УДК 001.38

Жаныс Арай Бошанкызы,
доктор философии PhD, профессор
Аманкелди Айжан,
Кокшетауский университет имени Абая Мырзахметова,
г. Кокшетау, Казахстан

ИНФОРМАЦИОННО-ПОИСКОВАЯ СИСТЕМА КАСКАД

Аннотация. *В условиях информационного общества обилие разработанных и разрабатываемых интеллектуальных и информационных систем обработки текста, к которым относятся информационно-поисковые системы и СМП в том числе, требует быстрой и адекватной оценки. Важно иметь представление о самом качестве систем, как на стадии начальной разработки, так и на стадии готового продукта. Создание формальной методики решения этой проблемы является актуальным, как в теоретическом, так и практическом смысле.*

Ключевые слова: *автоматизированная многоязычная база данных (АМБД) КАСКАД, прикладные программы, организатор разработчик, научно-техническая терминология, документо-графическая единица, интерфейс, автоматический словарь, лингвистическое обеспечение, десриптор, анализ, частный словарь, квази основа слов, семантический словарь, идентификатор документа, интерфесы файлов, эшолон.*

Автоматизированная многоязычная база данных (АМБД) КАСКАД, в разработке которой автор принимал непосредственное участие, предназначена для обеспечения зарубежных пользователей информацией об отечественных программных продуктах. Информационное и лингвистическое обеспечение АМБД формирует базу данных по программным средствам (ПС) на русском языке, осуществляет поиск данных по запросу на естественном языке (ЕЯ) (русском или иностранном, в существующем варианте – английском) и производит перевод русскоязычной БД на иностранные (английский) языки.

Информационную базу системы КАСКАД составляют описания программных средств (пакетов прикладных программ, ИПС, СУБД, СМП и т.п.). Эти описания выполняются в соответствии с анкетой, которая содержит сведения о наименовании ПС, его назначении, аннотацию, а также данные об организации-разработчике, годе окончания разработки, адресные данные, координаты для связи и пр [5, с. 6–7].

Основу АМБД составляет русский вариант БД по отечественным ПС, предназначенным для зарубежных пользователей. Средства доступа к ней обеспечивают возможность поиска по отдельным параметрам описания ПС с помощью подсказок и по запросам, сформулированным на ЕЯ (русском или английском). Базы данных на других языках образуются путем перевода символьных полей документов русскоязычной БД. Между русскоязычной и иноязычными базы данных сохраняется полное взаимно-однозначное соответствие – все файлы имеют одинаковую структуру, состав, наименование, но хранятся в разных каталогах.

Использование СМП для получения БД на иностранном языке или автоматических словарей дает возможность быстро осуществлять перевод данных на иностранные языки (что становится экономически выгодным для БД, насчитывающих сотни и тысячи документо-графических единиц), а также позволяет использовать единообразную научно-техническую терминологию, что, в свою очередь, существенно улучшает поисковые возможности системы. При установлении взаимно-однозначного соответствия между терминами русского и иных языков появляется возможность производить поиск в базе и выдавать его результаты на любом языке. Языки интерфейса и выходных документов определяется пользователем [10, с. 10–19].

Поскольку в системе используются автоматические словари, и структурой АМБД предусмотрена синонимичность файлов на различных языках, можно производить поиск по запросу

на Я только в русскоязычной базе. При этом запрос может быть сформулирован на любом, предусмотренном в АМБД, языке. Затем с помощью автоматического словаря он переводится на русский язык. После этого запрос обрабатывается по русскоязычной базе, а результаты поиска (по номерам найденных документов) выводятся на языке исходного запроса.

Лингвистическое обеспечение ЛМБД базируется на автоматическом выделении основ ключевых слов (дескрипторов) в текстовых полях (полное наименование и аннотация ПС). Для формирования словаря основ ключевых слов использовался словарь окончаний, разработанный при создании СМП ЛСПЕРА, и составленные предварительно вручную словари «стоп-слов» и «стоп-основ», которые включают предлоги, союзы, наречия, некоторые глаголы и неинформативные существительные.

Построение частотного словаря дескрипторов производится в несколько этапов. На первом этапе перед выделением основы из рассмотрения исключаются «стоп-слова», а затем формируется словарь основ путем отделения от словоформ максимального квазиокончания. После анализа лингвистами в получающемся словаре помечаются неинформативные основы (которыми образуют словарь «стоп-основ»), а основы информативных [3, с. 42–57] слов корректируются (укорачиваются, объединяются и т.п.). На втором этапе производится пополнение словаря дескрипторов. При этом сначала вновь исключаются из рассмотрения «стоп-слова», а затем в слове осуществляется поиск основы из словаря дескрипторов. Если найденная основа помечена как неинформативная, то слово пропускается, а если в слове не нашлось ни одной словарной основы, то слово выводится в словарь новых слов, который анализируется лингвистами. В результате анализа из новых слов выделяются основы и заносятся в словарь дескрипторов. Неинформативные основы вновь помечаются. После обработки всего текста формируется собственно частотный словарь дескрипторов, где для каждого слова указана его относительная частота. Все словари корректируются при пополнении БД с помощью соответствующих программ и редактирования результатов их работы специалистами [4, с. 35–61].

