ISSN 1992-6502 (Print) Вестник УГАТУ ISSN 2225-2789 (Online)
2014. Т. 18, № 2 (63). С. 85–92 http://journal.ugatu.ac.ru

УДК 004.65

СЕМАНТИЧЕСКАЯ ИНТЕГРАЦИЯ АВТОМАТИЗИРОВАННЫХ И ПОИСКОВЫХ СИСТЕМ В ПРОСТРАНСТВЕ ИССЛЕДУЕМОЙ ПРЕДМЕТНОЙ ОБЛАСТИ НА ОСНОВЕ СИСТЕМНЫХ МОДЕЛЕЙ И РОЛЕВОЙ МОДЕЛИ ДОСТУПА

А. А. БАРМИН[1], **Г. В. СТАРЦЕВ**[2], **С. Ф. БАБАК**[3], **Г. Г. КУЛИКОВ**[4]

[1] abarmin@outlook.com, [2] startsev@gmail.com, [4] molniya@molniya-ufa.ru

[1–3] ФГБОУ ВПО «Уфимский государственный авиационный технический университет» (УГАТУ)
[4] ОАО УНПП «Молния»

Поступила в редакцию 30.07.2013

Аннотация. Рассмотрены проблемы построения и интеграции информационно-поисковых систем с корпоративными информационными системами. Описывается модель информационного запроса и результатов поиска с учетом прав доступа пользователя и его личных предпочтений. Рассматривается реализация предложенной модели в системе электронного документооборота, построенной на платформе IBM Lotus Domino.

Ключевые слова: бизнес-процесс; модель предметной области; информационный поиск; ролевая модель доступа.

Современная цивилизация находится в стадии формирования информационного общества – особого общества, в котором большинство работающих занято производством, хранением, переработкой и реализацией информации, а также высшей ее формы – знаний. Для этой стадии развития общества и экономики характерно:

- увеличение роли информации, знаний и информационных технологий в жизни общества;
- возрастание числа людей, занятых информационными технологиями, коммуникациями и производством информационных продуктов и услуг;
- создание глобального информационного пространства, обеспечивающего эффективное взаимодействие людей, их доступ к мировым информационным ресурсам, удовлетворение потребностей в информационных продуктах и услугах.

Информационная система организации в процессе своего функционирования накапливает значительный объем данных, так что встает вопрос оперативного поиска информации. Многие компоненты информационной системы содержат встроенные механизмы поиска, но они позволяют осуществлять поиск только в рамках отдельных аспектов одной локальной системы. Для организации поиска по всему массиву доступной информации применяются информационно-поисковые системы, которые позволяют осуществлять поиск неструктурированной документальной информации на основе сформированных пользователем запросов. Запрос пользователя может быть представлен как в виде четко заданного логического выражения, так и в виде словосочетания или фразы на естественном языке.

Сложность поиска в корпоративной информационной системе обусловлена наличием различных источников и способов представления данных, необходимостью единообразного ранжирования результатов для различных представлений данных – веб-страниц, документов, вложенных в документы файлов и других форм представления данных. Также сложность поиска в корпоративной информационной системе обусловлена необходимостью дальнейшей обработки полученных данных, а не только предоставлением этих данных пользователю.

Цель данной статьи – рассмотреть подход к интеграции корпоративных информационных систем и информационно-поисковых систем на основе системных моделей и ролевой модели доступа.

Для реализации поставленной цели необходимо выполнение следующих задач:

1. Провести анализ рынка информационно-поисковых систем для анализа текущей ситуации;

2. Предложить математическую модель процесса поиска с учетом ролей пользователя в бизнес-процессах организации.

Также в статье рассматривается пример реализации предложенной модели на основе системы электронного документооборота на платформе IBM Lotus Domino.

