Intelligent Search System for Huge Non-Structured Data Storages with Domain-Based Natural Language Interface

Artyom Chernyshov, Anita Balandina, Anastasiya Kostkina and Valentin Klimov National research Nuclear University "MEPhI", Moscow, Russian Federation

a-chernyshov@protonmail.com, anita.balandina@gmail.com, anakost@bk.ru, vvklimov@mephi.ru

Abstract. Nowadays the number of huge companies and corporations has in their disposition various non-structured texts, documents and other data. The absence of clearly defined structure of the data makes the implementation of searching queries complicated and even impossible depending on the storage size. The other problem connected with staff, which may face the problem with misunderstanding of the special query languages, knowledge of which is necessary for the execution of searching queries. To solve these problems, we propose the semantic search system, the possibilities of which include the searching index construction, for queries execution and the semantic map, which would help to clarify the queries. In this paper we are going to describe our algorithms and the architecture of the system, and also to give a comparison to analogues.

Keywords: Semantic search, Semantic map, Non-structured data, Natural language, Domain-based Natural languages.

1 Introduction

Natural Language Understanding (NLU) is a set of tasks considered from the point of view of the semantic knowledge, which occurs in the natural language. NLU is one of the main problems in the area of natural language processing (NLP)).

NLP is a scientific direction in the area of artificial intelligence, which deals with the automated processing and understanding of the natural language. It is unofficially considered that the task of NLU AI is a complete task, which means that the complexity of solving this problem is as high as the solution of other central problems of artificial intelligence or how the creation of computers capable of thinking like a human.

The goals of the NLU are still far from reaching. One of the main tasks in modern NLP systems is the consideration of features and the ambiguous nature of the natural language. Despite this, recent studies have made significant progress in solving important NLU subtasks, such as syntactic analysis of dependencies and weak semantic analysis.

A wide range of applications, such as question-answer type systems, automatic abstraction or retrieval of information, can potentially benefit from the NLU study. In this paper, we propose the usage of syntactic and semantic analysis methods for natural language texts to construct a query structure that will be translated directly into query languages of specific database management system.

Generally, the question-answering system (QA-systems) may be divided into two parts: the closed-domain systems and open-domain systems.

The first ones give answers into a specific domain (for example, nuclear engineering), and can be considered as an easier task because NLP systems can exploit domain-specific knowledge frequently formalized in ontologies. Alternatively, closed-domain might refer to a situation where only a limited type of questions is accepted.

The other ones work with questions about almost anything, and can only rely on general ontologies and world knowledge. On the other hand, these systems usually have much more data available from which to extract the answer. But on the other hand the size of general ontology may negatively affect the accuracy of the answer.

In this case we can only talk about closed-domain QA-system, because the goal of our system is giving answers in a specific domain as relevant as possible. However, the domain may vary depending on the customer needs by modification of one of the components of the system. In this way the NLU task may be simplified, as we only work with domain-based natural language. Further, there will be explained the methods of the construction and rectification of searching queries which we are going to use in our system.

2 Natural Language Understanding and Processing

For the representation of the internal structure of the input request it is proposed to use the syntactical dependencies tree. In this case we use the part-of speech tags or POS-tags to mark the words of input sentence. The main word, normally verb, which is also called predicate is mark as the root of the tree. The other words of the input request are connected to the root with POS-tags. In this case, we can consider the nodes of the tree define the constraints of the query, and the leafs are parameters of the query which specify the certain conditions.

Formally, the input request model can be represented as follows:

$$I=<W,C>\text{, where} \tag{1}$$

- $W = \{w_1, \dots, w_n\}$ – is the set of words of the input request;
- $C = \{c_1, \dots, c_{n-1}\}$ – is the set of links between the words of the input query.

In turn, the sets W and C can be represented as follows:

$$w_i =< nf, pos, gr\_cat >\text{, where} \tag{2}$$

- Nf - normal form (infinitive for verbs, singular number of nominative case for nouns, etc.);
- Pos – part-of-speech tag (subj, dobj, etc.);
- Gram_category - grammatical categories or parameters (gender, number, case for nouns, form and time for verbs, etc.).

