ПОДХОД К ОРГАНИЗАЦИИ КОНТЕКСТНОГО ПОИСКА В СИСТЕМЕ ЭЛЕКТРОННОГО ДОКУМЕНТООБОРОТА С ПРИМЕНЕНИЕМ XML-ПРЕДСТАВЛЕНИЯ ФРЕЙМОВОЙ МОДЕЛИ ДОКУМЕНТА

Керносов М.А.
Харьковский национальный университет радиоэлектроники
61166, Харьков, пр. Ленина,14, каф. Информационных управляющих систем,
тел. (057) 702-14-51, E-mail: iasu@kture.kharkov.ua

The approach to organization of contextual searching in Electronic Document Management System (EDMS) is offered. For work with electronic document in EDMS is expected using as text, so and formalized presentations of the document. For formalization of document structure the frame-based document model is applied. This model can be easy presented as an XML-document, describing structure of the electronic document and containing contributed in it factographic data. Such XML-presentation allows to organize speediest contextual searching for in formalized presentations of the document due to exception of redundant information.

Современные системы электронного документооборота (СЭД) обеспечивают поддержку электронных документов, представленных в различных текстовых и графических форматах. Поэтому в настоящее время приобретают актуальность исследования способов организации быстрого и эффективного поиска в СЭД, оперирующей гетерогенными документами.

Для унификации работы с электронными документами в СЭД, как правило, помимо текстового или графического применяется также формализованное представление документа. Данное представление может быть описано на языке XML и использовано для решения задачи разработки подхода к организации контекстного поиска в СЭД, позволяющего не только организовать контекстный поиск в документах различных форматов, но и обеспечить высокую скорость его осуществления.

Одним из способов формализации структуры документа является применение фреймовой модели документа, согласно которой набор фреймов, связанных между собой отношениями подчиненности, образует структуру документа, включающую фактографические данные, содержащиеся в документе [1]. Поскольку фрейм представляет собой структуру данных и функциональность для работы с ними (реализуется в виде присоединенных процедур), то такая модель документа может быть легко представлена в виде XML-документа. Кроме того, данная модель адаптирована к реализации в среде объектно-ориентированного программирования, что снижает трудоемкость разработки и сопровождения СЭД, и может быть декомпозирована на составные части. Поэтому при разработке подхода к организации контекстного поиска в СЭД фреймовая модель документа выбрана в качестве базовой.

Формализованная структура документа, описанная на языке XML, будет содержать структурированные фактографические данные, которые в большинстве случаев включаются в запросы, применяемые для контекстного поиска документов, и являются критерием анализа хранимых документов. Задача формирования фреймовой модели электронного документа в соответствии с его текстовым или графическим представлением может быть решена вследствие документа анализа или непосредственно на этапе его разработки.

Таким образом, предлагаемый подход к организации контекстного поиска в СЭД будет включать в себя следующие два основных этапа: выполнение запроса пользователя СЭД применительно к хранилищу формализованных XML-представлений электронных документов и, в случае неудовлетворенности пользователя полученными результатами, выполнение контекстного поиска непосредственно в электронных документах.

Особенности реализации второго этапа данного подхода зависят от специфики применяемых текстовых и(или) графических форматов представления документов. Одним из вариантов организации полнотекстового поиска также может быть создание

унифицированных текстовых копий документов, лишенных разметки, форматирования и прочей служебной информации, специфической для конкретного формата представления документа.

На практике анализ предметной области и типовых запросов пользователей в большинстве случаев позволяют ограничиться только выполнением первого этапа данного подхода. При этом поиск в хранимых XML-представлениях документов, описывающих их структуру, обладает рядом преимуществ, позволяющих повысить его эффективность и обеспечить быстроту выполнения.

