Навигация по тезаурусам и поиск в распределенных гетерогенных информационных системах

© О. Л. Жижимов © С. А. Сантеева

Институт вычислительных технологий СО РАН (ИВТ СО РАН),
Новосибирский государственный университет (НГУ)
Новосибирск
zhizhim@mail.ru saya_santeeva@mail.ru

Аннотация

Обсуждаются вопросы, связанные с построением пользовательских интерфейсов для навигации по статьям тезаурусов и рубрикаторов в гетерогенных информационных системах. Приводятся некоторые алгоритмы формирования этих интерфейсов с учетом привязки внешних информационных ресурсов к выбранным статьям тезаурусов и рубрикаторов. Основной акцент сделан на динамическую привязку внешних ресурсов на основе текстового поиска по наборам характеристических терминов. Описываются стенд для проведения исследований и результаты исследований на тестовых экспертных наборах данных.

Работа выполнена при поддержке гранта ведущих научных школ НШ-7214.2016.9.

1 Навигация по рубрикаторам и поиск в гетерогенных информационных системах

С развитием технологий построения гетерогенных распределенных информационных систем, включающих в себя множество различных баз данных с различной структурой и содержанием, актуальным становится вопрос поиска информации в базах данных с использованием онтологий, тезаурусов и классификационных схем, представленных в виде отдельных баз данных (БДОТК - базы данных онтологий, тезаурусов и классификаторов).

Существует множество различных способов построения БДОТК, организации доступа к их содержимому и реализации явных и неявных связей между БДОТК и другими гетерогенными информационными ресурсами. Многие из этих способов основаны на строгих онтологических моделях [1,2] и для практической реализации предъявляют очень жесткие требования к организации информационных систем и баз данных вплоть до полной перегрузки информации в промежуточные хранилища, функциональные свойства которых позволяют обеспечить выявление всех семантических связей между информационными объектами на основе заданных онтологических моделей. Такой подход имеет право на существование, однако остается открытым вопрос о том, как включить поиск семантически связанной информации в существующих распределенных гетерогенных информационных ресурсах, причем в случае, когда они не могут быть перегружены в специализированные хранилища.

Настоящая работа посвящена описанию способов поиска семантически связанной информации в распределенных гетерогенных информационных системах (базах данных) без использования специализированных технологий семантического поиска, основанных на моделях Semantic WEB [3-6]. Описание способов будет иллюстрироваться их реализацией в существующих программных продуктах, в частности, на примере программной платформы ZooSPACE [7], предназначенной для интеграции разнородных распределенных информационных систем, успешно функционирующей в ИВТ СО РАН на базе распределенных узлов в городах Новосибирск, Томск, Красноярск и Иркутск и объединяющей сегодня более 70 различных баз данных с общим количеством записей более 60 миллионов.

Несмотря на привлекательность перспектив использования технологий Semantic Web для поиска информации [8], реальность сталкивается с фактом, что подавляющее большинство информационных ресурсов, организованных в виде различных баз данных (реляционных, иерархических, сетевых и пр.), поддерживают прежде всего ту или иную булеву модель атрибутивного поиска информации [9], т.е. поиска, основанного на использовании метаданных и предопределенных индексов (точек доступа).

Нашу задачу можно сформулировать и так: требуется найти все записи в некотором множестве гетерогенных баз данных, которые бы соответствовали определенной онтологической

Труды XVIII Международной конференции DAMDID/RCDL'2016 «Аналитика и управление данными в областях с интенсивным использованием данных», Ершово, 11-14 октября 2016

