Куликов Г.Г.[1], **Старцев Г.В.**[2], **Бармин А.А.**[3], **Бармина О.В.**[4] ©
[1]Доктор технических наук, профессор, заведующий кафедрой автоматизированных систем управления;
[2]доцент, кандидат технических наук; [3]аспирант; [4]магистрант,
[2,3,4]кафедра автоматизированных систем управления,
Уфимский государственный авиационный технический университет

МНОГОАСПЕКТНЫЙ МЕТОД СЕМАНТИЧЕСКОГО ПОИСКА В СЛАБОСТРУКТУРИРОВАННОМ КОНТЕКСТЕ ИНФОРМАЦИОННОГО ПРОСТРАНСТВА

Аннотация

В статье рассмотрены проблемы построения поисковых запросов к информационно-поисковым системам с использованием структурных моделей бизнес-процессов в качестве критериев поиска. Увеличение объема информации в корпоративных информационных системах не позволяет использовать простые поисковые запросы для оперативного поиска необходимой информации и документов, созданных в процессе выполнения бизнес-процесса. Предложен вариант построения поискового запроса с позиции анализа моделей бизнес-процесса, для которого выполняется поиск связанной информации.

Ключевые слова: бизнес-процесс, предметная область, информационный поиск.
Keywords: bussines process, semistructured subject area, information retrieval.

ВВЕДЕНИЕ

Современная цивилизация находится в стадии формирования информационного общества — особого общества, основными характеристиками которого являются:

• наличие информационной инфраструктуры, состоящей из большого числа разнородных информационных ресурсов как хранилищ данных и знаний;

• массовое применение персональных компьютеров, подключенных к высокоскоростным сетям передачи данных.

Информационная система организации в процессе своего функционирования накапливает значительный объем данных, так что встает вопрос оперативного поиска информации. Многие информационные 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 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 представлением этих данных пользователю.

Из-за сложности поиска в массивах неструктурированной информации широкое распространение получают информационно-поисковые системы, но проблема автоматизированного поиска документов, связанных с бизнес-процессами остается все еще недостаточно проработанным.

Статья посвящена решению вопросов, связанных с концепцией автоматизированного формирования поисковых запросов на основании структурных моделей с целью сокращения времени поиска и ошибок оператора.

1. СОСТОЯНИЕ ВОПРОСА

В настоящее время класс корпоративных систем информационного поиска представлен следующими наиболее распространенными в России продуктами[Ошибка: источник перёкрестной

ссылки не найден]:

1. Системы общего назначения:
• Custom Google Search
• Яндекс.Поиск
2. Специализированные системы:
• Sphinx
• Apache Lucene
• Apache Solr

Custom Google Search

В настоящее время существует два варианта Custom Google Search – пользовательский поиск для сайта и система поиска Site Search для организаций.

Система пользовательского поиска для сайта позволяет выполнять информационный поиск по документам, расположенным на одном или нескольких сайтах пользователя. Система не устанавливается как самостоятельное приложение, а предлагается как сервис – пользователь устанавливает код на свой сайт и получает возможность осуществлять поиск с использованием механизмов Google исключительно на своем сайте[Ошибка: источник перёкрестной ссылки не найден].

В качестве системы корпоративного поиска Google предлагает продукт Google Search Appliance – универсальное поисковое решение для поиска по внутренним документа, хранящихся в виде файлов, записей в базах данных информационных систем и документам, находящимся в глобальной сети[Ошибка: источник перёкрестной ссылки не найден].

Также Google предлагает решение для персонального информационного поиска документов на компьютере пользователя – Google Desktop. В настоящее время проект закрыт, выпуск новых версий приостановлен. В качестве поискового пространства могут выступать файлы документов и электронной почты, хранящиеся на компьютере пользователя.

Яндекс.Персональный поиск и Яндекс.Сервер

Яндекс также предлагает поисковые решения для персонального и корпоративного использования.

