ФОРМИРОВАНИЕ РЕКОМЕНДАЦИЙ В СЕМАНТИЧЕСКИХ ЭЛЕКТРОННЫХ БИБЛИОТЕКАХ

Ле Хоай, А. Ф. Тузовский

Институт кибернетики Национального исследовательского
Томского политехнического университета, 634034, Томск, Россия

УДК 004.02:004.82

Рассматривается задача формирования списков электронных ресурсов, рекомендуемых пользователям семантических электронных библиотек, сформированных на основе контекстных метаданных. Анализируются возможные методы решения и обосновывается решение данной задачи с использованием семантических технологий. Приведены экспериментальные результаты работы предложенного алгоритма.

Ключевые слова: метаописание документов, контекстные метаданные, семантическая близость, семантические технологии, семантическая электронная библиотека, рекомендация документов.

This paper examines the task of creating users's digital document recommendations in digital libraries based on contextual metadata. Methods for solving this problem are discussing, and arguing with selected semantic technologies, and also the experimental results of the proposed algorithm are showing.

Key words: document metadescription, contextual metadata, semantic similarity, semantic technologies, semantic digital library, document recommendation.

Введение. Применение семантических технологий для улучшения функциональности электронных библиотек и решения их основных задач является перспективным направлением [1]. Семантические электронные библиотеки (СЭБ) представляют собой новое поколение электронных библиотек, при этом семантика содержания информационных объектов – электронные ресурсы (ЭР) (документы, изображения, профиль пользователя и т. д.) – описываются явно с помощью специальных языков.

При разработке СЭБ необходимо решить ряд задач, таких как категоризация ЭР, их поиск и формирование рекомендаций пользователям [1]. Используемый в настоящее время подход к решению задачи формирования рекомендаций ЭР на основе ключевых слов имеет много недостатков, обусловленных омонимией, полисемией и синонимией языка. Кроме того, в традиционном подходе не учитывается семантика содержания ЭР.

Явное описание семантики содержания ЭР на основе онтологий подразумевает процесс составления набора утверждений в виде триплетов, состоящих из трех компонентов субъект – предикат – объект. Использование таких описаний позволяет выполнять вычисление семантической близости между электронными ресурсами библиотек.

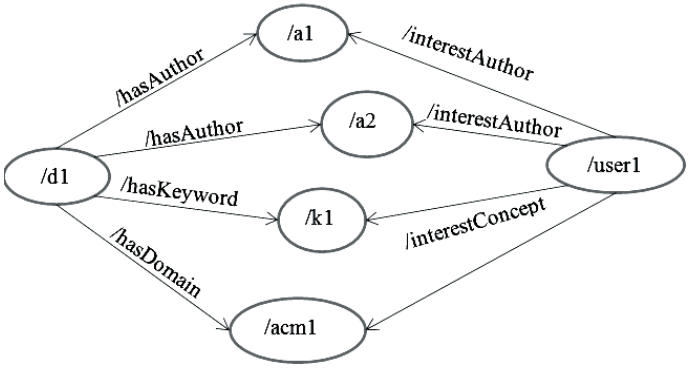

Рис. 1. Пример графа связей описания документа и пользователя в системе SemDL

В данной работе приводится теоретическая основа решения сформулированной задачи на основе семантики, предлагаются методы решения поставленной задачи в созданной семантической электронной библиотеке SemDL, на основе которой получены экспериментальные результаты, позволяющие выполнить оценку качества и эффективности разработанных алгоритмов.

Формирование рекомендаций в семантических электронных библиотеках. В каждой системе, разработанной на основе семантических технологий, используется набор определенных онтологий, позволяющий описывать в данной системе любые информационные объекты. Описания электронных ресурсов в СЭБ создают их метаописания, в которые входят системные (служебные), структурные метаданные и семантические данные [2]. Системные метаданные обеспечивают функционирование информационной системы (имена файлов, даты их создания, размеры и т. д). Структурные метаданные, как правило, содержат справочную информацию об объектах (наименование, статус, профили и т. д.) и используются для идентификации и классификации объектов в различных целях. Семантические метаданные представляют собой особый вид описаний, включающих концептуальное (аннотированное) описание содержания и смысла информации об объекте. В семантические данные ЭР входят контекстные и контентные метаданные. Контекстные метаданные описывают внешние взаимосвязи ЭР с другими объектами и понятиями СЭБ или имеют литеральные значения, а контентные метаданные ЭР описывают знания и информацию, содержащиеся в электронных ресурсах и различных объектах библиотеки (например, профили пользователей, содержание каталогов).

