УДК 681.3

Ю. М. Смирнов, А. М. Андреев,
Д. В. Березкин, А. В. Брик

ИСПОЛЬЗОВАНИЕ СТАТИСТИЧЕСКИХ МЕТОДОВ ДЛЯ СОЗДАНИЯ ЛИНГВИСТИЧЕСКОГО ОБЕСПЕЧЕНИЯ ИНФОРМАЦИОННО-ПОИСКОВОЙ СИСТЕМЫ

Рассмотрены проблемы создания информационно-поисковой системы с естественно-языковым интерфейсом запросов, в частности, подготовка словарей и поискового индекса, учитывающего синтаксическую структуру предложений документа. Предложен метод автоматического построения морфологического словаря и словаря словосочетаний, использующий статистический анализ достаточно большого множества текстов. Рассмотрен двухступенчатый алгоритм синтаксического анализа текста (использующий на первом этапе простой формально-грамматический анализ, а на втором — уточнение результатов его работы статистическими методами), а также алгоритм текстового поиска на основе результатов его работы. Приведены экспериментальные оценки качества работы предложенных методов.

Application of Statistic Methods for Development of Linguistic Support for Data Retrieval System / Yu.M. Smirnov, A.M. Andreev, D.V. Beryozkin, A.V. Brik // Vestnik MGTU. Priborostroenie. 2001. No. 2. P. 13–24.

Problems of the data retrieval system development with natural language interface of requests are considered, among them, the preparation of dictionaries and search index taking into account syntactic structure of the document sentences. A method of the automatic creation of both the morphological and word-combination dictionary is suggested using statistical analysis of the sufficient amount of texts. The two-stage algorithm of the text syntax analysis is considered (using the simple formal and grammatical analysis at the first stage and the statistical refinement of the analysis results — at the second stage), and the text search algorithm as well, based on results of the two-stage algorithm application. Experimental estimations of the suggested methods operation quality are given. Refs.8. Figs.1.

СПИСОК ЛИТЕРАТУРЫ

1. Андреев А. М., Березкин Д. В., Брик А. В. Лингвистический процессор для информационно-поисковой системы // Компьютерная хроника. – 1998. – № 11. – С. 79–100.

2. Смирнов Ю. М., Андреев А. М., Березкин Д. В., Брик А. В. Вероятностный синтаксический анализатор для информационно-поисковых систем // Вестник МГТУ. Серия "Приборостроение". – 2000. – № 2. – С. 34–54.
3. Попов Э. В. Общение с ЭВМ на естественном языке. – М.: Наука, 1982.
4. Перспективы развития вычислительной техники: В 11 кн.: Справ. пособие / Под ред. Ю.М. Смирнова. Кн. 2. Интеллектуализация ЭВМ / Е.С. Кузин, А.И. Ройтман, И.Б. Фоминых, Г.К. Хахалин. – М.: Высш. шк., 1989. – 159 с.
5. Сегалович И. В. Индексирование русских текстов с использованием словаря, представленного на основе разреженной хеш-таблицы // Труды Международного семинара по компьютерной лингвистике и ее приложениям "Dialogue'95". – Казань, 31 мая – 4 июня 1995. // www.comptek.ru/yandex/kazan.html.
6. Salton G. Automatic text processing. – Addison-Wesley, 1989.
7. Magerman D. M. Natural Language Parsing as Statistical Pattern Recognition. // A dissertation submitted to the department of computer science at the committee on graduate studies of Stanford University, 1994. // xxx.lanl.gov/cmp-lg.
8. Hrmjakob U. Learning Parse and Translation Decisions From Examples With Rich Context. // A dissertation presented to the Faculty of Graduate School of the University of Texas. Austin, 1997. // xxx.lanl.gov/cmp-lg.

Статья поступила в редакцию 22.02.01

Юрий Матвеевич Смирнов родился в 1923 г., окончил МВТУ им. Н.Э. Баумана. Д-р техн. наук, профессор кафедры "Компьютерные системы и сети" МГТУ им. Н.Э. Баумана. Лауреат Государственной премии СССР. Автор более 200 научных работ в области вычислительных средств и систем управления.

Yu.M. Smirnov (b. 1923) graduated from the Bauman Moscow State Higher Technical School. D.Sc. (Eng.), professor of "Computer Systems and Networks" department of the Bauman Moscow State Technical University. USSR State Prize winner. Author of over 200 publications in the field of computation means and control systems.

Арк Михайлович Андреев родился в 1943 г., окончил в 1967 г. МВТУ им. Н.Э. Баумана. Канд. техн. наук, доцент кафедры "Компьютерные системы и сети" МГТУ им. Н.Э. Баумана. Автор более 70 научных работ в области вычислительных средств и систем управления.

A.M. Andreev (b. 1943) graduated from the Bauman Moscow State Higher Technical School in 1967. Ph.D. (Eng.), ass. professor of "Computer Systems and Networks" department of the Bauman Moscow State Technical University. Author of over 70 publications in the field of computation means and control systems.

Дмитрий Валерьевич Березкин родился в 1966 г., окончил в 1990 г. МГТУ им. Н.Э. Баумана. Канд. техн. наук, директор Научно-производственного центра "Интелтек Плюс". Автор около 30 научных работ в области вычислительных средств.

D.V. Beryozkin (b. 1966) graduated from the Bauman Moscow State Technical University in 1990. Ph.D. (Eng.), director of Scientific and Industrial Center "INTELTEK PLUS". Author of about 30 publications in the field of computation means.

Алексей Владимирович Брик родился в 1974 г., окончил в 1997 г. МГТУ им. Н.Э. Баумана. Аспирант кафедры "Компьютерные системы и сети" МГТУ им. Н.Э. Баумана. Специализируется в области использования формальных грамматик и методов синтаксического анализа в информационных технологиях.

A.V. Brik (b. 1974) graduated from the Bauman Moscow State Technical University in 1997. Post-graduate of "Computer Systems and Networks" of the Bauman Moscow State Technical University. Specializes in application of formal grammar and syntax analysis methods in information technologies.