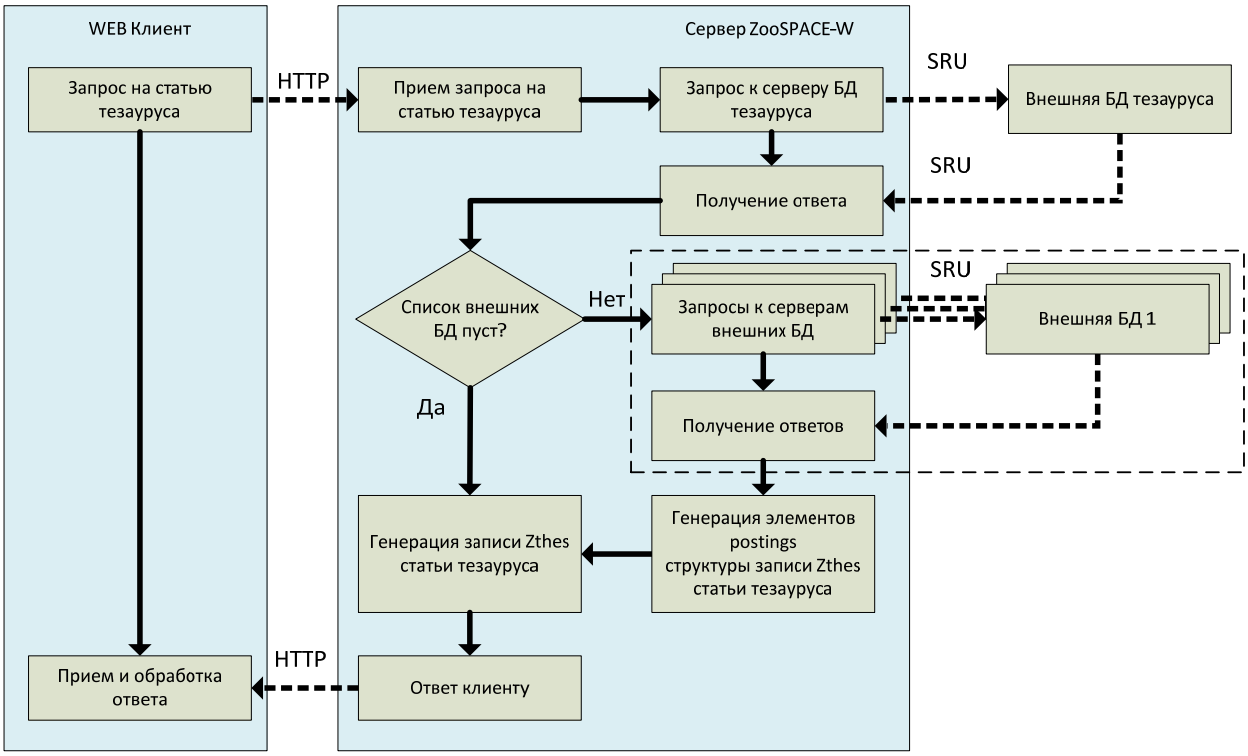

Рисунок 1 Формирование структуры записи статьи тезауруса с динамическими связями с внешними базами данных

сущности (статье тезауруса, рубрике, коду рубрикатора и пр.). Для определенности ниже эту онтологическую сущность мы будем ниже называть статьей тезауруса, понимая, что на ее месте может быть и другое. В качестве решения можно рассматривать алгоритм получения результата, реализованный в виде функционирующего серверного программного модуля для некоторой информационной системы. Эта задача практически полностью эквивалентна задачи навигации по статьям тезауруса, когда для текущей статьи тезауруса отображается информация о связанных с этой статьей записях из выбранного множества в общем случае гетерогенных баз данных. При этом «привязка» связанных записей баз данных к статье тезауруса должна быть динамической, т.е. формироваться в процессе формирования представления собственно статьи тезауруса.

Итак, клиент, используя WEB-браузер может просматривать тезаурус, перемещаясь по связанным статьям. Каждая выбранная статья тезауруса должна быть представлена клиенту в виде некоторой универсальной структуры, которая может быть однозначно интерпретирована, т.е. эта структура должна соответствовать какой-нибудь стандартной схеме данных, используемой для описания статей тезауруса. Ниже везде мы будем использовать схему данных ZThes [10] в формате XML [11]. Также мы будем подразумевать, что все необходимые обращения к серверам баз данных будут соответствовать спецификациям SRU [12] с языком запросов RPN [13] в синтаксисе PQF [14]. Этот язык запросов отличается от стандартного для SRU языка запросов CQL, но на наш взгляд он более удобен для формирования запросов и, что немаловажно, более нагляден.

