Информатика и вычислительная техника

УДК 004.657 **DOI: 10.14529/ctcr180101**

МЕТОД СТРУКТУРИРОВАНИЯ КОНТЕНТА ГЕТЕРОГЕННОГО ИНФОРМАЦИОННОГО ПРОСТРАНСТВА НА ОСНОВЕ ФОРМАЛИЗОВАННОЙ МОДЕЛИ ПРЕДМЕТНОЙ ОБЛАСТИ ДЛЯ РЕШЕНИЯ ЗАДАЧ ИНТЕЛЛЕКТУАЛЬНОГО ПОИСКА

Г.Г. Куликов, М.А. Шилина, А.А. Бармин, Г.В. Старцев, Д.Г. Шамиданов

Уфимский государственный авиационный технический университет, г. Уфа, Россия

Рассмотрены проблемы построения и интеграции информационно-поисковых систем с корпоративными информационными системами. Описывается модель информационного запроса и результатов поиска по базе данных и файловым хранилищам данных. Задача информационного поиска представляется как взаимодействие доменов с использованием адаптера. Задача адаптера заключается в преобразовании информационной потребности, выраженной на языке конкретного документа в поисковый запрос, представленный на языке информационно-поисковой системы. Представление поисковых запросов на языке поисковой системы предполагает следующий порядок операций: определение целевой сущности, преобразование критериев, заданных пользователем, и добавление критериев, заданных исходной системой. Предлагается алгоритм порядка преобразования в общем виде.

Рассматривается реализация предложенной модели в PHP-фреймворке веб-портала на базе поисковой системы Apache Solr.

Ключевые слова: полнотекстовый поиск, гетерогенное информационное пространство, информационно-поисковая система, корпоративное приложение, хранилища данных, модель предметной области, интеграция программных систем.

Введение

Информационные системы необходимы для поддержки принятия управленческих решений. Корпоративная информационная система не в полной мере адекватна информационной системе организации. Это связано как со сложностью формализации бизнес-процессов, так и с возрастающей сложностью используемого программного обеспечения. Вследствие этого создаваемая корпоративная ИС удовлетворяет только ограниченный круг информационных потребностей пользователей.

В процессе управления могут возникать вопросы, которые не были предусмотрены при проектировании программного обеспечения. Для того чтобы получить ответ на этот вопрос, нужно обратиться к хранилищу данных с новым запросом.

Структура хранилища данных накладывает ограничения на сферу применения системы, так как заранее сформированные таблицы, OLAP-кубы и витрины данных позволяют получать ответы только на заранее заданные вопросы пользователей.

Оперативно получать данные по требованию на основе произвольных запросов пользователей позволяют корпоративные информационно-поисковые системы. Эти системы поддерживают актуальный, периодически обновляющийся индекс объектов информационного пространства и предоставляют пользователю интерфейс для создания запросов на получение требуемых объектов [1].

Однако в силу гетерогенности информационного пространства большинства предприятий и организаций создание системы корпоративного поиска становится нетривиальной задачей, требующей применения научного подхода. Для решения задачи интеграции системы корпоративного поиска и корпоративных приложений требуется формализация представления предметной области, т. е. модель гетерогенного информационного пространства, а также формирование алго-

ритма взаимодействия корпоративных приложений с поисковой системой. Отдельные аспекты этой проблемы рассмотрены в работах [2–5].

Сравнительный анализ подходов к интеграции систем корпоративного поиска и корпоративных приложений

К решению задачи информационного поиска в гетерогенном информационном пространстве 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).