Персональный поиск Яндекса – это самостоятельное прикладное программное обеспечение, которое предназначено для полнотекстового поиска по файлам, находящимся на локальном компьютере. [Ошибка: источник перёкрестной ссылки не найден]

Яндекс.Сервер – это продукт для корпоративного поиска, который обеспечивает полнотекстовый информационный поиск с учетом морфологии русского языка на локальном веб-сервере или в корпоративной сети. Поддерживается поиск только по файлам документов и реляционным базам данных[Ошибка: источник перёкрестной ссылки не найден].

Sphinx

Sphinx – система информационного поиска по реляционным базам данных и нереляционным хранилищам, имеет программный интерфейс для интеграции с существующими приложениями. Также поддерживается создание кластеров поисковых серверов для повышения масштабируемости и отказоустойчивости, создание дополнительных пользовательских полей, морфологии и синонимии[Ошибка: источник перёкрестной ссылки не найден].

Apache Lucene

Apach Lucene – это библиотека для реализации поискового механизма, которая может быть встроена в существующие системы. В исходном варианте поддерживается только поиск по текстовым документам, но функциональность может быть расширена за счет дополнительных модулей. Также имеется интерфейс для работы с поисковым индексом, созданным Apache Lucene из других приложений[Ошибка: источник перёкрестной ссылки не найден].

Apache Solr

Apache Solr – это решение на базе Apache Lucene, которое представляет собой систему корпоративного поиска, реализованного в виде веб-сервиса. Apache Solr поддерживает полнотекстовый, фасетный поиск, динамическую кластеризацию и интеграцию с реляционными базами данных[Ошибка: источник перёкрестной ссылки не найден].

2. ПРОБЛЕМА СЕМАНТИЧЕСКОГО ПОИСКА В СЛАБОСТРУКТУРИРОВАННОМ КОНТЕКСТЕ ИНФОРМАЦИОННОГО ПРОСТРАНСТВА

Информационный поиск заключается в удовлетворения информационной потребности пользователя путем формирования и исполнения пользователем информационного запроса к поисковой системе. В результате исполнения информационного запроса пользователь получает результаты, релевантные его запросу.

Обычно уточнение поискового запроса производится пользователем после получения какого-либо результата и выполняется до тех пор, пока информационная потребность не будет удовлетворена. В рамках современного информационного общества предъявляются жесткие требования к скорости получения и обработки информации, так что подход к поиску на основе корректировки поискового запроса по результатам поиска может не дать удовлетворительного результата за короткое время.

Информационный поиск, выполняемый сотрудником в организации, чаще всего затрагивает документы, создаваемые и изменяемые сотрудником в рамках бизнес-процессов, с которыми работает сотрудник. Кроме того, найденные документы могут быть использованы в качестве оснований для создания новых документов и потоков исполнения и рассмотрения. Таким образом, пользователь, работающий с системой корпоративного поиска должен иметь возможность продолжить работу с найденными документами.

Корпоративная информационная система представляет собой совокупность подсистем, выполняющих автоматизацию определенных функций. Эти системы могут быть созданы как одной, так и несколькими компаниями-разработчиками, использовать в качестве хранилищ данных различные СУБД и форматы. Корпоративная информационно-поисковая система должна обеспечивать возможность поиска в разнородной среде и применять общие механизмы ранжирования для документов в различных представлениях. Также в результаты поиска должны попадать документы в соответствии с правами доступа пользователя, от имени которого выполняется поиск.

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

Представление документов в виде структур различных типов накладывает дополнительные ограничения — перед поиском элементы информационного пространства должны быть проиндексированы — создание поискового индекса значительно увеличивает скорость поиска. В случае представления документа в виде совокупности именованных атрибутов поисковый индекс с большей адекватностью отражает содержание документов, чем в случае представления документа в виде совокупности лексических единиц — в этом случае используется полнотекстовый индекс, использование которого снижает релевантностью поиска.