Для предоставления большей свободы при формулировке запроса частотный словарь пополняется словами-синонимами, которым ставится в соответствие одинаковая частота. Запрос на русском языке преобразуется с помощью частотного словаря в последовательность дескрипторов, после чего проверяется вхождение дескрипторов запроса в документы. Границы значений частот для формирования критерия выдачи определяются для высокочастотных, среднечастотных и низкочастотных основ.

При наличии частотных словарей основ ключевых слов критерием выдачи для выбора релевантных документов служит:

- вхождение слов запроса в документ,
- совпадение частот слов запроса в документе,
- попадание частот слов, входящих в запрос и в документ, в указанную категорию частоты (высокая, средняя, низкая) [5, с. 67–70].

Если пользователь в качестве языка общения использует иностранный язык, то запрос формируется на выбранном языке, после чего автоматически переводится на русский язык. Поиск производится в русскоязычной базе. По номерам документов, выбранным при поиске, выдаются документы из базы на иностранном языке.

АМБД практически не может содержать очень большого количества документов; пополнение и корректировка БД производится не чаще одного раза в полгода. При таких условиях, казалось бы, можно воспользоваться услугами переводчика для получения иноязычной базы данных. Но субъективизм переводчиков, неадекватность используемой ими лексики существенно снижают поисковые возможности системы, а потому при переводе с русского языка целесообразнее воспользоваться если не автоматическим переводом, то хотя бы автоматическим словарем, обеспечивающим пословный перевод, с последующим ручным редактированием. При разработке системы КАСКАД была использована ранее разработанная СМП АСПЕРА и предусмотренные ею

русско-английские словари [6, с. 276–280].

Русско-английский словарь состоит из трех частей:

- словаря основ русских слов;
- словаря английских эквивалентов слов;
- словаря словосочетаний.

Словарь русских основ состоит из записей, упорядоченных по алфавиту. Каждая запись представляет собой собственно основу слова, грамматическую характеристику (до семи наборов грамматических признаков без разделителей), относительный адрес перевода в словаре эквивалентов. В качестве основы слова может быть представлена квазиоснова или словоформа, если длина основы меньше 4-х символов.

В словаре английских эквивалентов каждый перевод слова представлен строками трех видов:

- первая содержит собственно перевод слова и его обобщенную грамматическую характеристику (на эту строку и дается ссылка в словаре русских основ);
- вторая содержит обозначение модели управления и признак ее обязательности (1) или факультативности (2);
- третья – грамматические характеристики русского слова и перевода, а также коды семантических признаков [7, с. 489].

В словаре словосочетаний статья представляет собой группу строк, в которых представлены:

- русское словосочетание с грамматической характеристикой у каждой основы;
- английский перевод словосочетания с грамматическими характеристиками;
- семантическая информация (номер слова в словосочетании, модель управления, грамматические характеристики и коды семантических признаков).

В информационных файлах системы базы данных каждая запись (документ) представляет собой последовательность полей, каждое из которых является символьным текстом или кодом, значение которого приведено в соответствующем словаре. Идентификатором документа является первый код – регистрационный номер документа [8, с. 69–80].

Инверсные файлы формируются после заполнения системы базы данных или ее актуализации и позволяют существенно сократить время поиска в системы базы данных по запросу пользователя. Инверсные файлы состоят из двух полей: первое поле содержит входной код или дескриптор – 20 символов; второе – регистрационный номер ППС – 5 символов.

В диалоговом режиме с помощью подсказок формируется запрос по следующим параметрам [9, с. 20–45]:

- организация-разработчик (сокращенное название);
- организация-владелец (сокращенное название);
- наименование (обозначение) ПС;
- тип (назначение) ПС (набор ключевых слов и словосочетаний);
- цена (в $);
- рейтинг (число осуществленных продаж от...);
- год разработки (от...).

В запросе значения, относящиеся к одному параметру, объединяются с помощью логической операции ИЛИ, а значения, относящиеся к разным параметрам, объединяются с помощью логической операции И.

Для обработки стандартного запроса разработан комплекс программ, который обеспечивает:

- формирование стандартного запроса в режиме диалога с пользователем с помощью подсказок;
- выбор документов, но стандартному запросу;
- формирование выходного документа и вывод найденных документов на экран или в текстовый файл [10, с. 12–16].