На Рисунке 1 схематично представлен алгоритм работы клиента и сервера при просмотре статьи тезауруса.

Выбор клиентом статьи тезауруса порождает обращение к WEB-серверу, который в свою очередь формирует запрос к серверу баз данных, хранящему информацию о текущем тезаурусе (БД тезауруса). Этот запрос соответствует запросу на поиск записи (статьи) по ее однозначному идентификатору, в результате его выполнения должна быть получена запись БД, соответствующая требуемой статье тезауруса и содержащей полную информацию о ней.

Если клиентом был сформулирован список баз данных, записи которых следует соотнести с текущей статьей тезауруса, должен быть включен механизм формирования запросов к каждой базе данных из выбранного списка, выполнения этих запросов на соответствующих серверах БД, получение ответов и формирование специальных элементов (postings) в записи статьи тезауруса, содержащих информацию об именах баз данных и количестве найденных записей [10]. Заметим, что выполнение запросов к внешним базам данных может происходить параллельно с асинхронным завершением.

```
<term>
 . . .
  <relation>
    <termID>31.27.20</termID>
```

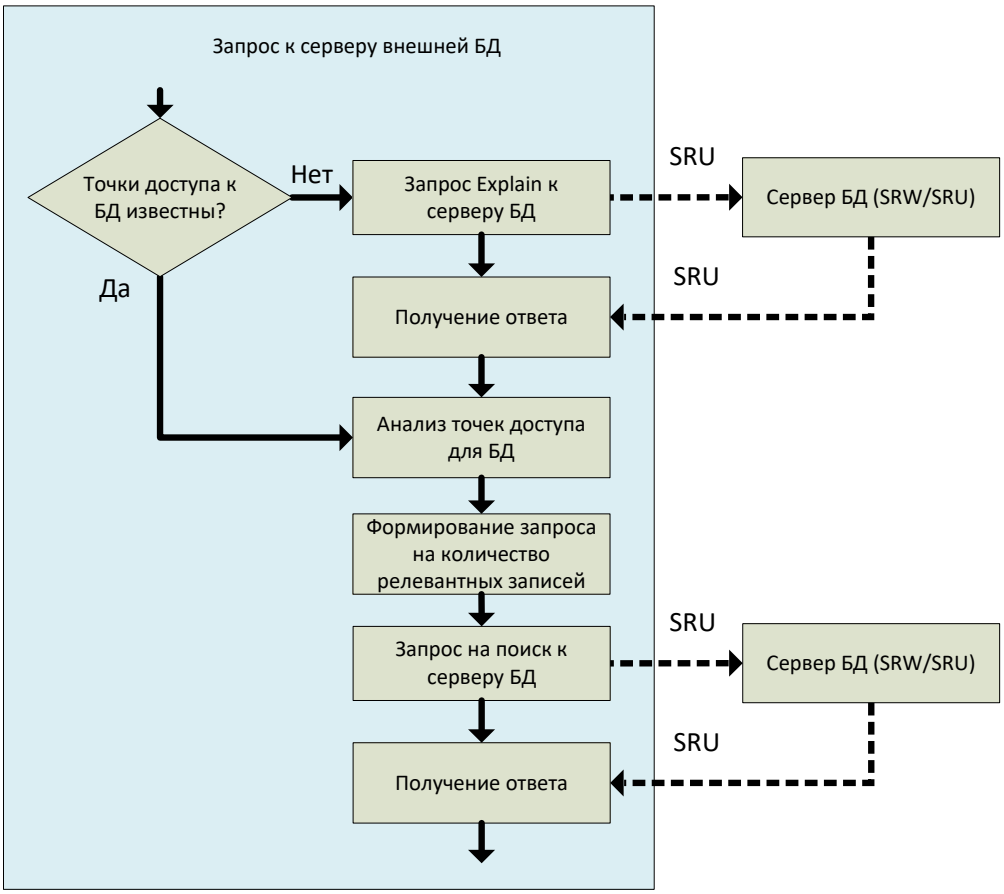

Рисунок 2 Формирование и исполнение запроса к внешней базе данных

```
    <relationType>RT</relationType>
    <termQualifier>31.27.20</termQualifier>
    <termName>Биохимия вирусов</termName>
    <termLanguage>rus</termLanguage>
  </relation>
  <postings>
    <SourceDB>AB</SourceDB>
    <hitCount>1022</hitCount>
  </postings>
</term>
```

Несомненно, самым критичным блоком этого алгоритма является блок формирования запросов к серверам внешних БД (на Рисунке 1 это блок обведен пунктиром). Именно от работы этого блока зависит качество динамической привязки записей внешних БД к текущей статье тезауруса. Работа этого блока представлена на Рисунке 2.

Прежде, чем сформировать запрос к серверу внешней БД, необходимо выяснить возможности этой БД в смысле поиска информации, т.е. в терминах SRU (или Z39.50) определить поддерживаемые поисковые атрибуты и варианты их комбинаций. Если отбросить тривиальные и маловероятные конфигурации с фиксированными точками доступа, существует только один регулярный способ – предварительно выполнить запрос explain (SRU, SRW, Z39.50) и проанализировать полученную структуру на предмет выявления поддержки требуемых поисковых атрибутов.