Документы, создаваемые в рамках определенной функции бизнес-процесса, могут инициировать создание или исполнение документов в других функциях этого же или другого процесса. Таким образом, необходимо выполнять поиск не самостоятельного документа или группы документов одного класса, а совокупности связанных документов, относящихся к одному бизнес-процессу.

С информационно-поисковой системой чаще всего работает пользователь, который не является экспертом в области информационного поиска — задачи поиска информации могут быть не основными в рамках его должностных обязанностей. Тем не менее, любой пользователь информационной системы должен решать задачи информационного поиска эффективно, что требует обеспечения пользователя механизмом, упрощающим формирование поисковых запросов. Пользователю может быть представлен интерфейс для ввода запроса на естественном языке, так и возможность формирования запроса по заранее заданным критериям в зависимости от бизнес-процесса, в рамках которого выполняется поиск.

Как видим, задача информационного поиска является сложной в виду большого количества

накладываемых ограничений, которые носят как технический, так и эргономический характер.

Задача автоматизированного формирования поискового запроса заключается в предоставлении пользователю возможности выбора заранее заданных бизнес-процессов и функций в них, для которых существуют элементы поискового запроса. Пользователь, выбирая нужные функции в интерактивном режиме, формирует строку запроса по заранее заданным критериям, что снижает возможность ошибки и изолирует пользователя от самой поисковой системы.

В рамках автоматизации бизнес-процессов кафедры АСУ, среди прочих, встала задача обеспечения оперативного поиска документов в слабоструктурированном контексте информационного пространства кафедры. В качестве критериев выбора информационно-поисковой системы были выдвинуты следующие:

1. Бесплатность для использования в образовательном учреждении.
2. Доступность документации на русском языке.
3. Простота администрирования.
4. Возможность интеграции с существующими сервисами портала.

В результате анализа приведенных выше альтернатив, в качестве поисковой системы для кафедры был выбран Яндекс.Сервер. В результате внедрения информационно-поисковой системы было проиндексировано хранилище документов, в котором находится 83817 документов общим объемом 64 гигабайта. Размер поискового индекса составил 202 мегабайта, среднее время исполнения запроса 63 миллисекунды.

3. ПРИМЕНЕНИЕ ПРИНЦИПОВ СИСТЕМНОГО ПОДХОДА ДЛЯ ФОРМИРОВАНИЯ ПОИСКОВЫХ ЗАПРОСОВ

Применение принципов системного подхода к процессу поиска позволяет сократить время поиска и уменьшить количество поисковых запросов. Информация, получаемая из структурных моделей бизнес-процесса, может быть использована в качестве критериев поиска.

Применение методологии SADT в процессе моделирования бизнес-процессов дает исходные данные для поиска документов, относящихся к моделируемому бизнес-процессу — информация, полученная из моделей, может быть использована в качестве критериев поиска.

Комплект структурных моделей в соответствии с методологией SADT включает в себя функциональную модель бизнес-процесса, информационную модель, описывающую структуру информации, обрабатываемой в бизнес-процессе и динамическую модель, описывающую последовательность работы и альтернативные варианты работы с информацией.

Данные этих моделей выступают в качестве критериев поиска на различных его этапах.

Функциональная модель отображает функциональную структуру бизнес-процесса в виде системы, то есть совокупности производимых действий и связей между этими действиями. Функциональная модель в контексте бизнес-процесса позволяет получить состав входных и выходных документов, точку зрения и цель выполнения бизнес-процесса. Цель и точка зрения позволяют сократить область поиска за счет выделения документов, относящихся к одной функции или действующему лицу.

Использование информации из функциональной модели позволяет выбрать конкретный класс документов, относящихся к какой-либо конкретной функции. Дальнейшее уточнение поиска идет за счет использования критериев из информационной модели. Информационная модель представляет документы предметной области в виде сущностей и связей. Сущности определяют состав атрибутов рассматриваемых документов, отношения между сущностями — отношения между документами в предметной области. Атрибуты и их значения выступают в качестве критериев поиска следующего уровня, то есть для поиска конкретных документов в выделенном ранее классе документов.

