ИРБИС64 как инструмент создания и ведения полнотекстовых баз данных

IRBIS64 as a Tool for Creation and Maintenance of Full-Text Databases

ІРБІС64 як інструмент створення і ведення повнотекстових баз даних

Бродовский А. И., Попов Е. В., Сбойчаков К. О.
Государственная публичная научно-техническая библиотека России, Москва, Россия

Alexander I. Brodovsky, Eugeny V. Popov, and Konstantin O. Sboichakov
Russian National Public Library for Science and Technology, Moscow, Russia

Бродовський О. І., Попов Є. В., Сбойчаков К. О.
Державна публічна науково-технічна бібліотека Росії, Москва, Росія

Изложены результаты работы по созданию оригинальной версии системы автоматизации библиотек ИРБИС64, предназначенной для работы с полнотекстовыми базами данных. Приводится подробное описание алгоритмов ранжированного поиска и поиска «схожих» документов.

The paper presents the results of the creation of an original version of IRBIS64 whose mission is to handle full-text databases. It describes in detail the ranged search algorithms and similar document search algorithms.

Викладено результати роботи по створенню оригінальної версії системи автоматизації бібліотек ІРБІС64, що призначена для роботи з повнотекстовими базами даних. Наводиться детальний опис алгоритмів ранжованого пошуку та пошуку «схожих» документів.

В предыдущих работах [3,4] в качестве основной задачи развития системы автоматизации библиотек ИРБИС64 определена поддержка полнотекстовых баз данных с обеспечением классификации и смыслового анализа текстов. В настоящее время в ГПНТБ России в рамках ИРБИС64 создана подсистема (далее система) для формирования и использования полнотекстовых баз данных. Тексты, составляющие такие базы данных, могут быть в форматах TXT, DOC, RTF, PDF, HTML. Никакой дополнительной обработки для включения полных текстов в базу данных не требуется. Тексты сохраняются в базе данных в специальном архивном файле или в виде ссылок.

Основой данной разработки служит уже неоднократно представленная на международных конференциях «Крым» и «Либком» система смыслового анализа текстов [1,2]. Вкратце подходы к решению задачи смысловой обработки текстов, применяемые в системе, могут быть сформулированы в виде следующих этапов:

1. Создание полнотекстовой базы данных из массива текстов.
2. Естественно-тематическая классификация текстов на основе выделения значимых терминов предметной области. Тематическая классификация позволяет сравнивать тексты между собой на предмет близости их по смыслу. Тематический классификатор — это набор тематических словарей, в который входят термины, значимые в данной предметной области.

Система полнотекстовых баз данных ИРБИС64 включает в себя:

1. Расширенный АРМ «Администратор», который помимо стандартных для системы ИРБИС функций администрирования включает в себя дополнительные режимы для работы с полнотекстовыми базами данных.
2. АРМ конечного пользователя (читателя) для поиска и просмотра полнотекстовых баз данных в локальной сети или на CD-ROM. В этом АРМе реализованы специальные поисковые алгоритмы:
   - Поиск по запросу на естественном языке с ранжированием найденных текстов.
   - Поиск «схожих» текстов в заданном пользователем тематическом контексте.
   - Кроме того, АРМ позволяет публиковать (представлять) Интернет-сайты, описывающие полнотекстовые базы данных в целом.
3. Шлюз для доступа к полнотекстовым базам данных по технологии WWW.

Остановимся подробнее на алгоритмах ранжированного поиска и поиска «схожих» текстов, представляющих собой оригинальные решения.

Алгоритм ранжированного поиска

В системе применяется алгоритм поиска с ранжированием близкий к классическому алгоритму TF*IDF [5,7], который расширен за счет:

- учета расстояний между словами;

- увеличения весов терминов, входящих в тематические словари.

При разборе строки запроса на естественном языке производится предварительная фильтрация с целью удаления стоп-слов, специальных символов и слишком коротких слов (менее 2-х символов). Далее слова подвергаются усечению (выделению основ), отбрасываются слова, не имеющие ссылок в словаре базы данных, и формируется окончательный список слов запроса, соединенных логикой «И». Вес каждого слова в запросе оценивается следующим образом:

$$TFIDF = TF*IDF,$$

где TF — частотность слова в тексте. В системе этот параметр равен 1. IDF — коэффициент уменьшения веса слова в зависимости от его распространенности в словаре базы данных.

$$IDF = \log_2(MaxMFN/df+1) / \log_2(MaxMFN+1),$$

где MaxMFN — число текстов в базе данных и df — число текстов, содержащих данное слово.

Если слово находится в словаре тематического классификатора системы, заданного в качестве контекста поиска (если контекст поиска не задан пользователем, по умолчанию используется общий контекст базы данных), то вес его умножается на эвристический коэффициент (1000).

При расчете ранга текста RANG используются веса слов запроса TFIDF и расстояния между ними в тексте. Причем для каждой пары слов берется минимальное расстояние между ними.

$$RANG = \sum\sum ij\ (TFIDFi*TFIDFj / MIN(Rij)^2),$$

где Rij — расстояние между парой слов. Если слова рядом, Rij = 1.