Например, из записи Explain можно сделать вывод, что внешняя база данных поддерживает поисковые атрибуты USE (type 1) 14 (УДК) и 21 (ключевые слова), операция сравнения - «равно» (type 2 = 3), поисковые термы интерпретируются как строки или слова (type 4=1,2,108), поиск возможен как по точному совпадению (type 5=100), так и по усечению справа (type 5=1). Поэтому к этой БД мы можем обращаться с поиском «по ключевым словам» и кодам рубрикатора УДК, т.е. если текущая статья нашего тезауруса (рубрикатора) является описанием рубрики УДК, то запрос к внешней БД должен выглядеть следующим образом (RPN в синтаксисе PQF):

```
@attr 1=14 @attr 5=1 {term}
```

где вместо «term» должен фигурировать код текущей рубрики. Следует заметить, что здесь запрос сформулирован с усечением справа, т.е. будут найдены все записи, коды УДК которых начинаются с символов «term». Для иерархических рубрикаторов это означает, что к текущей рубрике будут привязаны записи БД, содержащие коды УДК не только текущей, но и всех дочерних рубрик.

В случае тезауруса каждая статья идентифицируется ее заголовком, поэтому поиск во внешних БД следует выполнять по ключевым словам, причем по полному их совпадению:

```
@attr 1=21 {term}
```

где вместо «term» должен фигурировать заголовок текущей статьи тезауруса.

Строго говоря, такие запросы к внешним БД возможны только тогда, когда

1. Для рубрикаторов:
   a. для всех внешних БД возможен поиск по кодам текущего рубрикатора
2. Для тезаурусов:
   a. для всех внешних БД возможен поиск по ключевым словам
   b. ключевые слова для всех внешних БД сгенерированы из заголовков статей текущего тезауруса.

Последнее условие (2b) практически никогда не выполняется, поскольку разработчики той или иной внешней БД могут использовать тезаурусы, отличающиеся от нашего текущего, или не использовать вообще никакие, выбирая ключевые слова для записей БД в соответствии со своими правилами.

Возникает вопрос - как можно соотносить записи внешних БД с текущей статьей тезауруса при нарушении приведенных выше условий.