СОСТОЯНИЕ ВОПРОСА

Пользователи современных корпоративных информационных систем создают большой объем информации. Электронные письма, документы в электронном виде, аудио- и видеозаписи, файлы, системы автоматизированного проектирования используются различными компонентами информационной системы организации. Эта информация хранится в базах данных, файловых серверах, электронных архивах и рабочих станциях пользователей. Все эти документы являются накопленной интеллектуальной собственностью организации, но их доступность затруднительна, что снижает их ценность. Проблема доступности документов заключается в невозможности быстрого и интуитивно-понятного поиска по разнородным документам, находящимся на различных серверах в различных системах и форматах.

Существующие глобальные информационно-поисковые сервисы не могут быть полезны в данном случае, так как не могут проиндексировать информацию, находящуюся в локальной сети организации. Инструмент корпоративного поиска должен обладать возможностью поиска не только по содержимому интернет-документов, но и по содержимому документов распространенных форматов офисных документов. Задача поиска информации в содержимом документе является задачей полнотекстового поиска и решается с помощью корпоративных поисковых систем и систем управления данными.

К решению задачи информационного поиска в корпоративной среде есть несколько подходов. Первый подход заключается в использовании внешней системы поиска по распределенным массивам данных. В данном подходе используются информационно-поисковые подсистемы, встроенные в информационные системы корпоративной среды: подсистема поиска системы электронного документооборота, подсистема поиска файлового хранилища, подсистема поиска бухгалтерской системы и другие (рис. 1).

Достоинством данного подхода является минимальная модификация существующих подсистем поиска и хранения данных. К недостаткам можно отнести:

- разный уровень надежности хранения информации;
- использование разнообразных подсистем поиска. Для каждого из приложений корпоративной информационной системы будет необходимо реализовать программный интерфейс для интеграции с поисковой системой высшего уровня;
- разные уровни детализации результатов поиска и механизмы поиска. Для интеграции разнородных подсистем поиска потребуется преобразование результатов поиска к единому формату;
- сложность ранжирования результатов. Результаты поиска в каждой из подсистем ранжированы в соответствии с собственными критериями [1].

Второй подход заключается в использовании централизованной корпоративной информационно-поисковой системы. Выделенная подсистема корпоративной информационной системы индексирует всю информацию, находящуюся на серверах и рабочих станциях локальной сети предприятия и выполняет роль единого интерфейса поиска информации (рис. 2).

Достоинствами данного подхода являются:

- централизация функций поисковой системы;
- единая информационно-поисковая система позволяет выполнять ранжирование документов по одинаковым критериям для всех подсистем-источников данных;
- единая информационно-поисковая система предоставляет унифицированный формат представления результатов.

Тем не менее, подход с использованием централизованной информационно-поисковой системы обладает рядом недостатков:

- необходимость периодической индексации документов каждой из подсистем;
- документы в результаты поиска должны отбираться в соответствии с правами доступа пользователя, выполняющего поисковый запрос.

В настоящий момент существуют поисковые сервера, способные решить задачу организации подсистемы информационного поиска.

