ANALYSIS OF EXPERT STATEMENTS IN SEARCH SYSTEMS[1]

G.S. Lbov, N.L. Dolozov, P.P. Maslov
Institute of Mathematics (SB RAS), Novosibirsk State Technical University
Novosibirsk, RUSSIA
e-mail: lbov@math.nsc.ru, dnl@interface.nsk.su, alter@erasib.ru

Abstract

The work described in this paper, deals with textual data analysis application. The proposed approach which, is using statements coordination principle, is implemented to the described search system to improve accuracy of the search. This method allows to compile an ordered list of answers to the inquiry in the form of quotations from the document.

1 Introduction

In general, existing search systems are based on the analysis of index databases taking into account morphology; however they do not analyze semantic structures of sentences and interrelations of sentences in the document.

In this paper an approach of creating of the search system, based on the analysis of logic structures and interrelations of sentences in the document is offered. This approach improve search accuracy in proposed system.

The specified approach defined by criterion of selection of significant sentences of documents, based on accordance to a certain logic structure reflecting the sense of inquiry, improve search accuracy and allow revealing relevant documents in the order of inquiry degree accordance to the documents

2 Text Processing

The specified approach has been realized programmatically in search system Internal Search System3 further ISS3 [1] (In the creation of ISS3 technologies of known system Dialing [2] have been used), providing search service of documents on local and sharer network resources.

Offered system ISS3 is realized as two modules:

The module of the natural language processing (NLP). To perform of the subsequent procedure of the coordination of the statement, the developed system represents an input text as a sets of semantic relations for each sentences of the text. It is achieved by the multilevel natural language processing realizing initial, morphological, fragmentational, syntactic and superficial semantic analyses of input text documents.

The brief description of analysis stages of natural language (NL):

[1] This work was financially supported by RFBR-07-01-00331

1. Initial text analysis process input text and then forms two separate tables. The first describes elements of the text and their arrangement in the text, and the second defines interrelation of fragments in the input text.

2. At a stage of the morphological analysis, for each lexeme the set of lemmas with attributes is created. Each lemma is represented as a normal form of a word, and attributes - a set of descriptors (a part of speech, number, case etc.).

3. Fragmentational analysis performs derivation of fragment from the input text. Fragments are the main and dependent clauses in structure of complex one, participial, adverbial participial and other isolated turns.

4. The purpose of syntactic parse - automatic construction of a functional tree of a phrase. Syntactic parse process an outcomes of morphological and fragmentational analysis.

5. At a stage of the semantic analysis the semantic relations describing certain binary relations between dependent and operating members are formed. These binary relations are used in the subsequent algorithm of the coordination of statements.

Formed semantic graph of the sentence characterizes the interrelated binary relation in sentences of the input text.

As a result NLP forms two separate tables for each input document:

1) The content of the document. The table of semantic relation sets for each sentence of the input document, rows in which describe type and components of semantic relations.

2) Structure of the document. The table containing the descriptions of structure of the document (paragraphs, headers, etc.), derived at a stage of the fragmentational analysis and necessary to form the reply to inquiry.

The module of the coordination of expert statements. Basing on outcomes of NLP module the logic form is constructed for each sentence. This form is a model in the language of predicates calculus of two variables united in conjunctions. Each of such predicates is an elementary statement. Let X_i, Y_i, Z_i etc. are each predicate variable. As the predicate defines the relation between its variables, the sets of the one-type variables standing in a certain position in the predicate are designated by the same letter with different indexes.

On the basis of the derived models of sentences of the text and inquiry procedure of the coordination of statements in models is performed. For this purpose from each model the same predicates are derived. The set (for each type of a predicate), corresponding sentences of the text is a variety of coordinated statements, and a set corresponding inquiry is beforehand coordinated statement. By quantity of the coordinated predicates the level of its accordance to inquiry is defined, being based that each predicate is a part of model of the sentence.

3 Coordination of statements

Let some statement with known characteristics requires to define its accordance to inquiry [3]. The general formal writing of a sentence is done in the form of two-place predicates conjunction. We shall designate T_{ji}^{k} as area of the validity of function and argument variables in the initial sentences inquiry, where i, j, k are the numbers of predicates, statements and the links between argument and function variables, respectively. As variables are nominal the area of true statements is defined by variables satisfying the list of admissible values. As such list the dictionary of synonyms is used.

As predicates two-place and their variables are defined on different areas of the validity for the coordination of statements, it is necessary to consider variables in predicates separately.

For each predicates containing in the statement it is defined areas of the validity: T^1_{pi} is a truthful area of the first variable in the predicate i the inquiry p; T^2_{pi} is the same for the second variable. Let's designate T^1_{ji}, T^2_{ji} truthful areas of variables in predicates of the input text. Respectively, the statement satisfying:

1. $\frac{\mu(T^2_{ji} \cap T^2_{pi})}{\mu(T^2_{ji} \cup T^2_{pi})} \geqslant \beta_{r2}$ and $\frac{\mu(T^1_{ji} \cap T^1_{pi})}{\mu(T^1_{ji} \cup T^1_{pi})} \geqslant \beta_{r1}$ - true
2. $\frac{\mu(T^2_{ji} \cap T^2_{pi})}{\mu(T^2_{ji} \cup T^2_{pi})} \geqslant \beta_{r2}$ and $\frac{\mu(T^1_{ji} \cap T^1_{pi})}{\mu(T^1_{ji} \cup T^1_{pi})} < \beta_{r1}$ - not likely
3. $\frac{\mu(T^2_{ji} \cap T^2_{pi})}{\mu(T^2_{ji} \cup T^2_{pi})} < \beta_{r2}$ - contradictious

Where β_{rq} is a parameter. Thus, sets of statements are derived: true, improbable and denying.

To define the accordance of sentence to inquiry it is necessary to calculate ratio k, changing from 0 up to 1 where 0 corresponds to absence of accordance of semantic structures of the sentence and inquiry, and 1 - their full hit.

$\kappa = \frac{(N^i_{so})^2}{N^i_s N_r}$

Where N^i_s is number of all predicates in a sentence, N^i_{so} is number of the coordinated predicates of a sentence, N_r is number of predicates of inquiry.

4 Conclusion

In the paper the algorithm and its program realization for performance of search in the text documents, improving search accuracy, based on the analysis of semantic structure of sentences is offered. As a result of performance of algorithm sets of the coordinated statements for all types of predicates are formed, each of which describes the certain sentence. To define conformity of the sentence to inquiry the ratio specified above is calculated. The sentences derived by algorithm taking into account paragraphs and headings of documents formed a result in a natural language.

References

[1] P.P. Maslov. Proceedings of All-Russian scientific conference of young scientists in seven parts. Novosibirsk: NGTU, 2006. Part 1. - 291p. // pp. 250-251

[2] Automated Text Processing // www.aot.ru

[3] G.S.Lbov, T.I. Luchsheva. The analysis and coordination of expert knowledge in problem of recognition // 2'2004, NAS of Ukraine, pp. 109-112