ИСПОЛЬЗОВАНИЕ РЕЛЕВАНТНОСТИ В ПОЛНОТЕКСТОВОМ ПОИСКЕ В СИСТЕМАХ УПРАВЛЕНИЯ ПРОИЗВОДСТВОМ

Костыря Е.И.
Научный руководитель: Ковин Р.В, доцент
Томский политехнический университет, 634050, Россия, г. Томск, пр. Ленина, 30
E-mail: e.kostyrya@gmail.com

Автоматизация производственных и управленческих процессов с помощью ERP и MES систем в настоящее время является почти стандартом ведения бизнеса на рынке. Оптимизация работы сотрудников позволяет сокращать многие расходы, связанные с неточностью расчетов или отставанием от плана. Системы управления производством, обладая широким функционалом, позволяют решать различные задачи, которые стоят перед отдельно взятой компанией[1].

Выбор того, какой тип систем стоит внедрить на предприятии, основывается на целях, которые ставит руководство. Так, системы ERP решают задачи финансового менеджмента, управления трудовыми ресурсами, управления активами. В то время как MES является исполнительной системой, направленной на решение задач синхронизации, координации, анализа и оптимизации выпуска продукции в рамках какого-либо производства.

Эффективность использования таких систем в первую очередь зависит от того, насколько удобно можно найти требуемую информацию. Базы данных таких систем представляют хранилища, в которых находится более миллиона записей, таким образом, системы оперируют большим количество данных. Несмотря на качественную реализацию алгоритмов, которые выполняют основные операции, на которые нацелена отдельно взятая система, блок поиска, как часть системы, зачастую, прорабатывается менее детально. В итоге, поиск имеет очень маленький функционал, что вызывает трудности работы с системой в целом. Такой тип поиска как полнотекстовый позволяет решить эту проблему, однако сложность выбора необходимой информации остается, из-за слишком большого числа выведенных результатов. Поэтому реализация полнотекстового поиска с учетом релевантности в системе является оправданным шагом, который позволит оптимизировать работу с системой.

Реализация полнотекстового поиска была выполнена для MES-системы «Магистраль - Восток», которая является системой управления производством в нефтегазовой отрасли. Программное обеспечение «Магистраль – Восток» состоит из серверной (СУБД Microsoft SQL Server) и клиентской части [2]. Процесс поиска происходит следующим образом. Искомая информация, введенная клиентом в строку ввода, передается специальной процедуре-хранимой процедуре, которая является объектом базы данных, представляющий собой набор SQL-инструкций, который компилируется один раз и хранится на сервере. Хранимая процедура формирует текст запроса на языке TSQL и выполняет его. При этом происходит обращение к механизму полнотекстового поиска, который в свою очередь отвечает за хранимые данные. По полнотекстовому индексу происходит выборка требуемых данных. Запрошенные данные возвращаются через хранимую процедуру, которая передает результат запроса на клиентское ПО. В итоге клиент видит результат поиска. Поиск, предоставляет собой не одну хранимую процедуру, а как минимум две: одна - отвечает за индексирование таблиц (все полнотекстовые индексы хранятся в полнотекстовом каталоге), другая - непосредственный поиск. Автоматическая индексация таблиц является необходимым этапом, поскольку количество таблиц слишком велико для ручной индексации, именно поэтому модуль индексирования формируется в виде хранимой процедуры, так же как и модуль самого поиска.

Возможности полнотекстового поиска не ограничиваются морфологическим разбором слова, поиск можно интеллектуализировать, за счет введения ранжирования результатов. Ранжирование результатов поиска представляет некоторую сортировку, а именно вывод релевантных результатов к поисковому запросу.

Процесс вычисления ранга зависит от нескольких факторов. Средства разбиения по словам в различных языках по-разному разбивают текст на лексемы. Например, строку «dog-house» одно средство разбиения по словам может разбить на «dog» и «house», а другое — на «dog-house». Это означает, что соответствие и ранжирование будут зависеть от заданного языка, потому что в разных языках различаются не только слова, но и длина документа. Разница в длине документа может повлиять на ранжирование во всех запросах.

Такие статистические данные, как IndexRowCount (Общее число индексированных строк. Вычисляется на основе счетчиков в промежуточных индексах. Точность этого числа может быть различной.), могут различаться в широких пределах. Например, если полнотекстовый каталог имеет 2 миллиарда строк в главном индексе, то новый документ индексируется хранящимся в памяти индексом. Поэтому ранги для этого документа, вычисленные

на основе количества документов в индексе, хранящемся в памяти, могут отличаться от рангов для документов из главного индекса.