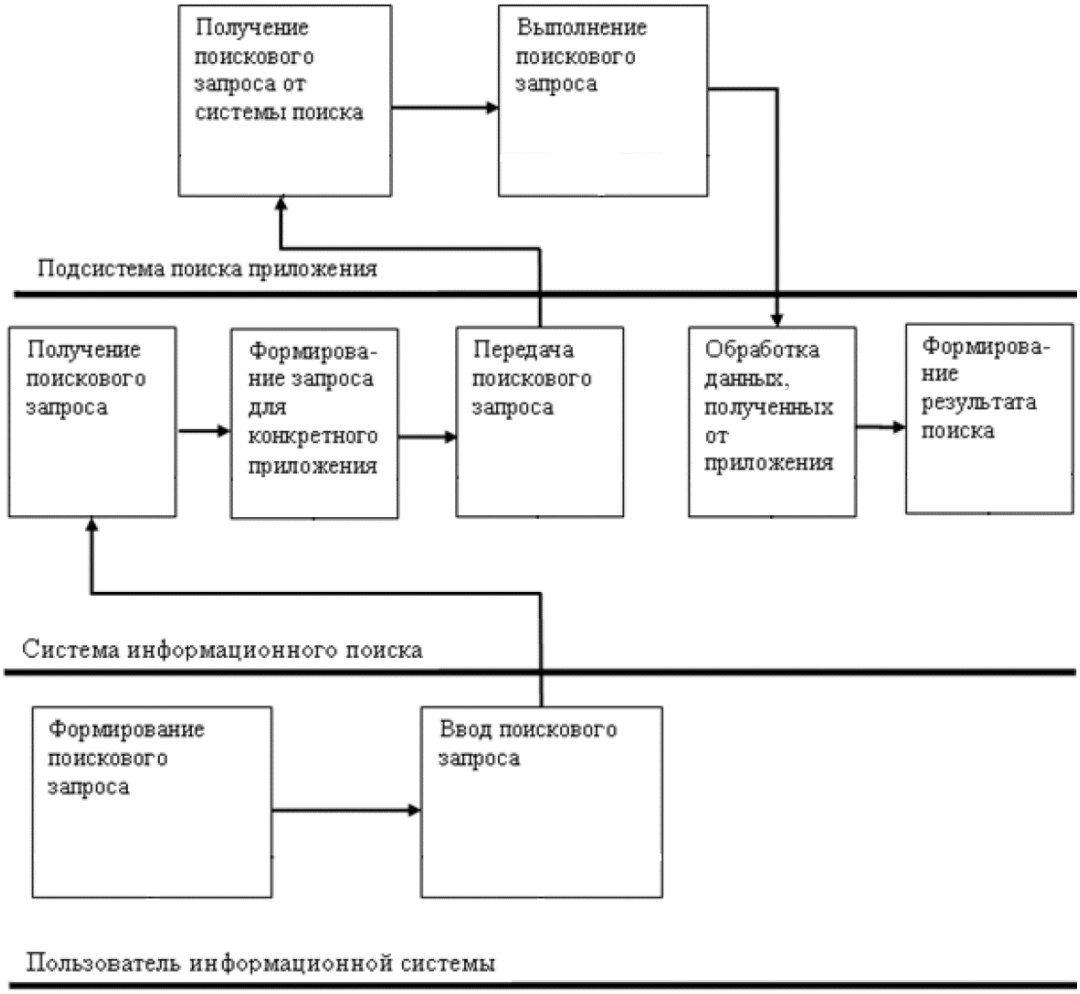

Рис. 1. Использование встроенной подсистемы поиска корпоративного приложения

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 [2].

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).

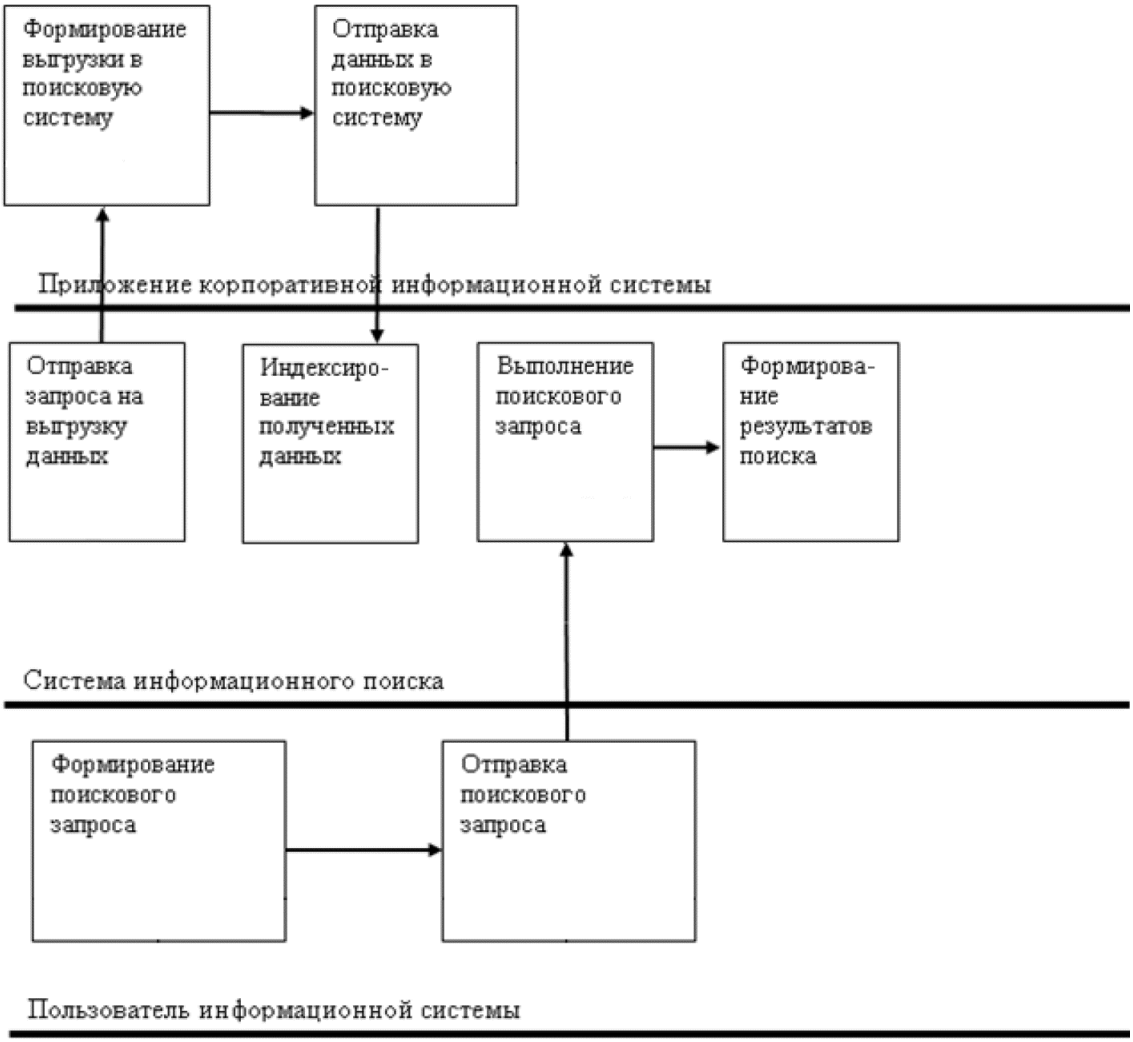

Рис. 2. Подход с использованием централизованной корпоративной информационно-поисковой системы

Достоинствами данного подхода являются:

- централизация функций поисковой системы;
- единая информационно-поисковая система позволяет выполнять ранжирование документов по одинаковым критериям для всех подсистем-источников данных;
- отсутствие необходимости модифицировать существующие системы;
- 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 [3].