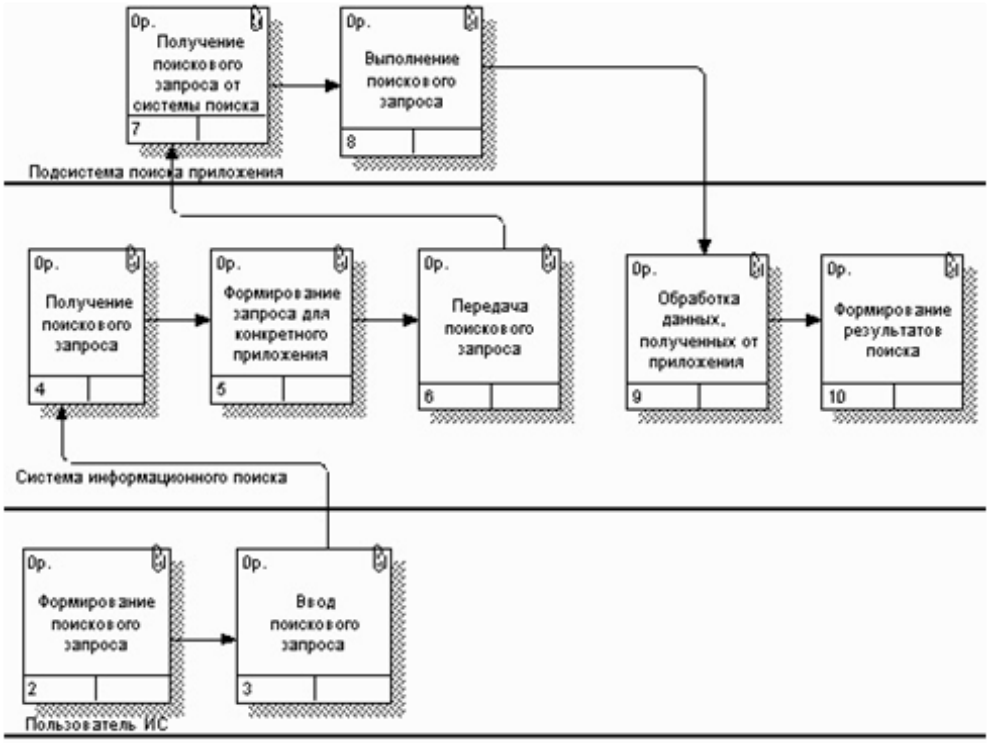

Рис. 2. Использование встроенной подсистемы поиска корпоративного приложения

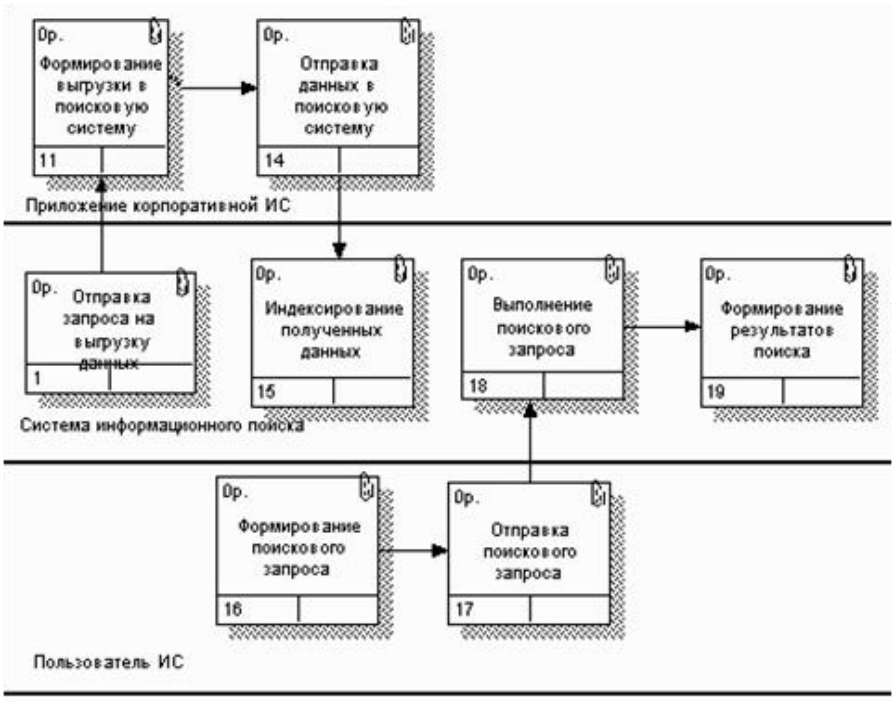

Рис. 2. Подход с использованием централизованной корпоративной информационно-поисковой системы

К таким системам относятся: Google Custom Search, Яндекс.Сервер, Apache Lucene, Apache Solr, Sphinx и др. Данные системы позволяют осуществлять полнотекстовый и фасетный поиск по файловым хранилищам, электронным документам и базам данным.

Сравнительная характеристика система информационного поиска приведена в табл.