Для рубрикаторов при нарушении условия 1a возможны два варианта:

1. поиск по связанным кодам других рубрикаторов
2. поиск по текстовым характеристикам статьи рубрикатора

1.1 Поиск по связанным кодам других рубрикаторов

Поиск по связанным кодам других рубрикаторов может быть полезен, когда внешняя база проиндексирована по этим кодам. Действительно, если внешняя БД не проиндексирована по кодам текущего рубрикатора, например, ГРНТИ, но проиндексирована по кодам УДК, наличие связи между статьей рубрикатора ГРНТИ и статьями УДК позволяет выполнить динамическую привязку записей из внешней БД не по кодам ГРНТИ, а по кодам УДК.

```
<Zthes>
. . .
<term>
  <termID>20.23.19</termID>
  <termQualifier>20.23.19</termQualifier>
  <termName>
     Процессы информационного поиска
  </termName>
  <termType>NT</termType>
  <termLanguage>rus</termLanguage>

  <relation>
    <termID>20.23</termID>
    <relationType>BT</relationType>
    <termQualifier>20.23</termQualifier>
    <termName>Информационный поиск</termName>
    <termLanguage>rus</termLanguage>
  </relation>
  <relation>
    <SourceDB>ruudc</SourceDB>
    <relationType>RT</relationType>
    <termQualifier>025.4.03</termQualifier>
  </relation>
</term>
</Zthes>
```

Технически динамическая привязка записей из внешней БД осуществляется также, как описано выше.

1.2 Поиск по текстовым характеристикам статьи рубрикатора

Если внешняя БД не проиндексирована по кодам текущего и связанных рубрикаторов, динамическая привязка ее записей к статьям текущего рубрикатора становится задачей нетривиальной.

Действительно, для того чтобы записи из внешних БД могли быть динамически привязаны к текущей рубрике, необходимо иметь поисковый образ документов, соответствующих этой рубрике. При этом почти очевидно, что для такого поискового образа практически бесполезна текстовая информация, которая обычно присутствует в описании статьи рубрикатора (название, описание, названия связанных рубрик и т.п.). Тем не менее, можно придумать схему динамической привязки, основываясь, например, на векторной модели поиска и дополнительной информации, которой должна быть дополнена каждая статья рубрикатора.

В векторной модели поиска в качестве поискового образа выступает некоторый уникальный для каждой статьи рубрикатора вектор, определенный в многомерном пространстве в декартовой системе координат, каждая ось которой соответствуют своему уникальному термину из фиксированного списка терминов, характеризующих данную рубрику (Q_1, Q_2, ..., Q_n) [15]. Если рассматривать каждую запись внешней БД как аналогичный вектор в пространстве встречающихся в ней терминов (X_1, X_2, ..., X_m), то можно говорить о скалярном произведении векторов **Q** и **X**. Чем больше это скалярное произведение, тем выше релевантность записи **X** запросу **Q**. Критерием отбора записей может быть выполнение условия

$$\frac{1}{n}(\boldsymbol{Q}\cdot\boldsymbol{X}) \geq s,\ \ s \leq 1$$

Таким образом, для реализации динамической связи записей из внешних БД со статьей текущего рубрикатора, необходимо:

Наличие для каждой статьи рубрикатора уникального характеристического вектора. Этот вектор может быть построен только в результате обработки большого количества документов, уже имеющих в результате экспертной оценки коды рубрик текущего рубрикатора. При этом для каждой рубрики количество обработанных документов должно быть достаточно большим. Вопрос о достаточной размерности вектора, т.е. о количестве необходимых характеристических терминов зависит от структуры рубрикатора и может быть решен в результате тестов.

Определение параметра s, характеризующего минимально допустимое значение скалярного

произведения векторов при поиске может быть произведено в результате тестов.

Наличие возможности серверами БД обрабатывать поисковые запросы, соответствующие векторной модели поиска. Это требование как правило выполняется для поисковых систем, ориентированных на неструктурированную и слабоструктурированную информацию. Серверы БД ориентированы на булеву модель [9] поиска, что затрудняет использование обсуждаемой технологии привязки записей. Тем не менее в простейшем варианте без использования частот встречаемости терминов в документе и в наборе документов, поисковый запрос, соответствующий векторной модели, может быть представлен в булевом виде.

В качестве примера рассмотрим характеристический вектор длиной n=4 с терминами a, b, c, d: **Q** = (a, b, c, d)/4.

