Шерстнёв Владислав Станиславович, канд. техн. наук, доцент кафедры вычислительной техники Института кибернетики ТПУ.
E-mail: vss@tpu.ru
Область научных интересов: разработка программного обеспечения.
Иванов Сергей Сергеевич, магистрант кафедры вычислительной техники Института кибернетики ТПУ.
E-mail: vss@tpu.ru
Область научных интересов: разработка программного обеспечения.
Акулин Иван Алексеевич, кафедры вычислительной техники Института кибернетики ТПУ.
E-mail: vss@tpu.ru
Область научных интересов: разработка программного обеспечения.

УДК 004.45:004.62

ИСПОЛЬЗОВАНИЕ ORACLE UNIVERSAL CONTENT MANAGMENT В КАЧЕСТВЕ КОРПОРАТИВНОГО ХРАНИЛИЩА ДОКУМЕНТОВ ТПУ

В.С. Шерстнёв, С.С. Иванов, И.А. Акулин

Томский политехнический университет
E-mail: vss@tpu.ru

Применительно к Томскому политехническому университету рассмотрены вопросы доступности корпоративной информации через собственные системы поиска информации ТПУ. Описаны варианты решения задачи консолидации и поиска корпоративной информации. Предположено, что наиболее целесообразным вариантом для управления накопленной информацией является создание централизованного корпоративного хранилища информации. Показана архитектура хранилища информации и роль взаимодействующих с ней модулей. Описаны перспективные задачи, решаемые на основе предложенной к использованию программной платформы.

Ключевые слова:
Oracle Universal Content Server, управление контентом, СУБД, полнотекстовый поиск, хранилище информации.

Key words:
Oracle Universal Content Server, content management, DBMS, full-text search, data warehouse.

С ходом времени и постепенного развития любая организация накапливает свой интеллектуальный капитал. В настоящее время весомой частью накопленных материалов организации является разнородная документация в электронном виде. Томский политехнический университет обладает огромными информационными электронными ресурсами, в связи с этим, задача эффективного использования и хранения электронных документов является актуальной для ТПУ.

Под эффективным использованием электронных документов стоит понимать их доступность для просмотра всем авторизованным заинтересованным лицам. Под эффективным хранением предлагается понимать хранение не только текущей версии электронных документов, но и предыдущих версий этих документов.

На данный момент электронные ресурсы Томского политехнического университета рассредоточены по следующим основным массивам:

- интернет-сайты ТПУ («Абитуриенту ТПУ», «Инновационная образовательная программа ТПУ на 2007–2008 гг.», «Известия ТПУ» и еще более 20 сайтов);
- интернет-решения на основе корпоративного портала ТПУ (фонд образовательных программ и т. д.)
- научно-техническая библиотека ТПУ;
- файловые ресурсы и базы данных отдельных подразделений ТПУ.

В целом, все эти документы форматов html, pdf, doc, xls, ppt и т. д., являются накопленной интеллектуальной собственностью ТПУ, но доступность их достаточно

затруднительна, что снижает их ценность. Основная проблема доступности документов заключается в невозможности быстрого и интуитивно понятного поиска по всем разнородным документам, хранимым на различных серверах в различных системах и форматах.

Существующие глобальные информационные сервисы не могут быть полезны в данном случае. Такие поисковые сервисы как Google и ему подобные не могут проиндексировать электронные документы, хранящиеся в корпоративной сети ТПУ и недоступные для неавторизованных пользователей.

Таким образом, ниша инструмента для поиска документов в корпоративной сети ТПУ остается свободной. Необходимый инструмент поиска должен обладать возможностью производить поиск не только в содержимом интернет-документов (html и т. д.), но и в содержимом документов распространенных офисных приложений (Microsoft Word, Excel, Power Point, Adobe Acrobat и т. п.). Задача поиска информации в содержимом документов носит название задачи «полнотекстового поиска» и активно решается современными средствами по управлению данными [1–2].

Для решения задачи поиска информации в корпоративной сети ТПУ возможны несколько вариантов, схематично представленных на рис. 1, 2. Первый вариант (рис. 1) можно назвать внешней системой поиска по распределенным массивам данных. В этом варианте для поиска документов используются подсистемы поиска встроенные в существующие массивы хранения электронных ресурсов: подсистема поиска по электронным ресурсам научно-технической библиотеки, подсистема поиска по содержимому корпоративного портала ТПУ, подсистемы поиска по собственным файловым ресурсам институтов ТПУ.