3 Rectification of Queries with Semantic Mapping

Rectification of Queries (RoQ) is the process of reformulating a seed query to improve retrieval performance in information retrieval operations [1]. Rectification of queries involves techniques such as: i) Finding synonyms of words, and searching for the synonyms as well; ii) Finding all the various morphological forms of words by stemming each word in the search query; iii) Fixing spelling errors and automatically searching for the corrected form or suggesting it in the results; iv) Re-weighting the terms in the original query [2].

The most part of the existing approaches, algorithms, techniques for rectification of queries are based on using WordNet. WordNet has a database that groups the words into sets of synonyms called synsets and provides definitions, comments, examples of usage of these words, and the actual meaning in each case. Therefore, it combines the elements of a dictionary (definitions and some examples) with those of a thesaurus (synonyms), resulting in an important support for the automatic analysis of text and words.

However, in Voorhees[3] the use of WordNet for performing rectification of queries did not actually increase the effectiveness of the information retrieval process without specific expansions and limitations. The numerous attempts to improve approach of thesaurus-based query expansion are mentioned in [4] and [5]. In [5] the improvement is the concept of the vector space model under WordNet and the computation of similarity of documents as terms.

It is considered that the vector space model gives the significant improvements in the case if we have a deal with analysis of documents. Here we should introduce the concept of a semantic map. In general, a semantic map can be defined as a topological or metric space, the topology and/or the geometry of which reflect semantic characteristics and relations among a set of representations (such as words or word senses) embedded in this space [6,7].

In the present time there are two types of semantic map. The first type named “strong semantic map” is based on the vector space model and the concept of dissimilarity (LSA, LDA) and we’ve mentioned it above. The second type is “weak semantic cognitive mapping” and consist in using such notion as “opposite relations”. Here it doesn’t take into account individual semantic characteristics of representations given a priori. Only relations, but not semantic features, are given as input. As a result, semantic dimensions of the map that are not predefined to emerge naturally, starting from a randomly generated initial distribution of words in an abstract space with no a priori given semantics and following [7]. The main relation that authors use in this approach is the relation “synonymy-antonymy” for representations. The most known approach of this type is Antomap developed by A.Samsonovich.

The disadvantages of the mentioned approaches based on “strong semantic map” are that they are not applied for some goals, for example, if it’s necessary to work with the narrow domain areas and graph databases, not documents. Our approach implements the synonym-based query expansion with the help of weak semantic map Antomap, which uses Microsoft Russian Thesaurus Core as a part of WordNet. //the ontology that will consist of concepts defined by the domain area.

The Microsoft Russian Thesaurus Core (MRTC), obtained in [7,8], represents a dictionary cluster of the Microsoft Russian Thesaurus with the major number of connections between words. There are no contradictory connections and duplicates inside it. In fact, the MRTC represents a cluster of the "strongest" relations of a synonimy and an antonymy. It contains the words which have not less than 2 and no more than 11 synonyms.

The Antomap is a superstructure over MRTC that allocates words with their connections on the map. The formal definition of Antomap consists in considering of that cognitive semantic map is represented as the dynamic system which is formed by N points in the space or vectors xi $\in \Re$ or, in other words, is a distribution of words in an abstract vector space (with no semantics preassociated with its elements or dimensions) that minimizes the following energy function [7,8]:

$$H(X) = -\frac{1}{2}\sum_{i,j=1}^{N} W_{ij} x_i \cdot x_j + \frac{1}{4}\sum_{i=1}^{N} \|x_i\|^4, \text{ where} \tag{3}$$

x_i is a 26-dimensionalvector that is representing the i^{th} word (out of N). The Wij entries of the symmetric relation matrix equal +1 for pairs of synonyms, –1 for pairs of antonyms, and zero otherwise. The energy function (1) follows the principle of parsimony: it is the simplest analytical expression that creates balanced forces of desired signs between synonyms and antonyms, preserves symmetries of semantic relations, and increases indefinitely at the infinity, keeping the resultant distribution localized near the origin of coordinates. More details of this approach are described in [7] and [8].