Включение в состав поискового запроса значений ключевых атрибутов сущностей позволяет выделить из области определения поиска конкретные документы, использование неключевых атрибутов и конкретных значений внешних ключей позволяет выделить документы, связанные с конкретным значением внешнего ключа.

Для поиска документов в информационном пространстве по значениям атрибутов необходимо определить значения атрибутов в документе. Документы в информационном пространстве могут быть представлены как в виде записей в базе данных (как реляционной, так и объектно-ориентированной), так

и в виде документов офисных приложений, например, документов Microsoft Word. В последнем случае поиск по атрибутам документа не может быть выполнен в явном виде. Для решения выявленной проблемы документ может быть структурирован и разделен на термины с использованием шаблона и механизма разбора, либо возможно выполнение полнотекстового поиска по полнотекстовому индексу.

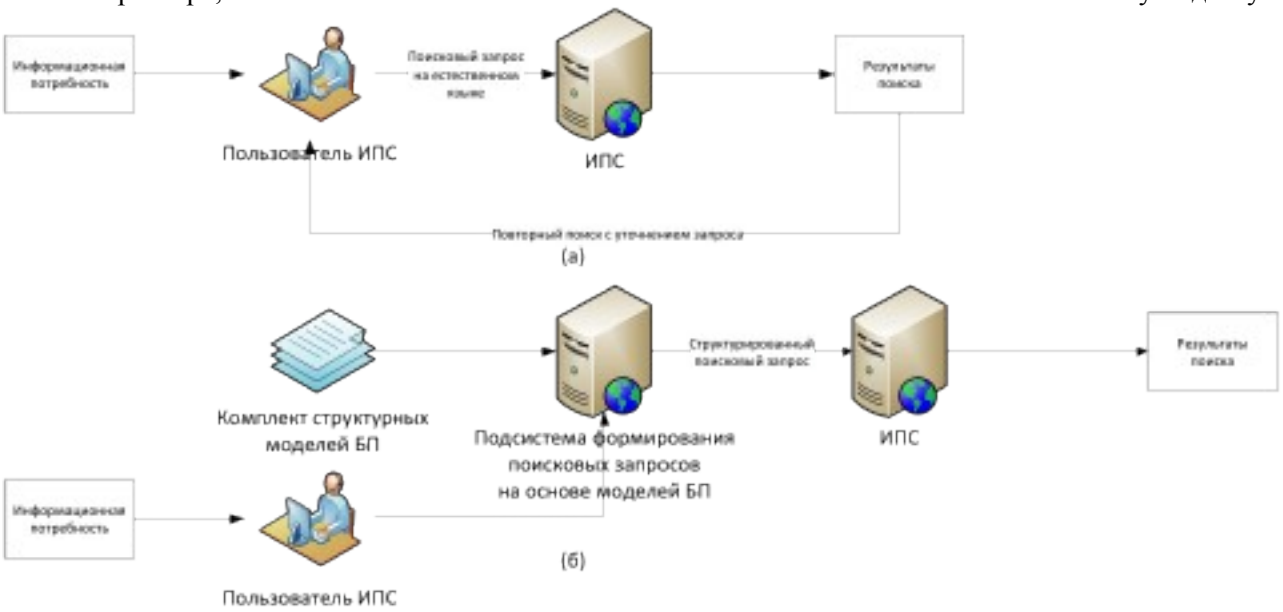

Рис. 1. **Существующий (а) и предлагаемый (б) подход к формированию поискового запроса**

Использование структурированных документов позволяет строить поисковые запросы с применением модели булева поиска — поисковый запрос представляет собой совокупность терминов и их значений, а также булевых операций.

Использование полнотекстового поиска значительно сокращает точность поиска, так как термины запроса могут встречаться во многих местах документа.

В процессе создания результирующего документа бизнес-процесса создается также и группа дополнительных, вспомогательных документов, которые являются промежуточными по отношению к целевому документу. Конкретный путь создания документа с учетом альтернатив определяется динамической моделью. Использование информации из динамической модели позволяет выделить совокупности документов, связанные причинно-следственными связями.