Также есть и третий подход к интеграции выделенной информационно-поисковой системы и частей автоматизированной информационной системы (рис. 3).

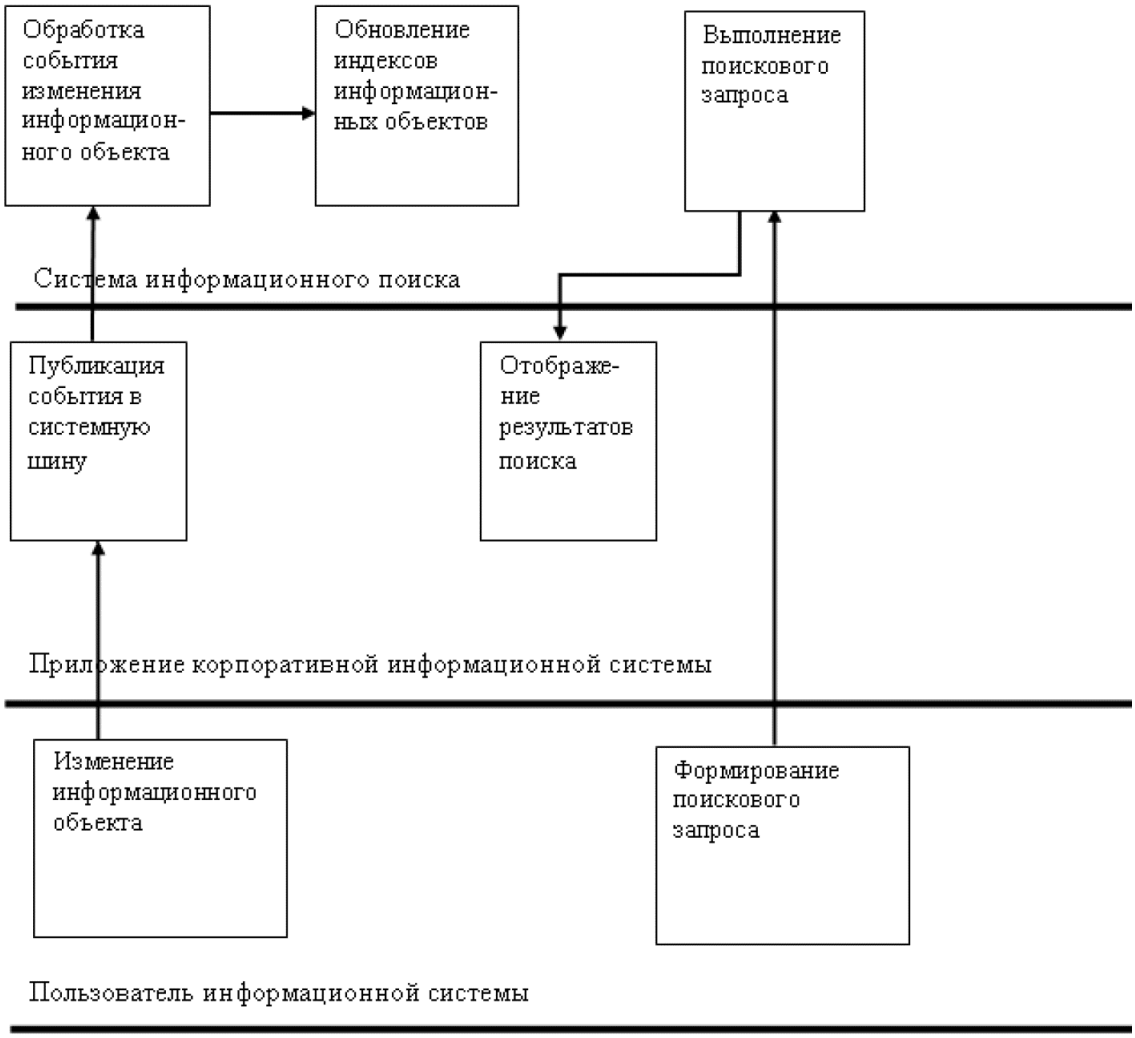

Рис. 3. Интеграция информационно-поисковой системы и модуля КИС с использованием системной шины

Подход основывается на использовании системной шины или брокеров сообщений. При данном подходе каждая из частей корпоративной информационной системы публикует в системную шину события создания, удаления или изменения информационных объектов, а информационно-поисковая система по мере их поступления обновляет поисковый индекс.

Достоинством данного подхода по сравнению с предыдущим является более оперативное обновление индекса информационных объектов, а также возможность более гибкого расширения за счет наличия единой точки входа – системной шины.

В то же время, основным недостатком данного подхода является использование вспомогательной промежуточной системы передачи сообщений и необходимость модификации существующих систем с целью добавления им функции публикации событий.

Алгоритм формирования поисковых запросов в гетерогенном информационном пространстве

Введем понятие домена (domain) с точки зрения предметно-ориентированного проектирования. Под доменом понимается область деятельности, в рамках которой организация выполняет свою функцию. Доменом может выступать разработка программного обеспечения, машиностроение, образовательная и другие области деятельности. Домен определяет операции, которые могут выполняться в рамках обозначенной области деятельности в терминах рассматриваемой предметной области, например, для образовательного процесса это могут быть операции «прием документов», «подготовка к проведению занятия», «проведение зачета» и другие.