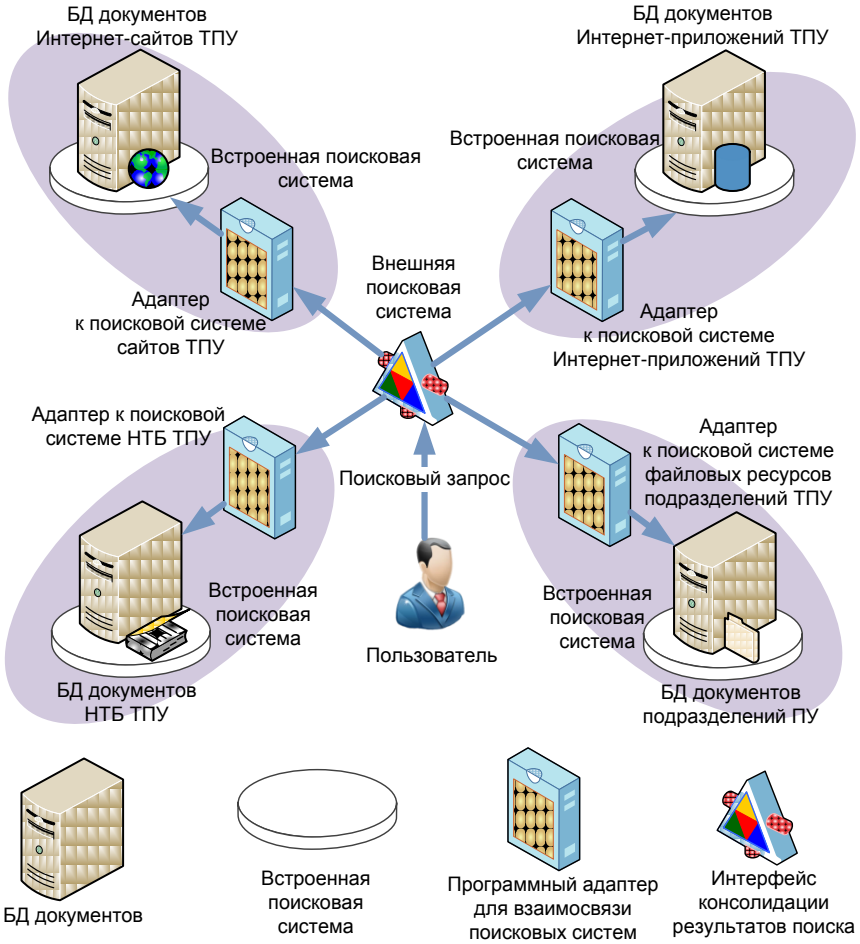

Рис. 1. Архитектура внешней системы поиска по децентрализованным массивам информации

Достоинством такого варианта является минимальная модификация существующих подсистем поиска и хранения данных. К недостаткам этого варианта можно отнести следующие особенности.

- Разный уровень надежности хранения информации. Файловые ресурсы подразделений ТПУ часто хранятся на собственных серверах подразделений и не обладают технической возможностью для формирования резервных копий данных для их восстановления после отказа.
- Использование разнообразных подсистем поиска. Для каждого существующего и вновь появляющегося электронного массива информации в подразделениях ТПУ потребуется реализовывать дополнительный программный интерфейс (адаптер), обеспечивающий интеграцию локальной системы поиска в данном массиве информации с общей (вышестоящей) системой поиска информации.
- Отсутствие возможности полнотекстового поиска практически во всех вышеупомянутых подсистемах поиска по массивам хранения электронных ресурсов. Разработка и внедрение в существующие массивы электронных ресурсов оригинальных механизмов полнотекстового поиска является достаточно трудоёмкой задачей и полностью отменяет имеющиеся достоинства данной архитектуры.
- Сложность внедрения существующей системы разграничения доступа на всех разнородных серверах подразделений ТПУ.

Таким образом, для устранения вышеперечисленных недостатков потребуется разработка достаточно большого объема частных программных решений, что в целом отменяет существующие достоинства данного варианта.

Другой вариант решения задачи поиска информации можно назвать системой поиска с централизованным хранением (рис. 2). В данном варианте предлагается перенести все электронные ресурсы в единую среду хранения.

Рис. 2. Архитектура системы поиска с использованием централизованного массива информации

Недостатком такого варианта можно считать проведение разовой операции переноса имеющихся разнородных электронных ресурсов в новый массив хранения информации. Тем не

менее, второй вариант, с точки зрения авторов, выглядит более перспективным, так как обладает следующими преимуществами:

- присутствует единый уровень надежности хранения информации. Вся документация хранится на единой высокопроизводительной и отказоустойчивой системе хранения, поддерживаемой специально подготовленным персоналом (Главный информационный узел ТПУ);
- используется единая подсистема полнотекстового поиска по электронным документам;
- единожды интегрируется с существующей в ТПУ системой разграничения прав доступа к информации.

Вышеописанные варианты, в некотором приближении можно сравнить с существовавшими ранее «одноранговыми» сетями, где производственная информация была бессистемно распределена по всем компьютерам, и более развитыми существующими ныне сетями «с выделенным сервером», на котором и хранится вся необходимая информация. В таком сравнении перспективность использования второго варианта становится еще очевидней.