The structure of the current map allows to compute a semantic similarity (SS) between words. The computation of SS is connected with the "the contextual quality" of the chosen synonyms and helps to dispose from synonyms which aren't close in the context with the original word. The rest of synonyms expands the query by forming new and similar queries. As for necessary components for SS, the main semantic dimensions of this map are used for computation of it and defined by the principal components of the emergent distribution of words on the map. For example, semantics associated with the first three PCs can be characterized as "good" versus "bad" (PC1), "calming, easy" versus "exciting, hard" (PC2), and "free, open" versus "dominated, closed" (PC3) [7,8]. These and next three semantic PCs are the most important for finding of "the strongest" synonyms for one or another word, despite the common number to equal 26 PCs.

The basis of our approach of computation of a semantic similarity is Weighted Euclidean distance in six dimensional space. The weight is the relation 1 to the squared standard deviation of every PC. The common formula is (n=6):

$$\boldsymbol{sim}(x,y) = \left(\sum_{i=1}^{n} w_i (x_i - y_i)^2\right)^{\frac{1}{2}} \tag{4}$$

In the present time the algorithm implies usage of the Microsoft Russian Thesaurus for finding all synonyms and only after it the list of synonyms is filtered by the Antomap. The algorithm is as follows:

Step 1. The input is the Token (word) X. The synonyms in the thesaurus are searched until the end is reached. In the process of searching all existing synonyms of the current word are extracted and entered in the list of synonyms. The size of the list is L, i.e. the list consists of L-synonyms.

Step 2. On the semantic map the Token and its synonyms based on the MTRC are randomly allocated. The Token and its synonyms represent 26-dimensional vectors in space. The first six coordinates are selected. The list of pairs is formed (word, vector coordinates).

Step 3. In turn a semantic similarity of the Token is calculated with each of the synonyms for each six components. The value of the semantic similarity is obtained with the common formula represented above. The negative values and zero values are not taken into account here, and the lower threshold of semantic similarity is set, i.e. sim (x, y) ≥ 0.5. The upper threshold is set to 1.If the semantic similarity of the Token with a synonym lies in the interval (0, 0.5), i.e. tt is less than the defined value for lower threshold, then the current synonym is “put aside”.

Step 4. A list of all pairs (synonym, semantic similarity) sorted by descending semantic similarity is formed.

$$c_i =< (w_i, w_j) >, \text{ where} \tag{5}$$

w_i, w_j - pairs of words related to each other by an oriented relationship, the beginning of which is in the first word, and the end in the second.

This method of representing the input request is extremely convenient, since the structure stores information about each word of the sentence. Thus, a good opportunity to find synonyms will be provided, as well as a format for issuing the results of the query.

4 System Description

The authors propose to use the described above algorithm in project related to semantic search over huge unstructured documents and other data.

Intelligent search engines are the natural continuation of the development of conventional search engines. In comparison with their predecessors, their capabilities can include many different features and functions, such as:

- It is necessary to obtain documents in some way associated with a specific considered document.
- Search for documents, which exact attributes are unknown, but can be formulated a query in natural language that can characterize them.
- It is necessary to compare two or more documents by their meaning.

Authors propose the developing an intelligent search system with an interface in a domain-specific natural language. Such an interface will help to get rid of the need to use special query languages like SQL. The user will be able to formulate a search query in natural language using vocabulary adopted in a particular subject area (for example, using terms from nuclear physics).