Таблица 1

	Булевый запрос (& - AND, | - OR)	**s**	**K**
1	a&b&c&d	1	0
2	(a&b&c) | (a&b&d) | (a&c&d) | (b&c&d)	0,75	1
3	(a&b) | (a&c) | (a&d) | (b&c) | (b&d) | (d&c)	0,5	2
4	a | b | c | d	0,25	3

при этом количество групп, объединенных операторами OR, равно количеству сочетаний из n элементов по k: $k!/(n-k)!$, причем $s = (n - k)/n$.

Из приведенного примера видно, что

1. При заданной длине n вектора запроса **Q** параметр критерия отбора a принимает дискретные значения в интервале ($0<s\leq1$) с шагом 1/n, причем $s = (n - k)/n$.
2. Каждый булевый запрос для фиксированного s (или k) перекрывает все запросы с большими s (меньшими k).
3. При фиксированном параметре s для поиска необходимо исполнить только один запрос, который содержит $n!/k!\,(n-k)!$ групп по $(n\text{-}k)$ термов. При этом количество участвующих в запросе термов равно $n!/k!\,(n-k-1)!$.
4. Группы объединяются оператором OR (ИЛИ), термы внутри группы объединяются оператором AND (И).

Наконец, можно сделать некоторое предположение для иерархических рубрикаторов. Если нас интересует запрос для рубрики N.M.L, для которой определен характеристический вектор $\mathbf{q}_{NML}$ и соответствующий частный запрос q_{NML}, то действующим запросом для рубрики N.M.L, будет запрос вида

$$Q_{LMN} = Q_{NM} \;\&\; q_{NML} = q_N \;\&\; q_{NM} \;\&\; q_{NML}$$

где частные запросы q_N и q_{NM} соответствуют характеристическим векторам $\mathbf{q}_N$ и $\mathbf{q}_{NM}$ для рубрик N и N.M соответственно.

Таким образом, поиск по текстовым характеристикам статьи рубрикатора возможен и может быть реализован в соответствии с упрощенной векторной моделью поиска конвертированием векторных запросов в булеву форму.

2 Экспериментальный стенд и результаты тестирования

Для проверки качества работы описанного выше механизма поиска во внешних ресурсах по текстовым характеристикам статей рубрикаторов были использованы:

1. База данных «Рубрикатор ГРНТИ», доступ к которой предоставлялся по протоколам Z39.50 и SRU в соответствии со спецификациями Zthes на платформе ZooSPACE.
2. Специализированная база данных (СБД), содержащая записи РЖ ВИНИТИ (Информатика, Автоматика, Вычислительные науки) с проставленными экспертами ВИНИТИ кодами ГРНТИ для групп кодов:
   a. 20.*.* - Информатика
   b. 28.*.* - Кибернетика
   c. 50.*.* - Автоматика и телемеханика. Вычислительная техника.

   по 200 записей для каждого кода. Для упрощения обработки эта БД была загружена в СУБД PostgreSQL с активизацией функций полнотекстового поиска в полях Title, Subject, Abstract.
3. Для числовых характеристик, описывающих качество поиска, использовались метрики [16]:

Таблица 2

	Релевантный	Не релевантный
Найдено	a	b
Не найдено	c	d

Полнота: $r = a/(a + c)$

Точность: $p = a/(a + b)$

Ошибка: $e = (b + c)/(a + b + c + d)$

F-мера: $F = 2pr/(p + r)$

Для каждой рубрики ГРНТИ в указанных выше группах на основании частоты встречаемости терминов в различных записях СБД и выполнения запроса к СБД по этому термину был определен ранжированный по убыванию F список слов из заголовков, ключевых слов и аннотаций для соответствующих записей СБД.

На основе этого списка для каждой рубрики ГРНТИ может быть построен наиболее эффективный

по вышеуказанным метрикам характеристический вектор. Мы использовали критерий максимального значения F при варьировании параметров n и s.