На сегодня в линейке программного обеспечения, решающего задачу хранения и организации доступа к разнородным документам на уровне предприятия, есть несколько программных продуктов [4–7]. Наиболее развитыми из существующих, с точки зрения авторов, относятся Microsoft SharePoint Server [4], Oracle Universal Content Management [5]. С учётом того, что ТПУ обладает правом использования многих продуктов Oracle (в том числе и Oracle Universal Content Management), именно этот продукт был принят за основу для построения корпоративного хранилища документов.

С учетом избранного в качестве основы Oracle Universal Content Management (Oracle UCM), архитектура системы выглядит так, как представлена на рис. 3. Oracle UCM выполняет роль ядра системы и практически не обладает пользовательским интерфейсом. Пользователи взаимодействуют с интерфейсом корпоративного портала или интерфейсом портальных приложений. Последние, в свою очередь, обращаются к хранилищу с запросами на работу с документами.

Внешние приложения могут общаться с хранилищем с помощью протокола SOAP (Simple Access Object Protocol), RIDC (Remote IntraDoc Client) [7, 8]. Посредством SOAP или RIDC-запросов внешние приложения могут производить все необходимые операции с документами хранилища:

- сохранение новых документов;
- внесение новой версии ранее созданного документа;
- выбор документа или одной из его версий;
- выбор множества документов, как результат поискового запроса.

Для организации взаимодействия портальных приложений с корпоративным хранилищем разработаны библиотеки на языке PL/SQL. Используя эти библиотеки, существующие приложения смогут пользоваться услугами корпоративного хранилища и сохранять в нём те файловые данные, что ранее сохранялись в базе данных (БД). Наличие соответствующих PL/SQL библиотек позволит вынести в доступное информационное пространство ТПУ те документы, что прежде являлись содержимым БД.

Тем не менее, задача поиска информации не решается полностью, так как в корпоративной сети ТПУ остаются ресурсы, которые не могут быть перенесены в Oracle UCM. Например, такой информацией является содержимое электронных каталогов научно-технической библиотеки (НТБ) ТПУ. Электронные каталоги НТБ содержат не сами электронные документы, а только метаданные о них: текстовые описания, ключевые слова и т. д. Реализация данного электронного каталога на платформе Oracle UCM не целесообразна, так как представляет собой законченную информационную систему, оптимизированную для работы с библиографической описательной информацией и разработанную в соответствии с международными требованиями к программам в этой предметной области.