Как уже говорилось, поиск по запросам на ЕЯ производится только в русскоязычной базе. На первом этапе запрос преобразуется в последовательность дескрипторов (по частотному словарю

основ ключевых слов). Если ни одного слова запроса в словаре не найдено, пользователю предлагается повторить формулировку запроса, воспользовавшись в качестве подсказки словарем дескрипторов. Запрос на английском языке предварительно переводится (пословно) на русский язык. Далее работа ведется аналогично, но в качестве подсказки приводится английский словарь дескрипторов. Поиск найденных всех дескрипторовно словарю дает входы в инверсный файл. В этом файле для каждого дескриптора указывается перечень номеров документов базы, в которых данный дескриптор встречается [101, с. 40]. Результаты поиска эшелонируются в соответствии **с** критерием выдачи (вхождение всех дескрипторов запроса в документ; получение пересечения номеров документов по всем дескрипторам; вхождение дескрипторов с высокой или низкой частотой и т.п.).

Описанная ИПС разработана в Комитете по информатизации при Президенте Казахстана.

ЛИТЕРАТУРА

1. Аверкин А.Н., Батыршин И.З. и др. Нечеткие множества в моделях управления и искусственного интеллекта / Под ред. Д.А Поспелова. – М.: Наука, 1986. – 311 с.
2. Анализ нечисловых данных в системных исследованиях // Сборник трудов ВНИИСИ. – М.: ВНИИСИ, 1982. – Вып. 10. – 155 с.
3. Аграев В.А., Казакевич Б.Л., Кобрин Р.Ю. Первый частотный словарь индексирования (рецензия) // НТИ : Сер.2. – М.: Наука, 1976. – № 4. – С. 36–38.
4. Алексеев П.М. Статистическая лексикография (типология, составление и применение частотных словарей). – Л.: ЛГПИ им. А.И. Герцена, 1975. – 221с.
5. Агеев В.Н., Узилевский Г.Я. Человеко-компьютерное взаимодействие: концепции, процессы, модели. – М.: Мир книги, 1995. – 352 с.
6. Аношкина Ж.Г. Подготовка частотных словарей и конкордансов на компьютере. – М.: МГУ, 1995. – 234 с.
7. Апресян Ю.Д. и др. Лингвистическое обеспечение системы: ЭТАП-2. – М.: Наука, 1989. – 295 с.
8. Баранов А.Н. Автоматизация лингвистических исследований: корпус текстов как лингвистическая проблема // Русистика сегодня. – Москва: Русский язык, 1998. – № 1–2. – С.179–191.
9. Баранов А.Н. Введение в прикладную лингвистику. – М.: УРСС, 2001. – 358 с.
10. Белоногов Г.Г., Зеленков Ю.Г. и др. Системы фразы логического машинного перевода – технология XXI века // Материалы конф. НТИ–97. – М: ВИНИТИ, 1997. – С. 255–283.

Жанис Арай Бошанкизи, Аманкелді Айжан. Інформаційно-пошукова система КАСКАД. – Стаття.

Анотація. *В умовах інформаційного суспільства велика кількість розроблених і розроблюваних інтелектуальних і інформаційних систем обробки тексту, до яких відносяться інформаційно-пошукові системи і СМП в тому числі, вимагає швидкої та адекватної оцінки. Важливо мати уявлення про саму якість систем, як на стадії початкової розробки, так і на стадії готового продукту. Створення формальної методики вирішення цієї проблеми є актуальним, як в теоретичному, так і практичному сенсі.*

Ключові слова: *автоматизована багатомовна база даних (АМБД) КАСКАД, прикладні програми, організатор розробник, науково-технічна термінологія, документо-графічна одиниця, інтерфейс, автоматичний словник, лінгвістичне забезпечення, десріптор, аналіз, приватний словник, квазі основа слів, семантичний словник, ідентифікатор документа, интерфес файлів, ешолон.*

Zhanys Aray Boshanqyzy PhD, Amankeldi Aizhan. Information Retrieval System Cascade. – Article.

Summary. *The abundance of information society developed and developing intelligent information systems and text processing, which include a retrieval system, and SMEs in particular, requires quick and adequate assessment. It is important to have an understanding of the systems, such as the initial development stage and at the stage of the final product. Create a formal method to solve this problem is the most relevant topic of the thesis, both in theoretical and practical sense.*

Key words: *automated multilingual database (DBA) CASCADE, applications, organizer of the developer, the scientific and technical terminology, documents, graphics unit, interface, automatic dictionary, linguistic support, desriptor, analysis, special dictionary, quasi the basis of words, semantic dictionary, the document identifier, user interface of files esholon.*