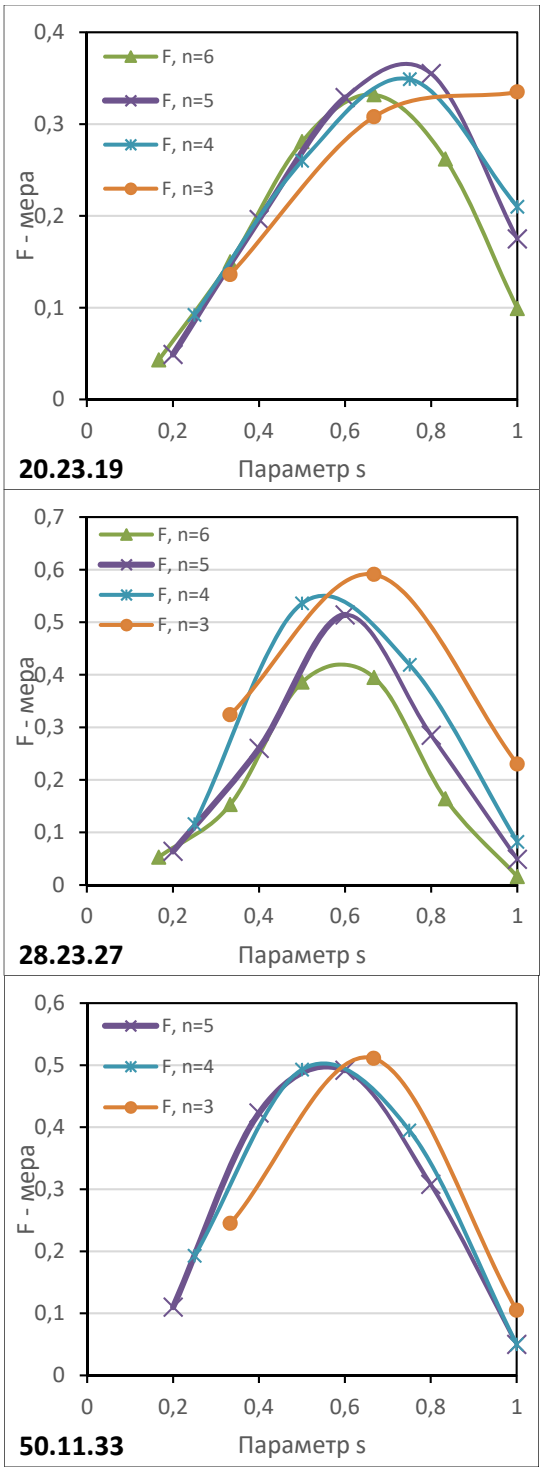

Рисунок 3 Характерная зависимость F от длины характеристического вектора (n) и параметра s для записей с разными кодами ГРНТИ.

В качестве примера приведем ранжированный список терминов для некоторых кодов ГРНТИ. Зависимость F-меры от длины характеристического вектора и параметра s представлена на рисунке 3 для трех кодов ГРНТИ. При этом для наиболее оптимальных значений n и s в таблице 4 приведены значения метрик.

Таблица 3

ГРНТИ	Термины
20.23.19 - Процессы информационного поиска	поисковый, запрос, поиск, документ, информационный, пользователь, обработка, база, …
28.23.27 - Интеллектуальные робототехнические системы	робот, мобильный, движение, алгоритм, управление, предлагаться, …
50.11.33 - Оптические запоминающие устройства	оптический, дисковод, воспроизведение, носитель, диск, запись, память, …

Таблица 4

	20.23.19	**28.23.27**	**50.11.33**
N	5	3	3
S	0,8	0,667	0,667
K	1	1	1
R	0.360	0,508	0,430
P	**0,350**	**0,707**	**0,629**
E	0,009	0,009	0,005
F	0,355	0,591	0,511

Таким образом, при наличии СБД можно определить для каждой рубрики:

1. Ранжированный список терминов
2. Длину и содержание характеристического вектора
3. Оптимальное значение параметра s (или k)

На основании этой информации можно построить булевый поисковый запрос по текстовым атрибутам, который наиболее полно будет соответствовать запросу по соответствующему коду рубрикатора. При этом вероятность нахождения нужных записей в найденном таким образом множестве записей предварительно известна и равна значению p.