Использование многоуровневых критериев поиска позволяет говорить о системе критериев семантического поиска. Каждый следующий уровень поиска позволяет сужать область поиска, что приведет к удовлетворению информационной потребности пользователя быстрее, чем использование совокупности простых запросов.

Рассмотрим пример поиска документов в слабоструктурированном контексте информационного пространства кафедры АСУ с использованием информационно-поисковой системы Яндекс.Сервер (адрес портала АСУ http://asu.ugatu.ac.ru/).

Область поиска представляет собой неструктурированную совокупность документов в виде файлов в формате Microsoft Office, в процессе индексации которых создан полнотекстовый индекс.

Поиск по полнотекстовому индексу с использованием модели ранжированного поиска позволяет строить поисковые запросы в упрощенном виде, но необходимо включать не только искомые ключевые слова, но и слова из шаблона документа, находящиеся в непосредственной близости ключевых — использование словосочетаний повышает точность поиска. Зачастую документы различаются незначительно, так что для выделения из области поиска конкретного типа документов необходимо использовать большое число параметров. Чтобы исключить возможность ошибки оператора, необходим алгоритм составления поискового запроса, который бы позволил в диалоговом режиме формировать запрос и отправлять его на исполнение поисковой системе.

В рамках автоматизации внутренних бизнес-процессов на кафедре автоматизированных систем управления был создан корпоративный портал, который является рабочим местом сотрудника кафедры и предоставляет централизованный доступ ко всем информационным ресурсам кафедры. Интеграция информационно-поисковой системы с порталом кафедры позволит обеспечить единую точку доступа и интеграцию с существующими сервисами.

ЗАКЛЮЧЕНИЕ

Использование информационно-поисковых систем позволяет осуществлять оперативный поиск требуемой информации в больших массивах разнородных данных, хранящихся как в виде файлов в файловой системе, так и в виде записей в базах данных информационных систем.

Среди рассмотренных приложений для построения систем информационного поиска можно выделить как готовые решения, которые содержат в себе необходимые компоненты для организации поиска и поддержания индекса в актуальном состоянии, так и библиотеки, которые реализуют поисковые механизмы и позволяют построить собственные поисковые системы с требуемыми параметрами.

Также был предложена методология формирования поисковых запросов на основе информации из структурных моделей бизнес-процессов, что позволяет построить многоуровневую систему критериев поиска и выполнять поиск необходимых документов за меньшее число итерации поиска.

Литература

1. Маннинг К.Д., Рагхаван П., Шютце Х. Введение в информационный поиск.: Пер. с англ. — М.: ООО «И.Д. Вильямс», 2011.- 528 с.: ил.- Парал. тит. англ.
2. Apache Lucene Core [Электрон. ресурс] // Apache Lucene.- Режим доступа: http://lucene.apache.org/java/docs/index.html
3. Apache Solr [Электрон. ресурс] // Apache Solr.- Режим доступа: http://lucene.apache.org/solr/
4. Google – система пользовательского поиска [Электрон. ресурс] // Google.- Режим доступа: http://www.google.ru/cse/
5. Google – универсальный поиск для организаций [Электрон. ресурс] // Google.- Режим доступа: http://google.softline.ru/gsa.php
6. Sphinx Open Source Search Server [Электрон. ресурс] // SphinxSearch.- Режим доступа: http://sphinxsearch.com/about/sphinx/
7. Персональный поиск Яндекса. [Электрон. ресурс] // Яндекс.- Режим доступа: http://desktop.yandex.ru/
8. Полнотекстовый поиск в веб-проектах [Электрон. ресурс] // Хабрахабр.- Режим доступа: http://habrahabr.ru/blogs/webdev/30594/
9. Яндекс.Сервер – порядок на Вашем веб-сервере. [Электрон. ресурс] // Яндекс.- Режим доступа: http://company.yandex.ru/technologies/server/