С точки зрения системного подхода каждый домен может быть разделен на совокупность поддоменов (subdomains), которые отражают только определенный аспект деятельности организации и ограничивают набор выполняемых в них операций. С точки зрения организационной структуры поддомены не обязательно представляют собой структурные подразделения, они также могут быть представлены сквозными бизнес-процессами, охватывающими несколько структурных подразделений. Также внутри домена существует собственный непротиворечивый язык (ubiquitous language), который определяет термины и связи между ними.

Также введем понятие ограниченного контекста (bounded context) как границу, которая определяет область применения доменного языка. Ограниченный контекст может включать в себя несколько доменов и поддоменов. Стоит отметить, что одни и те же понятия в разных ограниченных контекстах могут иметь различную семантику.

Для взаимодействия между несколькими ограниченными контекстами вводится понятие адаптера (adapter), который устанавливает правила преобразования и отношения между различными сущностями ограниченных контекстов.

Домены, ограниченные контексты и адаптеры могут также быть представлены в форме программного обеспечения. Тогда элементы доменов будут некоторыми сущностями в хранилища данных некоторой автоматизированной системы.

Представим механизм информационного поиска как взаимодействие нескольких доменов или систем предметной области с доменом информационно-поисковой системы путем передачи сообщений (рис. 4).

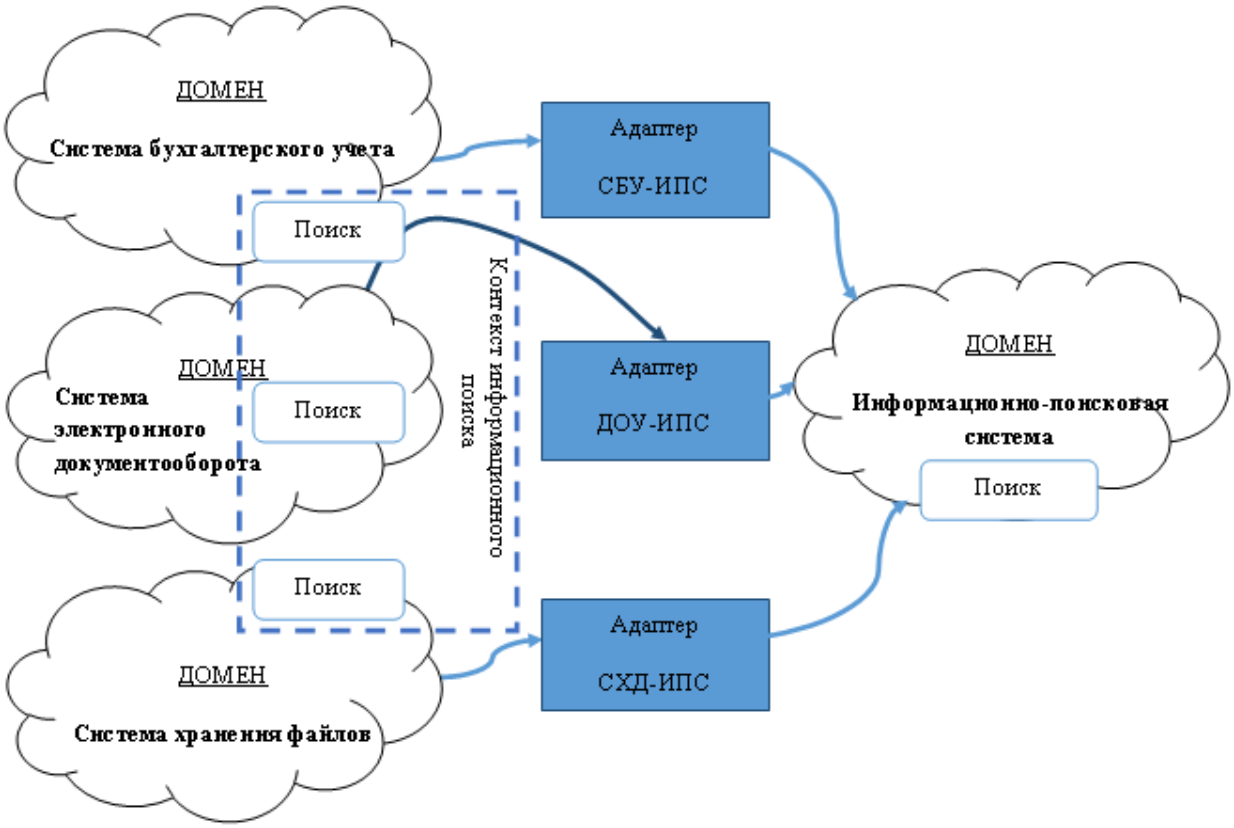

Рис. 4. Пример объединения доменов информационных подсистем предприятия (организации)

На рис. 4 представлено три домена, которые иллюстрируют три функции информационной системы – бухгалтерский учет, документационное обеспечение управления и хранение файлов. Также определен отдельный домен информационно-поисковой системы. В каждом из этих доменов определено действие «поиск», однако в каждом из доменов это действие имеет различную семантику.

