SELECTION OF THEMATIC NL-KNOWLEDGE FROM THE INTERNET

V.Gladun, A.Tkachev, V.Velichko, N.Vashchenko

Abstract: The paper deals with methods of choice in the INTERNET of natural-language textual fragments relevant to a given theme. Relevancy is estimated on the basis of semantic analysis of sentences. Recognition of syntactic and semantic connections between words of the text is carried out by the analysis of combinations of inflections and prepositions, without use of categories and rules of traditional grammar. Choice in the INTERNET of the thematic information is organized cyclically with automatic forming of the new key at every cycle when addressing to the INTERNET.

Keywords: semantic analysis, information search, INTERNET.

1. The purposes and base ideas

Among various variants of practical use of storehouses for the textual information the necessity to find the information having thematic unity prevails. These are needs of a scientist, a journalist, a politician, an official, a writer, a student. Usually the theme arises as one or several concepts, some initial situation having a number of blank valences and situational roles which serve as reference points for search of the new relevant information. The new information gives rise to new directions of search. This complex, sometimes psychologically painful, creative process requires the automated support. Thematic search needs laborious work with the texts stored in libraries, archives, the INTERNET, textual databases. Difficulty of this work consists, in particular, in necessity to select not the whole texts, but relevant to the theme fragments of texts. The contents of many texts are an interlacing of a number of themes. Thus, a problem arises of search inside textual documents of fragments, relevant to the given theme.

In the paper methods, software and results of selection of the thematic textual information are considered. The researches submitted in the paper continue the works published in [1-3].

The solving of the problem unites the following actions:

1) selection of texts or fragments of texts relevant to an investigated theme;
2) selection from the relevant information of the most important, first of all, such which defines and connects the most essential terminology of a theme;
3) representation of the chosen information in the user-friendly form.

Implementation of the specified actions is based on the following ideas:

1) to focus a technique of selection of the thematic textual information on the INTERNET, as on the most full storehouse of the textual data;
2) to combine search by key words with the semantic analysis of NL-texts;
3) to use semantic criteria for selection of the most important thematic information;
4) to organize automatic cyclic process of key words formation to investigate the theme as complete as possible.

2. A technique

The initial stage of the thematic information selection consists in search of textual documents in the INTERNET using the given key. Existing methods of information search in the INTERNET give out a lot of unnecessary for user “garbage” information which filtration takes too much time. The way out consists in use of the semantic criteria providing selection of the most essential characteristics of concepts concerning which the information is gathered.

The offered method is based on the assumption, that the most important user information is contained in *kernel constructions* of sentences. The term “kernel constructions” is used in transformation grammar for designation of simple base judgment by which transformation the sentence as a whole is formed. In our case the kernel construction consists of a subject, a predicate and a link.

The method represents cyclically repeating sequence of the following operations:

1. Selection of the given quantity (parameter) of texts using a key. A set of used search systems is unlimited. Now the program can use the following search systems: Yandex, Rambler, Meta-Ukraine, Aport, Google.
2. Selection in the found texts of the sentences containing a given key.
3. Selection in set of the sentences that were chosen in item 2, the sentences containing kernel constructions. For item 3 performing the natural-language semantic analyzer is used.
4. Formation *of n-step expansions of the kernel* of the selected sentences. *n*-step expansion of the kernel is a part of the sentence containing its kernel, and also the words connected in a tree of dependencies with elements of the kernel by paths which length does not exceed *n. n* is a user-given parameter.

The item 4 is performed on the basis of the semantic analysis of the sentence.

5. Selection in the set of the sentences chosen in item 3, such sentences in which *n*-step expansions of the kernel contain the given key.
6. Formation of a new key on the basis of the analysis of semantic representations of before selected sentences. Transfer to the item 1.

The initial key word is given by a user. New keys on the subsequent cycles of the algorithm are chosen among *terms* that are significant words used only within the limits of investigated domains. The terms are marked in the dictionary.

When choosing a new key, the degree of its relevance to the given theme is taken into account. The relevance is defined on the basis of results of semantic analysis of sentences. At the following cycle of the algorithm the term, that was not used earlier and has the greatest relevancy coefficient, is chosen as a key.

After a choice of a new key, actions 1 - 6 are repeated.