Например, вместо запроса по коду ГРНТИ 28.23.27 можно выполнять запрос вида

```
(робот & мобильный)
| (робот & движение)
| (мобильный & движение)
```

Результат выполнения этого запроса будет содержать нужные данные с вероятностью 0,7.

Следует заметить:

1. Описанный механизм привязки внешних ресурсов к кодам рубрикаторов хорошо работает для «грубых» рубрикаторов.
2. Для иерархических рубрикаторов и рубрикаторов с «родственными» рубриками качество поиска является удовлетворительным. При этом поисковые метрики сильно зависят от длины

характеристических векторов и значения критерия отбора. Обе этих характеристики могут быть получены на основе анализа экспертных данных.

В заключение следует заметить, что на основе изложенных выше методов и алгоритмов в настоящее время разрабатываются программные модули для системы ZooSPACE, реализующие графические пользовательские интерфейсы для навигации по тезаурусам и рубрикаторам с привязкой информации из разнородных источников.

Литература

[1] Web Ontology Language (OWL), http://www.w3.org/2004/OWL/

[2] Онтология (информатика) — Материал из Википедии - свободной энциклопедии — http://ru.wikipedia.org/wiki/Онтология_(информатика)

[3] Semantic Web, http://www.w3.org/2001/sw/

[4] Tim Berners-Lee, James Hendler and Ora Lassila, The Semantic Web, http://sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21

[5] Metadata Architecture, http://www.w3.org/DesignIssues/Metadata

[6] W3C standards, http://w3.org/sw/

[7] Жижимов О.Л., Федотов А.М., Шокин Ю.И. Технологическая платформа массовой интеграции гетерогенных данных // Вестник Новосибирского государственного университета. Серия: Информационные технологии. - 2013. - Т.11. - № 1. - С.24-41. - ISSN 1818-7900.

[8] Guha R. Semantic search / R. Guha, R. McCool, E. Miller // Proceedings of the 12th international conference on World Wide Web. – N.Y. ACM Press, 2003. – P. 700–709.

[9] Шарапов Р.В., Шарапова Е.В., Саратсвцева О.А. Модели информационного поиска. http://vuz.exponenta.ru/PDF/FOTO/kaz/Articles/sharapov1.pdf

[10] The Zthes specifications for thesaurus representation, access and navigation - http://zthes.z3950.org/

[11] Extensible Markup Language (XML), http://www.w3.org/XML/

[12] SRU - Search/Retrieve via URL// The Library of Congress. - USA - http://www.loc.gov/standards/sru/

[13] RPN - https://www.loc.gov/z3950/agency/markup/09.html

[14] Mike Taylor. PQF - http://search.cpan.org/dist/Net-Z3950-PQF/lib/Net/Z3950/PQF.pm

[15] Э. Мбайкоджи, А.А. Драль, И.В. Соченков. Метод автоматической классификации коротких текстовых сообщений. http://elib.ict.nsc.ru/jspui/bitstream/ICT/1396/1/93_102.pdf

[16] М. Агеев, И. Кураленок, И.Некрестьянов. Официальные метрики РОМИП 2006 http://romip.ru/romip2006/appendix_a_metrics.pdf

Thesaurus navigation and search in the distributed heterogeneous information systems

Oleg L. Zhizhimov, Saya A. Santeeva

The issues related to creation of the user interfaces for navigation through the articles of thesauruses and rubricators in heterogeneous information systems are discussed. The algorithms of formation of these interfaces taking into account a binding of external information resources to the chosen articles of thesauruses and rubricators are given. The main emphasis is placed on a dynamic binding of external resources based on text search in the sets of characteristic terms. The workbench for studies as well as the research results obtained on the testing expert data sets are described.