Для того чтобы домены могли взаимодействовать между собой, необходимо преобразование сообщений из терминов одного домена в термины другого домена. Чтобы выполнять эти дейст-

вия, определим адаптеры между доменами. Задача адаптера в данном случае заключается в преобразовании информационной потребности, выраженной в терминах конкретного домена к терминам, используемым в домене информационного поиска, т. е. его можно представить следующей функцией:

$$\text{domain2} = \text{adapter}(\text{domain1}). \quad (1)$$

Для того чтобы иметь возможность представлять данные из домена поиска, в предметных доменах должно иметь место и обратное преобразование, т. е. функция adapter должна обладать свойством симметричности:

$$\text{domain1} = \text{adapter}(\text{domain2}). \quad (2)$$

Таким образом, задача функции adapter заключается в формировании поискового запроса для информационно-поисковой системы на основе запроса, выраженного в терминах конкретной предметной области.

Обычно функция adapter имплементируется в программном обеспечении с помощью одноименного паттерна и реализует логику, соответствующую следующему алгоритму (рис. 5).

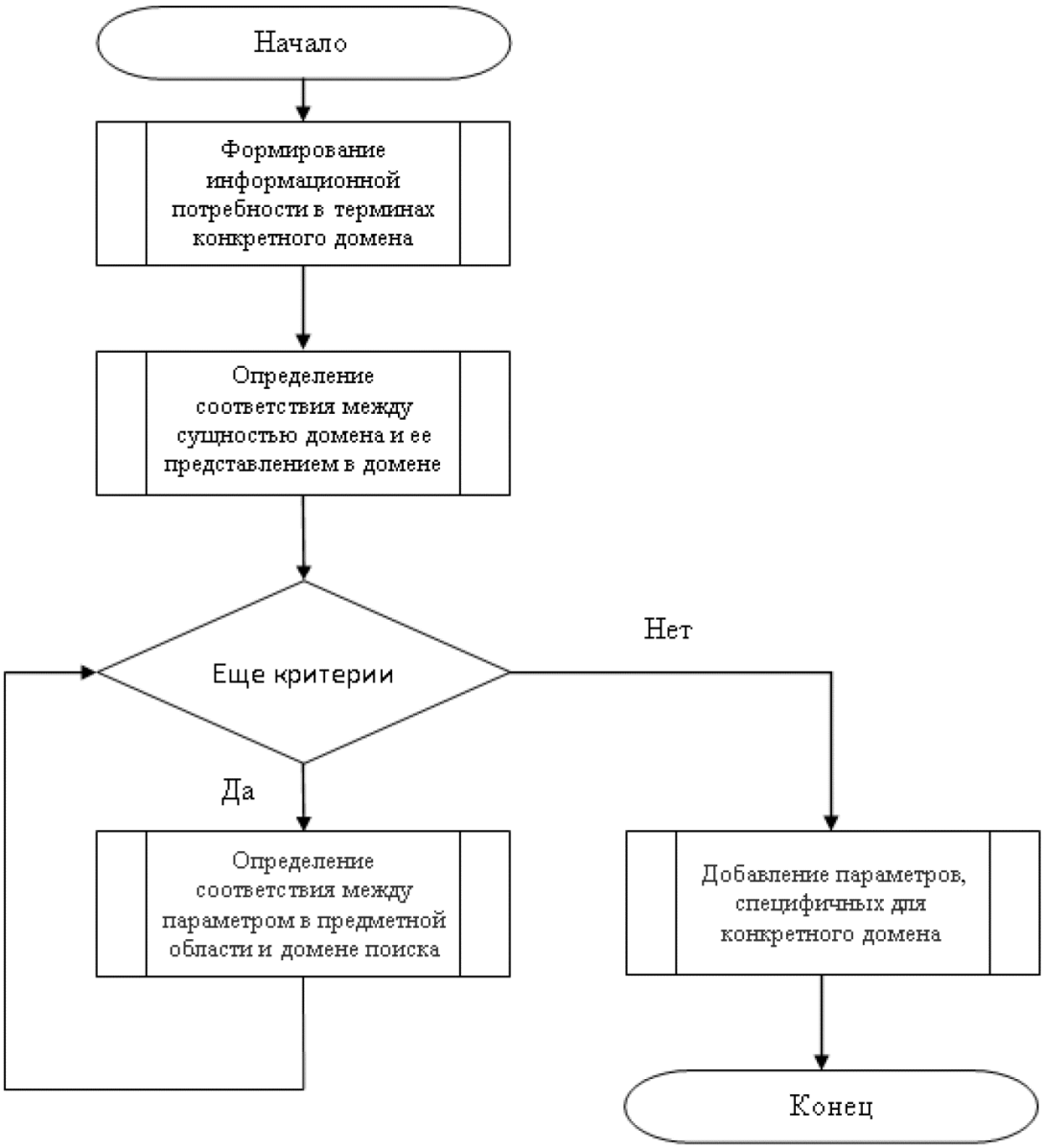

Рис. 5. Алгоритм преобразования информационной потребности из конкретного домена в поисковый запрос, выполняемый в конкретной информационно-поисковой системе

Установим следующие правила преобразования данных между доменами (см. таблицу).

Правила преобразования данных между доменами

Термин в домене ДОУ	Термин в домене ИПС
ИЩУ	SELECT
И	AND
Тип документа	Field «docType»
Ключевое слово	Field «docTitle»
Вид документа	Field «docKind»

Применяя описанный выше алгоритм, сформируем следующий запрос:

```
SELECT document
WHERE
   docType = “Поручение” AND
   docTitle = “О предоставлении доступа” AND
docKind = “Служебная записка”
```

Также проиллюстрируем описанный выше подход с использованием мнемосхемы, приведенной на рис. 6.

