Kirillov A. (2009); Search Systems: Components, Logic, and Methods of Ranging; Business-informatics (Moscow), No. 4 (10) (pp. 51-59)

[6] Lei, Y., Uren, V. and E. Motta (2006); Semsearch: A search engine for the semantic web. In Proc. 5th International Conference on Knowledge Engineering and Knowledge Management Managing Knowledge in a World of Networks, Lect. Notes in Comp. Sci., Springer, Podebrady, Czech Republic (pp. 238-245).