Таблица

Сравнительная характеристика систем информационного поиска

Наименование	Тип	Поддерживаемые типы файлов и хранилищ данных	Поддержка русского языка, наличие документации	Условия использования
Google Custom Search	Облачный сервис	Только веб-страницы	Русский язык поддерживается, обширная документация	Платный продукт, есть ограниченная бесплатная версия
Google Search Appliance	Программно-аппаратное решение	Доработка специалистами Google под любые форматы	Русский язык поддерживается, обширная документация	Платный продукт
Google Desktop	Пользовательское программное обеспечение	Документы MS Office, электронная почта на сервисе Google Mail	Русский язык поддерживается	Бесплатный продукт
Яндекс. Персональный поиск	Пользовательское программное обеспечение	Документы MS Office, электронная почта на сервисе Yandex Mail	Русский язык поддерживается, поиск с учетом морфологии	Бесплатный продукт
Яндекс.Сервер	Серверное программное обеспечение	Документы MS Office, PDF, изображения, реляционные хранилища данных	Русский язык поддерживается, поиск с учетом морфологии	Бесплатно для некоммерческого использования и образовательных учреждений
Sphinx	Серверное программное обеспечение	Реляционные хранилища данных	Русский язык поддерживается, поиск с учетом морфологии	Бесплатно для некоммерческого использования
Apache Lucene	Набор библиотек для разработки поисковой системы	Только текст	Поддержка с использованием сторонних модулей	Свободное программное обеспечение
Apache Solr	Серверное программное обеспечение	Только XML	Поддержка с использованием сторонних модулей	Свободное программное обеспечение

ПРОБЛЕМА ПОИСКА В ГЕТЕРОГЕННОЙ СРЕДЕ

Корпоративные информационно-поисковые системы требуются в первую очередь компаниям, чья деятельность связана с обработкой данных. Обычно такие компании имеют гетерогенную IT-среду: помимо работы с документами на обычных файловых серверах сотрудники постоянно работают с базами данных, CRM-системами, ERP-системами, внутренним порталом. Без внедрения общей системы поиска по этим ресурсам получить информацию из всех источников одновременно не получится.

По результатам исследования, проведенного компанией IDC, 38 % своего времени офисный сотрудник тратит на поиск информации, причем 21 % из этого времени уходит на подбор нужных документов, а 17 % тратится безрезультатно – на неудачный поиск и создание не найденной информации повторно [2].

Круг пользователей корпоративной информационной системы ограничен сотрудниками организации. Защита, распределение и доступ пользователей к информации регламентируется политикой безопасности организации – совокупностью руководящих правил, принципов, процедур и практических приемов в области безопасности. Политика безопасности зависит:

- от конкретной технологии обработки информации;
- используемых технических и программных средств;
- расположения организации и других условий его функционирования.

Далеко не все источники корпоративной информации являются открытыми. Скорее, наоборот, – большая их часть составляет коммер-

ческую тайну. Соответственно, поисковое решение должно учитывать систему прав доступа.

В соответствии с существующими подходами принято считать, что информационная безопасность ИС обеспечена в случае, если для любых информационных ресурсов в системе поддерживается определенный уровень:

1. Конфиденциальности (невозможности несанкционированного получения информации);
2. Целостности (невозможности несанкционированной ее модификации);
3. Доступности (возможности за разумное время получить требуемую информацию)[3].

Кроме того, есть ряд факторов, которые усложняют интеграцию информационно-поисковых и корпоративных систем:

1. Циркулирующие в организации документы могут иметь различную структуру, информационно-поисковая система должна одинаково хорошо работать как со структурированным, так и со слабоструктурированным контентом;
2. Существующие политики доступа в организации должны учитываться при формировании результатов поискового запроса: пользователь должен получать в поисковой выдаче только те документы, к которым он имеет доступ;
3. Наличие собственных адаптеров для интеграции установленных систем осложняет использование информационно-поисковых систем: многие компоненты корпоративной системы имеют интерфейс для взаимодействия с другими системами. Использование типовых решений позволяет их интегрировать без доработок или с минимальными изменениями, в то время как использование собственных нестандартных решений исключает такую возможность;
4. Ограничение доступа на уровне бизнес-правил: пользователям могут потребоваться документы, к которым они не имеют доступа, но создаваемые в бизнес-процессах, в которых пользователи являются непосредственными участниками.