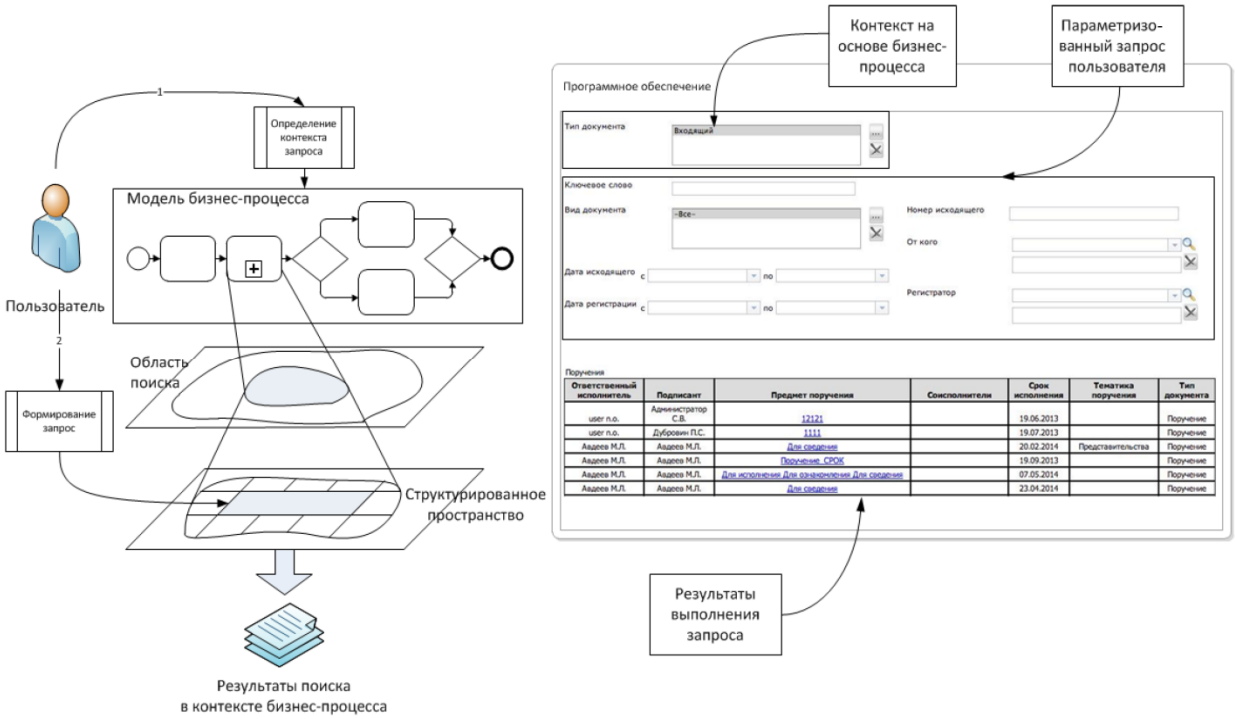

Рис. 6. Схема выполнения поискового запроса в структурированном информационном пространстве

Пример реализации алгоритма формирования поисковых запросов для структурирования гетерогенного информационного пространства крупной выпускающей кафедры

Рассмотрим реализацию модуля поиска для системы веб-портала, служащего интеграционной средой при построении информационного пространства кафедры вуза. В рамках единого интерфейса отображения и управления информацией веб-портал объединяет различные задачи по ведению документооборота учебного процесса.

Веб-портал организован на базе PHP-фреймворка «Фреймворк.АСУ». Ориентируясь на принципы объектно-ориентированного программирования, в «Фреймворк.АСУ» организованы классы, описывающие предметную область кафедры АСУ. Представление информационного

пространства кафедры в виде совокупности классов позволяет реализовывать и легко представлять новые сущности предметной области и устанавливать связи между ними. «Фреймворк.АСУ» использует в качестве хранилища данных базу данных MySQL и позволяет производить поиск по ней с помощью полнотекстовых SQL-запросов непосредственно в базе данных, проиндексировав данные с помощью MySQL Fulltext.

Сложность предметной области и объем обрабатываемых документов не позволяет большинству пользователей формировать поисковые запросы самостоятельно, поэтому был произведен анализ существующих решений для реализации полнотекстового поиска.

Поисковый модуль представляет собой набор заранее заготовленных критериев поиска, соответствующих полям документа, элементов управления и словарей, предоставляющих пользователю возможность выбора значения в интерактивном режиме.

Представим модуль поиска в виде теоретико-множественной модели.

$$S = \{S_{docType}, V_{docType}\}, \quad (3)$$

где S – модуль поиска веб-портала;

$S_{docType}$ − совокупность поисковых подсистем для каждого типа документов информационного пространства;

$V_{docType}$ − совокупность подсистем отображения результатов поиска для каждого типа документов.

Поисковая подсистема позволяет выполнять поиск документов определенных типов как в базе данных веб-портала, так и во внешних хранилищах. Разделение типов документов внутри веб-портала выполняется на уровне атрибутов, поэтому поисковая подсистема имеет заранее заданный набор критериев для выбора документов только указанных типов. Также поисковая подсистема позволяет пользователю указать критерии поиска, характерные для выбранного типа.

$$S_{docType} = \{Db, C_{docType}, C_{user}\}, \quad (4)$$

где Db – множество хранилищ данных информационного пространства;

$C_{docType}$ – заранее заданные критерии для выбора документов конкретного типа;

C_{user}– пользовательские критерии отбора документов.

