UDC 62-506.1

M. V. Karaseva, E. P. Bachurina, P. V. Zelenkov, V. V. Brezitskaya

INFORMATION SEARCH MODULE BASED ON MULTILINGUISTIC THESAURUSES

The paper is devoted to the module for search, ranking and determination of document relevance by means of meta-search multilinguistic algorithms of data processing.

Keywords: ranking, relevance level, meta-search.

At present time a lot of models and algorithms for information representation in the distributed information systems are developed. As special case of similar systems are informating-controlling systems, corporate information systems and intensively developing systems for decision-making support. However the majority of distributed systems are developed on the basis of monolingual representation of information or consider multilinguistic aspect implicitly [1].

Now the activities on development of new algorithms for search, ranking and relevance determination of the information both in global Internet network and in local informating-controlling systems are highly advanced. One of the promising trends in development of new models and algorithms of data processing is application of subject dictionaries, or thesauruses. A thesaurus presents as much as possible full volume of the lexicon arranged by a thematic (semantic) principle with representation of a certain set of basic semantic relations, which is full systematized data set on the specific field of knowledge which allows a person or a computer to be guided in it. It is necessary to mention that in modern systems such dictionaries-thesauruses are rarely presented in multilinguistic frequency realization. The authors of this paper apply thesauruses developed on the basis of multilinguistic technology for searching procedure in information systems.

The given approach is focused firstly on the problem of multilingual information representation in informating-controlling systems. In up-to-date conditions even small corporate information systems operate in a multilinguistic mode. Often administrative personnel at their decision-making needs prompt providing with documents of various language sets. The requirements to efficiency and quality of multilinguistic information search systems used by a decision-making person have considerably increased.

Within the proposed module the basic work on search, ranking and determination of relevance level is made by use of meta-search multilinguistic algorithms for data processing and control [2]. Firstly it is necessary to determine the searching process parameters. They are: functions of a knowledge domain choice and options of language sets, within which it is necessary to search.

Besides, it is necessary to show separately possibility of work with information search line both in Internet and in a corporate network. According to the offered approach, operation with a search line can be realized in two modes:

– a mode of manual input of a search line;

– a mode of automated arrangement of a search line.

At input of a search line in a manual mode the system checks presence of the entered terms in the frequency multilinguistic thesaurus. In case of a term failing in the dictionary the user is offered to enter a search line with change of terms in it.

Let's consider in more detail the process of inquiry generation on the given subject domain at input of a search line in an automatic mode [3]. The information search module is based on application of frequency multilinguistic thesauruses which raise the quality of documents relevance definition at inquiries. These thesauruses allow to allocate a document directivity, up to a subject domain which the document belongs to. The quality of relevance level definition in the offered method corresponds to relevance level in catalogue systems of manual indexation [1]. Based on the frequency characteristics of terms it is possible using the given algorithm to generate the search line adjustable by the user.

It is necessary to mention that in modern corporate information systems it can be stored multilinguistic information. However the user of the search module cannot know all the languages presented in the network. Therefore it is necessary to consider propriety of indication of the language sets necessary for the user.

After finishing the generation of a search line and indication of languages of information search, it is necessary to start directly the procedure of search [3]. As a result the consistent inquiry of all information corporate resources is realized and the file of references to documents interesting the user is formed, and also splitting of all set of references by a principle of language set membership is performed.

Besides, the user can see the following additional information which is considered at ranking of documents and definition of relevance level of each found document:

– document heading;

– document volume;

– quantity of the terms found in the document.

It is the first step of search procedure processing.

On the second step there is a definition of relevance level and ranging of a multilinguistic file of references. Here the user is provided with the additional information about:

– level of relevance of the found document;

– overall weight of relevant terms in the document.

The third step is a direct viewing of the found documents. On the given step it is possible not only to see the document, but also to receive the expanded information about it, presented, for example, in the form of the table, made at finishing of processing of the English-language document:

Term	Frequency	Term Weight in the Document	Term Weight in the Thesaurus
Activity	8	0.002	0.00000817
Process	3	0.000 9	0.0000108
Search	1	0.000 01	0.00000176

These characteristics are important at calculation of relevance level and ranking of the found documents [2]. Let's consider the structure of the given table:

– "Term" means the list of terms which have met in the thesaurus and the document;

– "Frequency" shows, how many time the given term has met in the document;

– "Term Weight in the Document" calculates concerning frequency of a term and total of terms in the document;

– "Term Weight in the Thesaurus" calculates as the relation of the frequency characteristic of a term in the dictionary to the general total frequency characteristic of all terms of the dictionary. The proposed module of information search and processing in corporate systems of decision-making support completely meets the requirements to systems of such level and allows to solve a problem of arranging, storage and processing of information in the modern distributed multilinguistic corporate systems of decision-making support.

Realization of metasearch principles promotes coverage of indexes of the most popular search web services. At the same time the number of irrelevant references resulted at searching is reduced, the quality of inquiries processing of the user essentially raises and the traffic volume decreases while forming a personal base of relevant documents.

Besides, the presented multilinguistic models allow to make more flexible multilinguistic answers even on monolingual inquiries in comparison with simple distributed-informational system, taking into account uncertainty of the description both multilinguistic documents and inquiries.

References

1. Multilinguistic Model of Distributed System Based on Thesaurus / P. V. Zelenkov, I. V. Kovalev, M. V. Karasva, S. V. Rogov // Bulletin of Siberian State Aerospace Univ. №. 1(18). Krasnoyarsk, 2008. P. 26–27.

2. Meta-search Multilinguistic System for Highly Specialized Information : software registration certificate of All-Russia Center of Scientific Information № 50200701673 / I. N. Kartsan, P. V. Zelenkov, D. A. Ragzin et al. M., 2007.

3. Zelenkov P. V., Kovaleva T. A. A problem of Meta-search Technologies Development // Bulletin of Scientific Institute of Wave Processes Control Systems. Vol. 8. Krasnoyarsk, 2004. P. 95–103.

М. В. Карасева, Е. П. Бачурина, П. В. Зеленков, В. В. Брезицкая

МОДУЛЬ ПОИСКА ИНФОРМАЦИИ НА ОСНОВЕ ИСПОЛЬЗОВАНИЯ МУЛЬТИЛИНГВИСТИЧЕСКИХ ТЕЗАУРУСОВ

Статья посвящена новому модулю поиска, ранжирования и определения уровня релевантности многоязычных документов. Описанные методы основаны на метапоисковых мультилингвистических алгоритмах обработки информации.

Ключевые слова: ранжирование, уровень релевантности, метапоиск.