Вводится следующее различие в работе алгоритма в зависимости от длины запроса:

- При работе с короткими запросами (до 4-х слов), если не найдены тексты, содержащие все слова запроса, запускается цикл формирования последовательности новых запросов путем отбрасывания слов согласно убыванию их веса. Если и в этом случае ничего не найдено, система пробует отбрасывать два и, наконец, три слова. Таким образом, ранг текста строго связан с числом содержащихся в нем слов запроса.
- При работе с длинными запросами формируется битовая шкала текстов, содержащих слова из запроса, специально организованная для последующей сортировки. Для отбора текстов в эту шкалу используется критерий (необходимый кворум [7]), который рассчитывается для данного текста как сумма весов слов, найденных в нем. После отбора и сортировки по значениям кворумов производится сортировка по рангу текста, рассчитываемого по вышеизложенному алгоритму в зависимости от расстояний между словами. Полученный результат уже не так строго зависит от числа слов найденных в тексте.

При поиске можно указать дополнительный параметр — максимальное расстояние между словами. В этом случае необходимым условием выдачи является наличие в тексте фрагментов, включающих все найденные слова из запроса. Если результат поиска нулевой, делается попытка применить менее строгий критерий отбора за счет учета неполных фрагментов. При расчете ранга текста используется минимальное расстояние между словами [см. выше]. То есть постулируется принцип — лучше один «хороший» фрагмент, чем несколько «плохих».

Максимальный размер фрагмента (для включения текста в выдачу) составляет величину:

$$F = NumWords*MaxDistance,$$

где NumWords — число слов в запросе, MaxDistance — максимальное расстояние между словами.

Алгоритм поиска «схожих» текстов

В расширенном АРМе «Администратор» предлагается ориентированная на пользователя-эксперта автоматизированная технология создания тематических словарей (контекстов), составляющих основу естественно-тематической классификации [2]. Причем основной тематический словарь базы данных, составляющий ее основной тематический контекст и включающий наиболее общую терминологию, создается автоматически.

Алгоритм поиска «схожих» текстов базируется на использовании естественно-тематической классификации. При этом сначала производится пересечение заданного контекста поиска (тематических словарей) и текста, выбранного в качестве образца. В результате формируется список слов общих для контекста и текста-образца. Объем этого списка относительно контекста ограничивается снизу параметром настройки «Степень сходства», который принимает следующие значения:

- >= 5 % — слабо.

- >=10% и <15% — приблизительно.
- >=15% точно.

Слова из этого списка составляют поисковый запрос. В этом случае ранжирование результатов поиска производится без учета расстояний между словами, играет роль только вес слова TFIDF. Скорость работы алгоритма значительно оптимизирована за счет кэширования ссылок терминов контекста.

Примером практического применения представленной системы служит CD-ROM с полнотекстовой БД материалов конференции «Крым 2005», который получили все участники конференции.

Литература

1. Макагонов П. П., Сбойчаков К. О. Интерактивные методы решения слабо-формализованных задач в гуманитарных и естественно научных приложениях: (Визуальный эвристический кластерный анализ) // Материалы симпозиума по компьютерным приложениям CIC'98, Мексиканский Национальный Политехнический институт. — Мехико,1998. — С.346-358. — Англ. яз.
2. Макагонов П. П., Александров М. А., Сбойчаков К. О. Программное обеспечение для создания предметно-ориентированных словарей и кластеризации текстов в полнотекстовых базах данных // Компьютерная лингвистика и интеллектуальная обработка текстов. — Б. г.:Шпрингер,2001. — С.454-456. — Англ. яз.
3. Бродовский А. И., Сбойчаков К. О. Новое поколение системы автоматизации библиотек ИРБИС — ИРБИС64: от электронного каталога к полнотекстовым базам данных // Библиотеки и ассоциации в меняющемся мире: новые технологии и новые формы сотрудничества: Тр конф. — М., 2004.
4. Бродовский А. И., Сбойчаков К. О. Полнотекстовые базы данных в системе ИРБИС64 // Информационные технологии, компьютерные системы и издательская продукция для библиотек: Тр конф. — М., 2004.
5. Callan J. P., Croft W. B. and Harding S. M., The INQUERY Retrieval System // A. M. Tjoa and I. Ramos (eds.), Database and Expert System Applications. Proceedings of {DEXA}-92, 3-rd International Conference on Database and Expert Systems Applications. — Springer Verlag, New York. — 1992. — pp.78-93.
6. Илья Сегалович, Михаил Маслов, Яндекс на РОМИП 2004. Некоторые аспекты полнотекстового поиска и ранжирования в Яндекс // Ромип 2004: Тр конф. — М., 2004.
7. М. С. Агеев, Б. В. Добров, Н. В. Лукашевич, А. В. Сидоров, Экспериментальные алгоритмы поиска/классификации и сравнение с «basic line» // Ромип 2004: Тр конф. — М., 2004.