В данной работе предлагается подход к интеграции корпоративных информационных систем и информационно-поисковых систем на основе системных моделей бизнес-процессов и ролевой модели доступа.

ПОДХОД К ИНТЕГРАЦИИ НА ОСНОВЕ СИСТЕМНЫХ МОДЕЛЕЙ БИЗНЕС-ПРОЦЕССОВ И РОЛЕВОЙ МОДЕЛИ ДОСТУПА

Комплект системных моделей, в соответствии с методологией SADT, включает в себя функциональную, информационную и динамическую модели. Каждая из этих моделей описывает реальную систему в определенном аспекте: как систему элементов и отношений, систему сущностей и связей, причинно-следственные связи системы, семантический аспект системы.

Представим функциональную модель в виде следующей теоретико-множественной модели:

$$B = (I, C, O, M), \quad (1)$$

где F – совокупность функций модели, I – входные данные, C – нормативные документы, O – выходные данные, M – механизмы и исполнители.

В качестве механизмов могут выступать пользователи, информационные системы и роли.

$$M = \{U_B, R_B, S_B\}, \quad (2)$$

где U_B – пользователь-исполнитель бизнес-процесса, R_B – роль-исполнитель бизнес-процесса, S_B – система-исполнитель бизнес-процесса.

Представим каждого пользователя информационной системы в виде следующей модели:

$$U = \{R_U, S_U\}, \quad (3)$$

где R_U – роли, которыми обладает пользователь в рамках всех бизнес-процессов организации, S_U – семантическая информация о пользователе, например, ФИО, адрес электронной почты и др.

Тогда бизнес процессы, к которым пользователь имеет доступ, описываются следующей моделью:

$$B_U = R_U \cap R_B, \quad (4)$$

где B_U – бизнес-процессы, участником которых пользователь является.

В случае разделения бизнес-процессов по отдельным хранилищам данных, подсистемам, базам данных (4) представляет собой набор подсистем, к которым пользователь имеет доступ и данные, из которых должны включаться в результаты пользовательского запроса.

Представим информационную модель в виде следующей теоретико-множественной модели:

$$M = \{E, R\}, \quad (5)$$

где M – информационная модель предметной области, E – сущности предметной области,

R – отношения между сущностями предметной области.

Представим сущность предметной области в виде совокупности ключевых, неключевых атрибутов и атрибутов контроля доступа:

$$R = \{A_K, A_S\}, \quad (6)$$

где A_K – ключевые атрибуты сущности, A_S – неключевые атрибуты.

Неключевые атрибуты могут содержать как непосредственно данные, так и списки контроля доступа:

$$A_S = \{A_D, A_A\}, \quad (7)$$

где A_D – данные, A_A – список контроля доступа.

Списки контроля доступа ограничивают видимость записей для конкретных групп контроля доступа, ролей и пользователей. При использовании списков контроля доступа мы можем выделить конкретные экземпляры сущностей, к которым каждый конкретный пользователь может иметь доступ.

$$E_U = \{E|A_A \cap U\}, \quad (8)$$

где E_U – экземпляры сущностей, доступные пользователю, т. е. только те экземпляры, в полях контроля доступа которых есть упоминание об указанном пользователе.

Таким образом, на основе (7) и (4) запрос, который должна выполнять информационно-поисковая система, можно представить в виде следующей модели:

$$Q = B_U \cap E_U \cap Q_{\mathrm{U}}, \quad (9)$$

где Q – поисковый запрос, выполняемый информационно-поисковой системой, Q_U – поисковый запрос, сформированный пользователем.