3. The semantic analysis

The basic operation of the semantic analysis of natural-language texts is recognition of the syntactic and semantic relations connecting words of the text. Recognition of relations is carried out on the basis of their descriptions (models). Such models are necessarily present at all methods of the analysis though it is not always obvious. In the majority of the analysis methods the process of recognition of relations is preceded with translation of initial natural-language representation of relations to be recognized in the language of categories of traditional grammar (gender, case, time, etc.). Rules of recognition of syntactic and semantic relations operate with grammatical descriptions of words. Binding to grammatical descriptions of elements of the text results in the following imperfections: heterogeneity of ways of processing separate words and word combinations; bulkiness of processing; complexity of adaptation to changes of lexicon and a user's domain; laboriousness of the research. Meanwhile, transition to grammatical descriptions is not an obligatory condition for performance of the semantic analysis of natural-language texts. The information necessary for recognition of syntactic and semantic relations is contained directly in the text. As a proof to that, there are "human" processes of the analysis of the natural-language texts, which are not connected with grammatical categories and rules. Therefore, it is competent another approach based on use of conformity between relations and means of their expression in natural-language texts. Recognition of syntactic and semantic connections between words is carried out by the analysis of combinations of inflections and prepositions, without using categories and rules of traditional grammar. By virtue of its basic features, such approach allows to exclude the imperfections named above.

Models of relations in which elements of natural-language texts are used for recognition of syntactic and semantic relations, we shall refer to as *lexical models of relations*. The algorithm of the semantic analysis of natural-language sentences on the basis of lexical models of relations is described in [1-3].

4. Implementation and results

The structure of the program complex realizing processes of thematic knowledge formation consists of the programs which are carrying out the following actions:

1. Selection in the INTERNET of the textual fragments containing a given key.
2. Formation of semantic representations of sentences (the linguistic processor).

3. Selection of sentences, relevant to a theme, on the basis of the analysis of semantic representations of sentences.
4. Choice of a new key.

At the present time lexical data and knowledge bases of the complex are created for Russian language.

As a result of working a program complex the text is formed which consists of separate sentences that are relevant to a theme designated by an initial key which is given by a user. For each sentence, the address of corresponding document is indicated. The set of sentences selected from one document allows generating a conception about its thematic relevance as a whole. The high level of relevance of the document may induce a user to choose this document for detailed studying. The set of all selected sentences throws light on an investigated theme as a whole. The degree of completeness of the selected information on a theme depends on efficiency of the used search machine and quantity of the texts chosen in the INTERNET. Experience of the complex exploitation shows that the set of sentences selected by a program on the basis of the thematic analysis, well correlates with result of “manual” selection of “useful” sentences by an end user. The complex provides the high degree of elimination of the information that is unnecessary for a user.

Conclusion

Above described method of thematic selection of information can be used for the information search not only in the INTERNET, but in any textual databases. We also consider it as the instrument for creation of ontology’s. The merit of the method is effective filtration of the information on the basis of criteria of relevancy to the given theme that is obtained at the cost of semantic analysis of sentences and a cyclic process of automatic selection of a new key at every cycle. The method allows comparatively simple adaptation to changes of a text language.

The literature

1. Gladun V.P. Processes of formation of new knowledge. - Sofia: СД "Педагог". 1994. - 192p. (in Russian).
2. Gladun V.P. Planning of decisions. Kiev: Наукова думка, 1987.-168p. (in Russian).
3. Gladun V.P. Natural language in purposeful systems.//DIALOG-2000. Applied problems. 2000, p.99-102. (in Russian).

Author information

Victor Gladun - V.M.Glushkov Institute of cybernetics of NAS of Ukraine, Prospekt akad. Glushkova 40, 03680 Kiev, Ukraine; e-mail: glad@aduis.kiev.ua

Alexander Tkachev - V.M.Glushkov Institute of cybernetics of NAS of Ukraine, Prospekt akad. Glushkova 40, 03680 Kiev, Ukraine; e-mail: glad@aduis.kiev.ua

Vitaly Velichko - V.M.Glushkov Institute of cybernetics of NAS of Ukraine, Prospekt akad. Glushkova 40, 03680 Kiev, Ukraine; e-mail: glad@aduis.kiev.ua

Neonila Vashchenko - V.M.Glushkov Institute of cybernetics of NAS of Ukraine, Prospekt akad. Glushkova 40, 03680 Kiev, Ukraine; e-mail: glad@aduis.kiev.ua