$C_{docType} \in C_{user}$, но задаются администратором веб-портала и не доступны для редактирования пользователям. Пользователю доступен набор полей для каждого из типов документов, значения которых можно использовать в качестве критериев поиска.

С точки зрения программной реализации каждое поле представляет собой объект, который инкапсулирует часть запроса. При генерации поискового запроса отбираются непустые частичные запросы, из которых затем формируется общий запрос для параметризированного поиска.

В веб-портале для каждого типа документа создается документ настроек, в котором указывается, по каким хранилищам данных выполняется поиск. Также администратор задает форму и тип объекта хранилища данных для поиска. Эти настройки используются для каждого поискового запроса.

В модуле поиска имеется набор заранее заготовленных полей поиска для каждого типа документа. Эти поля настраиваются и хранятся в специальном файле schema.xml в системе Apache Solr. Используя эти поля, пользователь может формировать и сохранять собственные шаблоны поисковых запросов, подставляя значения, в которые формируются поисковые запросы к базе данных (рис. 7).

При настройке поиска по базе данных, пользователем указывается множество полей, по которым осуществляется поиск:

$$C_{user} = \{C_{field}\}, \quad (5)$$

где C_{field} – множество полей, которые пользователь использует в качестве критериев отбора документов.

$$C_{field} =< f, w, D >, \quad (6)$$

где f – поле документа, которое является критерием отбора документов;

w – элемент управления, используемый в веб-портале для представления пользователю возможности заполнить это поле;

D – доступные для выбора значения указанного поля.

Рис. 7. Настройка множества полей поискового индекса Apache Solr

Таким образом, пользователь использует простой интерфейс из доступных в веб-портале элементов управления, каждый из которых позволяет выбрать из значений, доступных в базе данных.

Реализация поискового модуля во фреймворке использует интерфейс ISearchSource, который позволяет задавать несколько разных источников – хранилищ данных, по которым может выполняться поиск. Например, это может быть сама база данных MySQL, а также файловые хранилища, такие как FTP или Samba-сервер.

В список необходимых настроек поискового модуля для поиска файлов по различным хранилищам данных входят: путь к хранилищу данных, пользователь и пароль для доступа к хранилищу, форматы файлов для индексирования.

Поиск по файлам хранилищ данных осуществляется по фразе из искомого документа. Пользователь вводит в строку поиска слово или фразу из документа, и в результате получает список файлов из всех проиндексированных хранилищ, в которых была найдена введенная фраза (рис. 8).

```
// строка поиска
$userQuery = CRequest::getString("stringSearch");
// параметр, указывающий, что поиск ведется по содержимому файлов
$params = array(
"content" => 1
);
// объект, содержащий результаты поиска по индексу Apache Solr
$resultObj = CSolr::search($userQuery, $params);
// массив с результатами поиска
$result = newCArrayList();
foreach ($resultObj->getDocuments() as $doc) {
$result->add($doc->id, $doc);
}
```

Рис. 8. Пример кода для поиска по запросу пользователя

При поиске по запросу пользователя по фразе «В работе рассматриваются вопросы» на странице результата поиска отображается ссылка на файл, а также фрагмент текста файла с подсветкой искомой фразы (рис. 9).

#	Результат поиска	Файл
1	ХОМСКОГО *В работе рассматриваются вопросы* выявления общих правил структурирования контента в	Kulikov.docx
2	бизнес-процессов в соответствии с иерархией Хомского. *В работе рассматриваются вопросы* выявления общих	doklad.docx
3	ХОМСКОГО *В работе рассматриваются вопросы* выявления общих правил структурирования контента в	Kulikov.docx

Рис. 9. Результат поиска по фразе «В работе рассматриваются вопросы»

При этом в индексе Solr файл представляется следующим образом (рис. 10).


```
"id": "ef494411b3db61363eb1ae01d2fd6a1d",
"_is_file_": "1",
"filepath": "test/doklad.docx",
"filename": "doklad.docx",
```


Рис. 10. Представление файла в индексе Solr

Заключение

Использование предметно-ориентированного моделирования позволяет представить предметную область как совокупность взаимодействующих подсистем.

Взаимодействие между подсистемами выполняется путем передачи сообщений. В каждом из доменов могут быть сходные операции и сущности, таким образом, актуальной является задача установления отношений семантической эквивалентности между сущностями разных доменов. Данная задача успешно решается с использованием типового проектного решения «адаптер». Задачей адаптера в данном случае является отображение сущностей одного домена в сущности другого.

Задача информационного поиска также может быть представлена как взаимодействие доменов с использованием адаптера. Задача адаптера в данном случае заключается в преобразовании информационной потребности, выраженной на языке конкретного документа в поисковый запрос, представленный на языке информационно-поисковой системы.

Представление поисковых запросов на языке поисковой системы предполагает следующий порядок операций: определение целевой сущности, преобразование критериев, заданных пользователем и добавление критериев, заданных исходной системой. Предлагаемый алгоритм описывает порядок преобразования в общем виде.

Рассмотрен поисковый модуль фреймворка «Фреймворк.АСУ», построенный на базе поискового механизма Apache Solr. Дружественный интерфейс модуля позволяет получать информацию в разрезах, не предоставляемых другими модулями веб-портала, что обеспечивает адаптивность к изменяющимся условиям.

Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 16-37-00064 мол_а.

Литература

1. Корпоративный поиск: технологии Google на службе вашей компании // Каталог программных решений Softline direct, февраль 2013-2(132)-RU.