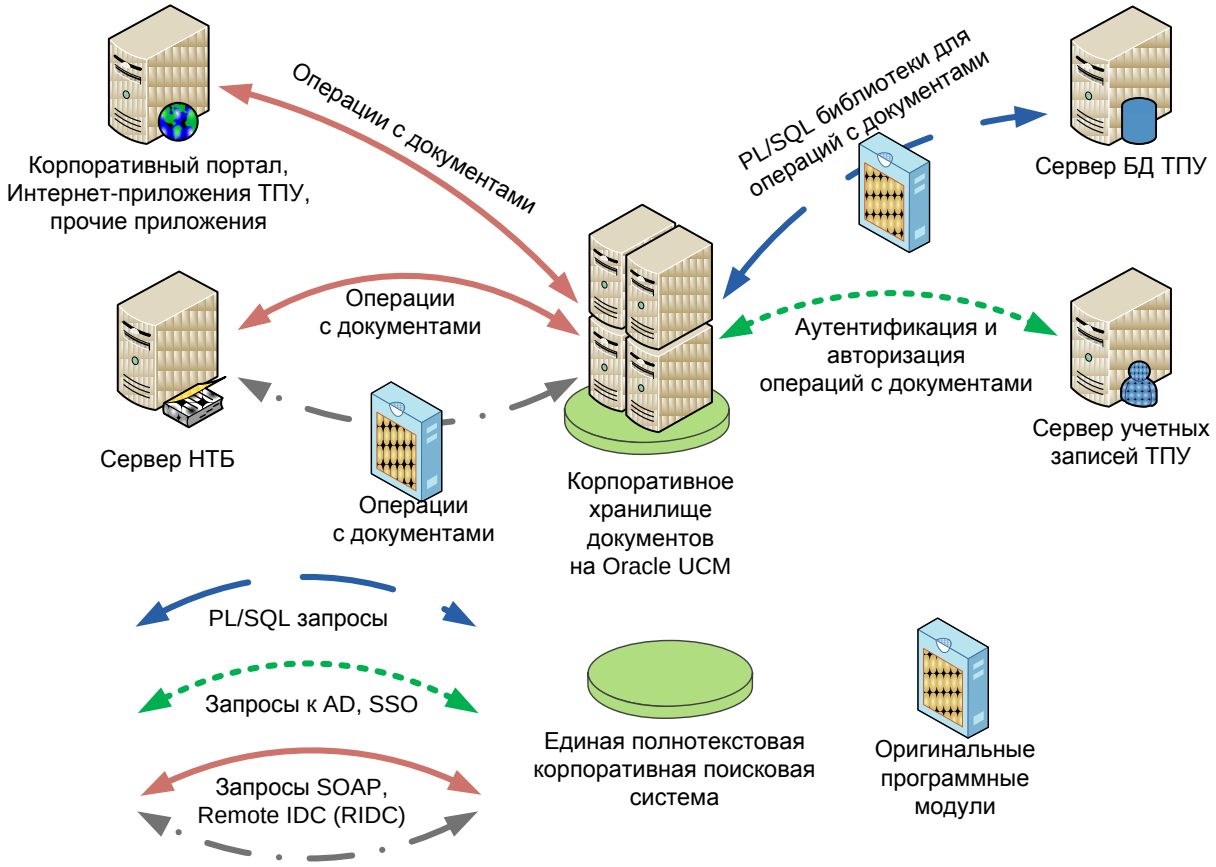

Рис. 3. Oracle Universal Content Management, как центральный элемент корпоративного хранилища документов

Для успешного поиска документов в информационном пространстве ТПУ требуется задействовать и поиск информации в этом каталоге. Интеграцию электронного каталога публикаций НТБ с поисковой системой Oracle UCM предлагается решить за счёт разработки соответствующих SOAP-адаптеров, транспортирующих поисковые запросы и ответы на них между подсистемой поиска электронного каталога НТБ и поисковой системой Oracle UCM.

Таким образом, в предлагаемой архитектуре появляются соответствующие программные SOAP-адаптеры, обеспечивающие интеграцию содержимого различных баз данных ТПУ в доступное для осуществления поиска информационное пространство.

Полученные результаты: предложена концепция организации единого корпоративного хранилища документов ТПУ, рассмотрены возможности интеграции хранилища с существующими системами, реализованы соответствующие программные интерфейсы и библиотеки для интеграции Oracle UCM и некоторых корпоративных приложений, рассмотрены перспективы работы с корпоративным хранилищем.

Выводы

Показана актуальность использования единого корпоративного хранилища документов. Предположено, что наиболее целесообразным вариантом для управления накопленной информации является создание централизованного корпоративного хранилища информации. Показана архитектура хранилища информации и роль взаимодействующих с ней модулей. Описаны перспективные задачи, решаемые на основе предложенной к использованию программной платформы.

СПИСОК ЛИТЕРАТУРЫ

1. Ермаков Е.А. Полнотекстовый поиск: проблемы и их решение // Журнал для пользователей персональных компьютеров «Мир ПК». 2001. URL: http://www.osp.ru/pcworld/2001/05/161575 (дата обращения 01.11.2011).
2. Лавренова О. Полнотекстовый поиск // Российская ассоциация электронных библиотек. 2009. URL: http://www.aselibrary.ru/blogs/archives/27/ (дата обращения 01.11.2011).
3. Корпорация Microsoft. Управление контентом в SharePoint 2010 // Интернет-сайт Microsoft® SharePoint® 2010. 2010. URL: http://sharepoint.microsoft.com/ru-ru/product/capabilities/content/Documents/FINAL%20Business%20Value%20of%20ECM%20Whitepaper.pdf (дата обращения 01.11.2011).
4. Michelle Huff, Brian Dirking. The Benefits of a Unified Enterprise Content Management Platform // Интернет-сайт Oracle Corporation. 2011. URL: http://www.oracle.com/technetwork/middleware/webcenter/content/wp-owc-benefits-ecm-platform-427847.pdf?ssSourceSiteId=ocomen (дата обращения 01.11.2011).
5. Компания Cognitive Technologies. Система электронного документооборота E1 Евфрат // Интернет-сайт компании Cognitive Technologies. 2011. URL: http://www.evfrat.ru/about/ (дата обращения 01.11.2011).
6. Brian Proffitt. Alfresco: An open-source ECM alternative for SharePoint // Электронный журнал ComputerWorld. 2011. URL: http://www.computerworld.com/s/article/9221265/Alfresco_An_open_source_ECM_alternative_for_SharePoint (дата обращения 01.11.2011).
7. World Wide Web Consortium (W3C). SOAP Specifications // Интернет сайт консорциума WWW. 2007. URL: http://www.w3.org/TR/soap/ (дата обращения 01.11.2011).
8. Oracle Corporation. Oracle Remote Intradoc Client // Интернет-сайт корпорации Oracle. 2008 г. URL: http://download.oracle.com/docs/cd/E10316_01/ ContentIntegration/ridc/ridc-developer-guide.pdf (дата обращения 01.11.2011).

Поступила 26.09.2010 г.