A distinctive feature of the developed system is its flexibility and ability to work in different subject areas, using only the vocabulary and terminology that is adopted in a particular area (or areas) of the customer. Flexibility and ability to work in different subject areas will be achieved through the usage of semantic maps - a tool for comparing commonly accepted terminology and vocabulary adopted in a specific subject area, which in turn is built by using the synonymy relationship on which the semantic map itself is based. The algorithms for processing natural language algorithms are based on the methods of constructing the syntactic trees of dependencies, by which constructs query structure, that is suitable for further translations into specific program queries.

The developed system will be designed to facilitate the search process in subject areas with extensive databases, or large volumes of unstructured documents, and has no limitations in the field of application. The system will not depend on the language, so it will be enough to simply enter the international market. To introduce a prototype system into a specific area of activity, certain semantic maps will be created that take into account the specific vocabulary of this particular area. The need to modify one single component for the full-fledged operation of the system in any subject area makes the developed complex flexible and multipurpose.

5 Conclusion

In the area of semantic search at the moment there are a lot of gaps and unsolved problems, such as highlighting the meaning of words and sentences, searching in different subject areas, communicating with the user in a language that is close to natural (using specialized vocabulary).

These tasks can be considered practically solved for certain industries and spheres of activity. For example, in the field of medicine, there are lots of powerful tools and knowledge bases that allow you to receive answers to various questions based on knowledge stored in the database. An example is the IBM Watson supercomputer, which works well in the field of medicine, but it is rather weakly applicable in other subject areas.

Our solution is to use a combination of such technologies as machine learning, semantic mapping and ontologies. One of the important tasks is to improve the accuracy of query processing, in which a term-specific terminology is used. For the processing of such terminology and the synonyms highlighting both at the level of words and at the level of whole phrases it is proposed to use semantic maps. This technology has received some distribution abroad, but in Russia the work on this topic began relatively recently and is conducted at a rather slow pace, nevertheless, there are already exist some algorithms that allow building semantic maps for the Russian language. Within this project, the algorithm will be finalized and adapted for working in the area of technical documentation.

Described queries rectification algorithm with semantic mapping shows practical application of semantic maps. That will allow to increase the theoretical significance and show practical importance of semantic maps.

6 Acknowledgements

This work was supported by Competitiveness Growth Program of the Federal Autonomous Educational Institution of Higher Professional Education National Research Nuclear University MEPhI (Moscow Engineering Physics Institute). The funding for this research was provided by the Russian Science Foundation, Grant RSF 15-11-30014.

References

1. D. Abberley, D. Kirby, S. Renals, and T. Robinson. The THISL broadcast news retrieval system (1999). In Proc. ESCA ETRW Workshop Accessing Information in Spoken Audio, (Cambridge), 14 – 19.
2. Leung, C. H. C., et al. Collective Evolutionary Concept Distance Based Query Expansion for Effective Web Document Retrieval (2013). In Proceedings of the 13th International Conference on Computational Science and Its Applications (ICCSA-2013), LNCS, 657-672.
3. Voorhees, E. M. Query expansion using lexical-semantic relations (1994). In Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 61–69.
4. Navigli, Roberto, and Paola Velardi. An analysis of ontology-based query expansion strategies (2003). Workshop on Adaptive Text Extraction and Mining, held in conjunction with ECML 2003, Cavtat Dubrovnik, Croatia, September 22.
5. Francisco Joao Pinto et al. Joining automatic query expansion based on thesaurus and word sense disambiguation using WordNet (2009). International Journal of Computer Applications in Technology, Volume 33, 271-279.
6. Klimov V., Chernyshov A., Balandina A., Kostkina A.. A new approach for semantic cognitive maps creation and evaluation based on affix relations (2016). FIERCES on BICA, 99-105.
7. Samsonovich A., Ascoli G.. Augmenting Weak Semantic Cognitive Maps with an ‘‘Abstractness’’ Dimension (2013). Hindawi Publishing Corporation Computa-tional Intelligence and Neuroscience, 10 pages.
8. Samsonovich A. V., Ascoli G. A. (2010). Principal semantic components of language and the measurement of meaning. PLoS One, 5(6), 1-17.