2. Шерстнев, В.С. Использование Oracle Universal Content Management в качестве корпоративного хранилища документов ТПУ / В.С. Шерстнев, С.С. Иванов, И.А. Акулин // Вестник науки Сибири. – 2011. – № 1 (1). – С. 302–307.

3. Методика интеграции информационно-поисковых и корпоративных информационных систем на основе системных моделей бизнес-процессов / Г.Г. Куликов, Г.В. Старцев, А.А. Бармин, О.В. Бармина // Прикладная информатика. – 2014. – № 1 (49). – С. 6–15.

4. Domain-specific area formal model of content categories based on set-theoretic concepts / G.G. Kulikov, M.A. Shilina, A.A. Barmin, D.G. Shamidanov // Proceedings of the Workshop on Computer Science and Information Technologies (18th CSIT'2016), Czech Republic, Prague, Kunovice. September 26–30, 2016. – Vol. 1. – P. 50–56.

5. The algorithm for generating search queries for structuring the content of heterogenous information space / M.A. Shilina, G.V. Startsev, A.A. Barmin, D.G. Shamidanov // Proceedings of the Workshop on Computer Science and Information Technologies (19th CSIT'2017), Germany, Baden-Baden, 2017. – Vol. 1. – P. 104–110.

Куликов Геннадий Григорьевич, д-р техн. наук, профессор, Уфимский государственный авиационный технический университет, г. Уфа; gennadyg_98@yahoo.com.

Шилина Мария Анатольевна, канд. техн. наук, доцент, Уфимский государственный авиационный технический университет, г. Уфа.

Бармин Александр Александрович, канд. техн. наук, Уфимский государственный авиационный технический университет, г. Уфа.

Старцев Геннадий Владимирович, канд. техн. наук, доцент, Уфимский государственный авиационный технический университет, г. Уфа.

Шамиданов Дмитрий Геннадьевич, аспирант, Уфимский государственный авиационный технический университет, г. Уфа.

Поступила в редакцию 1 декабря 2017 г.

DOI: 10.14529/ctcr180101

METHOD OF STRUCTURING THE CONTENT OF THE HETEROGENEOUS INFORMATION SPACE BASED ON THE FORMALIZED MODEL OF THE SUBJECT DOMAIN FOR SOLVING THE PROBLEMS OF INTELLECTUAL SEARCH

G.G. Kulikov*, M.A. Shilina, A.A. Barmin, G.V. Startcev, D.G. Shamidanov

Ufa State Aviation Technical University, Ufa, Russian Federation

**gennadyg_98@yahoo.com*

The problems of construction and integration of information retrieval systems with corporate information systems are considered. A model of information query and search results for the database and file data stores is described. The task of information retrieval is represented as the interaction of domains using an adapter. The task of the adapter is to convert the information needs, expressed in the language of a particular document into a search query, presented in the language of the information retrieval system. The representation of search queries in the language of the search engine assumes the following order of operations: the definition of the target entity, the conversion of the criteria specified by the user, and the addition of criteria specified by the source system. An algorithm for the order of the transformation is proposed in the general form.

The implementation of the proposed model in the PHP framework of the web portal based on the Apache Solr search engine is considered.

Keywords: full-text search, heterogeneous information space, information retrieval system, corporate application, data warehouses, object domain model, integration of software systems.

References

1. Enterprise information retrieval: working with Google Search Appliance, in Softline direct, February 2013-2(132)-RU.

2. SherstnevV.S., Ivanov S.S., Akulin I.A. [Oracle Universal Content Management as Enterprise Document Storage TPU]. *Siberian Science Journal*, 2011, no. 1 (1), pp. 302–307. (in Russ.)

3. Kulikov G.G., Startsev G.V., Barmin A.A., Barmina O.V. [Method of Integration of Information Retrieval and Enterprise Systems on the Basis of Business Process System Models]. *Applied Informatics*, 2014, no. 1 (49), pp. 6–15. (inRuss.)

4. Kulikov G.G., Shilina M.A., Barmin A.A., Shamidanov D.G. [Domain-Specific Area Formal Model of Content Categories Based on Set-Theoretic Concepts]. *Proceedings of the Workshop on Computer Science and Information Technologies (18th CSIT'2016)*, Czech Republic, Prague, Kunovice, 2016, vol. 1, pp. 50–56.

5. Shilina M.A., Startsev G.V., Barmin A.A., Shamidanov D.G. [The Algorithm for Generating Search Queries for Structuring the Content of Heterogenous Information Space]. *Proceedings of the Workshop on Computer Science and Information Technologies (19th CSIT'2017)*, Germany, Baden-Baden, 2017, vol. 1, pp. 104–110.

Received 1 December 2017

ОБРАЗЕЦ ЦИТИРОВАНИЯ

Метод структурирования контента гетерогенного информационного пространства на основе формализованной модели предметной области для решения задач интеллектуального поиска / Г.Г. Куликов, М.А. Шилина, А.А. Бармин и др. // Вестник ЮУрГУ. Серия «Компьютерные технологии, управление, радиоэлектроника». – 2018. – Т. 18, № 1. – С. 5–16. DOI: 10.14529/ctcr180101

FOR CITATION

Kulikov G.G., Shilina M.A., Barmin A.A., Startcev G.V., Shamidanov D.G. Method of Structuring the Content of the Heterogeneous Information Space Based on the Formalized Model of the Subject Domain for Solving the Problems of Intellectual Search. *Bulletin of the South Ural State University. Ser. Computer Technologies, Automatic Control, Radio Electronics*, 2018, vol. 18, no. 1, pp. 5–16. (in Russ.) DOI: 10.14529/ctcr180101