Рассмотрим приведенную выше модель на примере системы электронного документооборота Логика ECM.СЭД версии 3.3.1 (бывший Босс-Референт). Система электронного документооборота построена на платформе IBM Domino и использует в качестве хранилища данных Notes Storage Facility (nsf-хранилища), которое имеет встроенные механизмы контроля доступа на основе групп Domino Directory. Каждый пользователь в системе идентифицирован и обладает собственным иерархическим notes-именем.

Доступ к документам, хранящимся в nsf-хранилище, может предоставляться как отдельным пользователям, ролям, так и группам пользователей за счет использования полей контроля доступа. Таким образом, в системе есть возможность ограничить доступ к каждому документу как на уровне хранилища целиком, так и на уровне отдельного документа. При смене типа хранилища система прав доступа будет унаследована.

Подсистема информационного поиска содержит настройки для поиска документов, созданных по конкретным моделям бизнес-процессов и формат вывода результатов поиска:

), (10)

где – параметры поиска документов, созданных по конкретному бизнес-процессу, – формат вывода результатов поиска.

Параметры поиска для каждого типа документа содержат ограничения на хранилища данных, заранее заданные критерии и пользовательские критерии:

), (11)

где – множество хранилищ данных, в которых выполняется поиск, – множество заранее заданных критериев поиска.

В отличие от Q_U, Q_{docType} задаются разработчиком системы заранее и недоступны для редактирования пользователю – обычно это ограничения на тип документа, используемый бизнес-процесс.

Реализация модели (11) в СЭД представлена на рис. 3.

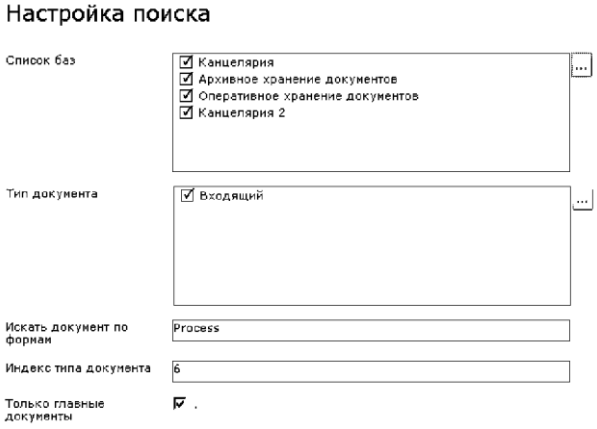

Рис. 3. Настройка поиска для конкретного бизнес-процесса

Поисковый запрос, формируемый пользователем, состоит из структурированной и неструктурированной частей. В качестве неструктурированной части пользователю предлагается ввести ключевое слово для поиска, в качестве структурированной предлагается выбор из заранее заданных критериев поиска для конкретного типа документа.

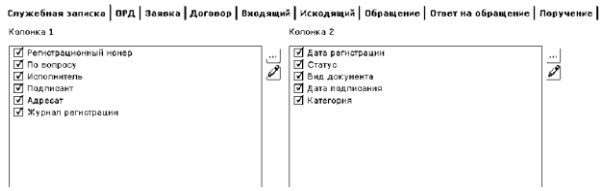

Рис. 4. Критерии поиска, доступные для конкретного типа документа

На основе доступных полей формируется интерфейс пользователя подсистемы информационного поиска.

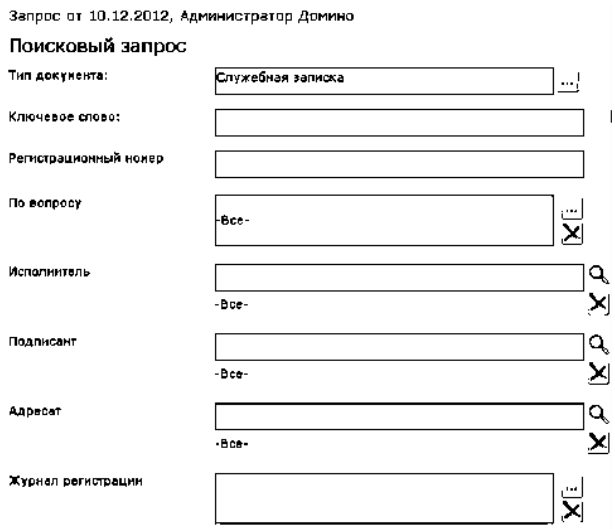

Рис. 5. Интерфейс пользователя подсистемы поиска

Поисковый запрос, выполняемый подсистемой поиска (9), в данном случае:

(11)

В качестве критерия *U* выступает иерархическое notes-имя пользователя, от лица которого формируется поисковый запрос. Ограничение (8) реализуется хранилищем данных в автоматическом режиме: документы, в полях контроля доступа которых нет упоминания пользователя *U*, не попадут в результаты поиска.

Мнемоническая схема процесса поиска представлена на рис. 6.

На основе полученной коллекции документов пользователь может принять решение о повторном поиске или продолжении работы с найденными документами.

ЗАКЛЮЧЕНИЕ

Использование информационно-поисковых систем позволяет осуществлять оперативный поиск требуемой информации в больших массивах разнородных данных, хранящихся на различных носителях и устройствах.

Поиск в корпоративной системе в значительной степени отличается от поиска в глобальной сети. Особое внимание при поиске в сети организации уделяется разграничению доступа пользователей к документам. Использование данных системных моделей позволяет уточнить поисковый запрос таким образом, чтобы в выборку попали только документы, доступные пользователю.

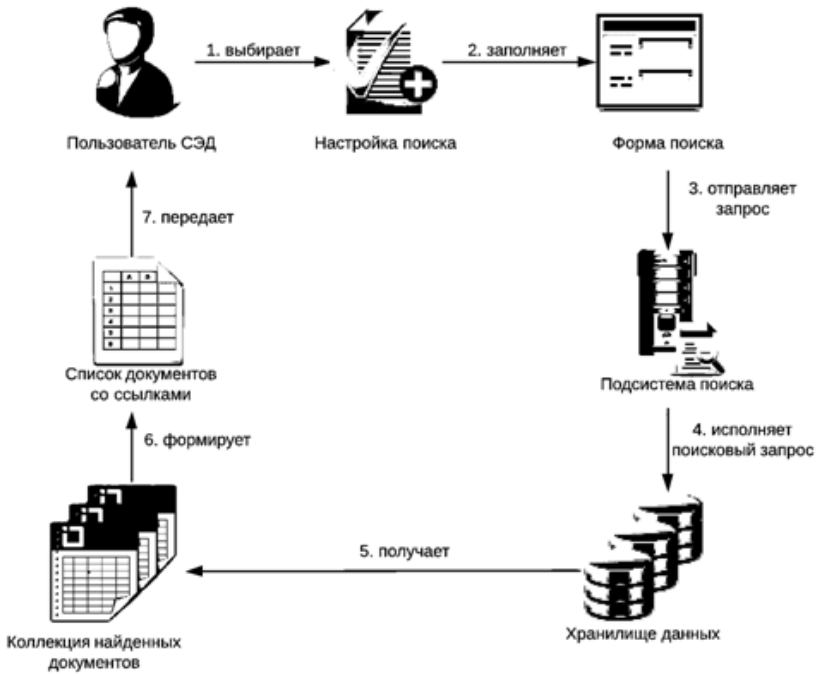

Рис. 6. Мнемоническая схема процесса поиска

Также в статье рассмотрена реализация предложенной модели в системе Логика ECM.СЭД. В поисковом модуле имеются настройки для одиннадцати типов документов, для которых определено 96 критериев поиска и 40 полей отображения. Среднее время поиска в коллекции из 3000 документов составляет 4 с.