Средства MS SQL Server ранжирования:2008R2 позволяют использовать четыре различные предикаты[3]:

- FullText
- Containstable
- SABOUT
- Freetexttable

Предикаты ранжирования основываются на разных алгоритмах. Наиболее используемые являются алгоритмы ранжирования на основе формулы Жаккарда и формулы ранжирования OKAPI BM25.

Так, ранжирование предикатом FullText позволяет сравнить строку поискового запроса с ключевым полем, после сделать сортировку по убыванию. FullText является наиболее простым механизмом ранжирования. В то время как Containstable позволяет ранжировать неточные совпадения каждого из слов, входящих в строку поиска[4].

Наиболее эффективными являются предикаты Freetexttable и параметр ISABOUT предиката Containstable.

ISABOUT — это запрос в векторном пространстве, если пользоваться традиционной терминологией извлечения данных. Ранжирование вычисляется для каждого термина в запросе, а затем результаты объединяются.

Ранжирование предикатом Freetexttable основывается на формуле OKAPI BM25 [5]. Запросы FREETEXTTABLE добавляют к запросу словоформы, полученные по исходным словам запроса. Эти слова обрабатываются как отдельные и независимые, не относящиеся к словам, производными которых они являются. Синонимы, сформированные с помощью тезауруса, обрабатываются как отдельные, независимые и взвешенные выражения. Каждое слово в запросе вносит свой вклад в ранжирование.

В результате исследования, было выявлено, что предикат Freetexttable выполнял наиболее эффективный поиск, в рамках поставленных задач. Сравнение происходило двух наиболее успешных в полнотекстовом поиске предикатов Freetexttable и Containstable. Так, при работе с БД «Магистраль-Восток», при вводе строки ‘Бурение*’ (см. табл.1.) запрос с предикатом Containstable выдал все результаты по порядку, в то время как предикат Freetexttable выдал численное значение ранга релевантного результата.

Таким образом, полнотекстовый поиск с учетом релевантности позволяет существенно повысить эффективность поиска. За счет несложных алгоритмов, оптимизация поиска увеличивается в несколько раз. Выдавая ранжированные списки результатов запроса, полнотекстовый поиск помогает сократить время на выборку требуемой информации. В масштабах предприятий небольшая, с технической точки зрения, модернизация модуля поиска, позволяет повысить производительность труда в целом.

Таблица 1. Ранжирование таблиц

Название таблицы	Предикат Containstable	Предикат Freetexttable
Class._Геолого-технические мероприятия	1	65
Class._Геолого-технические мероприятия	2	65
Class._Геолого-технические мероприятия	3	65
Class. Мероприятия в процессе работ	7	21
Class. Мероприятия в процессе работ	8	21
Class. Мероприятия в процессе работ	9	21
Class. Мероприятия в процессе работ	10	21

Список литературы

1. Ковин Р.В., Копнов М.В., Кудинов А.В., Мирошниченко Е.А., Шерстнев В.С. Корпоративная геоинформационная система для управления производством ОАО “Востокгазпром” //Труды межрегиональной конференции “Газораспределительные системы. АГНКС. АГЗС. Проектирование. Строительство. Эксплуатация.” - Томск, 24-26 сентября 2003. - : , 2003. - с. 112-119 (36752207)

2. Богдан С.А., Кудинов А.В., Марков Н.Г. Опыт внедрения MES «Магистраль-Восток» в нефтегазодобывающей компании //Автоматизация в промышленности, 2010. -№ 8 -с. 53–58 (66276283)

3. Обзор SQL Server. [Электронный ресурс]. URL: http://msdn.microsoft.com/ru-ru/library/ms166352(v=sql.90).aspx (дата обращения 10.02.2014)

4. Блог о технологиях .NET. [Электронный ресурс]. URL:http://djekmusic.blogspot.ru/2012/03/ms-sql-2008-2.html

(дата обращения: 10.02.2014)

5. Okapi BM25: a non-binary model. [Электронный ресурс]. URL: http://nlp.stanford.edu/IR-book/html/htmledition/ okapi-bm25-a-non-binary-model-1.html (дата обращения: 10.02.2014)