XML-представление фреймовой модели электронного документа содержит фактографические данные, которые были внесены в документ, описывает его структуру и не содержит избыточной информации. Поэтому его размер в несколько раз меньше размера исходного электронного документа, что позволяет значительно увеличить скорость контекстного поиска. Кроме того, современные СУБД, например Oracle версии 8 и выше, поддерживают специфический тип данных для работы с XML-документами (например, XMLType), обеспечивают проверку корректности их формирования (XMLSchema), а также позволяют разбить XML-документ на секции и осуществить их индексацию [2], что позволяет в еще большей степени повысить скорость контекстного поиска. При этом индексация секций XML-документа позволяет индексирование только данные, содержащихся в конкретных тегах XML-документа, исключая сами теги. Вследствие этого при выполнении запросов к индексированным секциям, т.е. к определенной части структуры XML-документа, не обрабатывается текст всего XML-документа, а осуществляется поиск только в проиндексированных данных [3].

Создание контекстных индексов (CONTEXT INDEX) для полей таблиц, содержащих XML-документы, дает возможность использования SQL-оператора CONTAINS, указывающего анализируемую секцию XML-документа, в условии WHERE относящемся к SELECT-запросу на выборку данных (рис. 1).

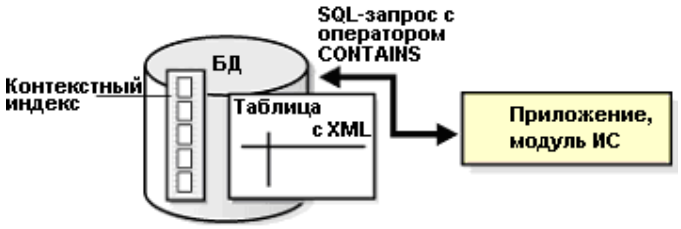

Рис. 1 – Механизм использования контекстных запросов в приложении

Данная иллюстрация представляет собой обобщённую схему выполнения запросов приложения, осуществляющих контекстный поиск в хранимых XML-документах. Контекстный индекс и таблица, содержащая XML-документы, являются частью БД, а двусторонняя связь между БД и приложением обозначает SQL-запрос, содержащий оператор CONTAINS [4].

Алгоритм работы такого приложения будет состоять из следующих шагов, большая часть которых реализуется без участия пользователя СЭД:

– пользователь выполняет действия, вызывающие формирование приложением запроса для контекстного поиска;
– СУБД выполняет CONTAINS-запрос;
– приложение отображает пользователю список документов, соответствующих запросу. Этот список обычно сортируется по рангу документов, отображающему степень соответствия документа запросу пользователя;
– пользователь выбирает документ из предоставленного списка;
– приложение предоставляет пользователю документ для просмотра [4].

Предложенный подход к организации контекстного поиска в СЭД с применением XML-представления фреймовой модели документа позволяет обеспечить практическую реализацию высокоскоростного контекстного поиска архиве электронных документов СЭД с учетом их структуры. Фреймовая модель структуры документа позволяет снизить трудоемкость реализации иерархию классов для работы с формализованными представлениями электронных документов в средах объектно-ориентированного программирования. Применение технологии XML для формализованного описания электронных документов позволяет не только возможности современных СУБД по обработке XML-документов, но и использовать XML-представления документами для обмена данными СЭД с другими информационными системами.

Литература

1. Левыкин В.М., Керносов М.А. Исследование и разработка фреймовой модели структуры документа // Нові технології. – 2008. – № 1 (19). – С. 149-154.
2. Керносов М.А. Применения технологии XML для хранения структурированных данных в объектно-реляционных системах управления базами данных // Управління розвитком. – 2006. – № 1 – С. 42-43.
3. Том Кайт. Oracle для профессионалов. Книга 2. Расширение возможностей и защита: Пер. с англ. – К.: ООО «ТИД «ДС», 2003. – 848 с.
4. Oracle® Text Application Developer's Guide 10*g* Release 2 (10.2), Part Number B14217-01 (Oracle Database 10*g* Documentation Library), www.oracle.com.