СПИСОК ЛИТЕРАТУРЫ

1. **Шерстнев В. С., Иванов С. С., Акулин И. А.** Использование Oracle Universal Content Management в качестве корпоративного хранилища документов ТПУ // Вестник науки Сибири. Томский политехнический университет, 2011. № 1 (1). С. 302–307. **[** V. S. Sherstnev, S. S. Ivanov, and I. A. Akulin, “Oracle Universal Content Management as enterprise document storage TPU,” *Siberian science journal*, no. 1 (1) pp. 302-307, 2011. **]**

2. **Корпоративный** поиск: технологии Google на службе вашей компании // Каталог программных решений Softline direct, февраль 2013-2(132)-RU. **[“**Enterprise information retrieval: working with Google Search Appliance,” in *Softlie direct*, February 2013-2(132)-RU. **]**

3. **Медведев Н. Д., Гришин Г. А.** Модели управления доступом в распределенных информационных системах // Наука и образование: электрон. науч.-техн. изд-е. №1 январь 2011. С. 1 **[** N. D. Medvedev and G. A. Grishina, “Access control in distributed information systems models,” *Science and education*, no.1, January 2011. **]**

4. **Куликов Г. Г., Старцев Г. В., Бармин А. А.** Подход к построению информационно-поисковых систем для систем электронного документооборота // Актуальные проблемы в науке и технике. Т. 1. Информационные и инфокоммуникационные технологии: сб. науч. тр. 8-й Всерос. зимн. шк.-сем. аспирантов и молодых ученых (Уфа, 14-16 февраля 2013). Уфа: УГАТУ, 2013. С. 184–187. **[** G. G Kylikov, G. V. Startsev, and A. A. Barmin, “Approach of development of information retrieval system for enterprise content management systems,” in *Actual problems of science and technic*. *Vol. 1: Information and network technologies*: Proc. 8th winter workshop of postgrade students and young scientists, Ufa State Aviation Technical University. Ufa: USATU, 2013. **]**

ОБ АВТОРАХ

БАРМИН Александр Александрович, асп. каф. АСУ. Дипл. информатик-экономист (УГАТУ, 2011). Готовит дис. в обл. инф.-поисков. систем.

СТАРЦЕВ Геннадий Владимирович, доц. той же каф. Дипл. м-р техн. и технол. по инф.-упр. системам (УГАТУ, 2003). Канд. техн. наук (УГАТУ, 2006). Иссл. в обл. проектир. инф. систем на основе веб-технологий.

БАБАК Сергей Федерович, доц. каф. АСУ. дипл. инженер (УАИ, 1970), канд. техн. наук (УАИ, 1978).

КУЛИКОВ Григорий Геннадьевич, нач. отд. АСУ на ОАО УНПП «Молния».

METADATA

Title: Integration of automated information systems and information retrieval systems based on system models.

Authors: G.G. Kylikov[1], G.V. Startsev, A.A.[1], Barmin[1], S.V. Babak

Affiliation: Ufa State Aviation Technical University (UGATU), Russia.

Email: [1]gennadyg_98@yahoo.com, [2]startsev@gmail.com, [3]abarmin@outlook.com.

Language: Russian.

Source: Vestnik UGATU (scientific journal of Ufa State Aviation Technical University), vol. 18, no. 2 (63), pp. 85-92, 2014. ISSN 2225-2789 (Online), ISSN 1992-6502 (Print).

Abstract: Article describes problems of development and integration information retrieval systems and enterprise information systems. There are model of informational query and results, based on access lists and private preferences described in article. Model illustrated on an example of search subsystem of enterprise content management system based on IBM Lotus Domino platform.

Key words: business-process; domain model; information retrieval; role-based access model.

About authors:

BARMIN, Aleksandr Aleksandrovich Postgrad. (PhD) Student, Dept. of Automated Systems. Informatics and economist (UGATU, 2011).

STARTSEV, Gennady Vladimirovich, Dept. of Automated Systems. Master of technic and technology (USATU, 2003). PhD (USATU, 2006)

BABAK, Sergey Fedorovich, Docent, Dept. of Automated Systems. Dipl. Engineer, PhD.

KULIKOV, Grivory Gennadievich, Head of Automated Systems Department.