Семантические метаданные формируют граф, связывающий между собой описания электронных ресурсов и профилей пользователей библиотеки. Например, можно выделить следующие связи показанного на рис. 1 графа, которые являются важными для решения задачи формирования рекомендаций:

1. Связи с профилями авторов документа: документ имеет некоторых авторов или соавторов, представляющих интерес для пользователя библиотеки.

2. Связи с ключевыми словами: документ аннотирован определенными авторами документа ключевыми словами, представляющими интерес для пользователя. Ключевые слова могут быть выбраны из онтологий предметных областей.

3. Связи с предметными областями в некоторой онтологии типа таксономии. Содержание документов включает понятия и объекты из этих областей, представляющих интерес для пользователя.

В системе могут использоваться также другие связи между объектами, аналогичные связям между описаниями авторов и ключевых слов. Различие состоит только в трудоемкости вычисления близостей. Другие связи могут не интересовать пользователей данной системы. Идеальным решением является возможность выбора пользователями связей для выполнения анализа. В настоящей работе основное внимание уделяется использованию контекстных метаданных, связанных с рассматриваемым объектом (профиль пользователя или документ).

Постановка задачи. Формирование рекомендаций формально можно описать следующим образом:

1. На основе онтологий O в СЭБ для каждого документа d набора $D = \{d_i\}$ создаются семантические метаописания $m(d) = \{t_i | i{:}=1 \div k,\ t_i \in T\}$, где t_i – триплет (RDF); k – количество триплетов в $m(d)$; T – множество возможных триплетов, составленных на основе онтологий O.

2. Для каждого шаблона s из множества шаблонов S (метаописания профиля пользователя или документа) в системе также составляется метаописание $m(s) = \{t_j | j{:}=1 \div h,\ t_j \in T\}$, где k – количество триплетов в $m(s)$.

3. Задается весовая функция w, которая присваивает значимость любому триплету при описании документа d и шаблона s: $0 \le w(t, d) \le 1$ и $0 \le w(t, s)$.

Задача формирования рекомендаций заключается в вычислении семантической близости между метаописаниями шаблона s и документами d набора D. В результате должно быть получено упорядоченное по убыванию значения близости множество документов R, для которых значение близости превышает заданное пороговое значение ε. В настоящей работе вычисляется близость контекстных метаданных, в случае когда при рассмотрении каждого триплета, используемого для вычисления близости, учитывается только его объект v ($t_i = \{s_i, p_i, v_i\}$) (v_i могут быть авторами, ключевыми словами и предметными областями, имеющими весовые коэффициенты). Имеется следующая упрощенная формула для вычисления близости:

$$Sim_{sem}\left(m(d), m(s)\right) = Sim_{sem}\left(mdv, msv\right) \tag{1}$$

(mdv – множество значимых частей триплетов метаописания документа d; msv – шаблон s).

Задача формирования рекомендации заключается в предоставлении ЭР заинтересованным пользователям на основе метаописания семантических данных шаблона. Шаблонами в данном случае могу быть метаописания интересов пользователя или документа. Иными словами, решение данной задачи дает возможность пользователю СЭБ получать список рекомендованных документов, связанных с описанием его интересов, а также предоставляет возможность просмотра набора документов, сходных с открытым документом. Такая задача в электронной библиотеке является типичной и востребованной пользователями.

Метод решения задачи формирования рекомендаций. Имеется следующая формула для вычисления семантической близости между mdv и msv:

$$Sim_{sem}\left(mdv, msv\right) = \frac{\sum\limits_{v_i \in mdv}^{k} \sum\limits_{v_j \in msv}^{h} sim(v_i, v_j)}{\sqrt{\sum\limits_{v_i \in mdv}^{k} \sum\limits_{v_j \in mdv}^{h} sim(v_i, v_j)} \sqrt{\sum\limits_{v_i \in msv}^{k} \sum\limits_{v_j \in msv}^{h} sim(v_i, v_j)}}. \tag{2}$$

В формуле (2) семантическая близость метаописаний документа d и шаблона s на основе значимых частей их триплетов в векторном пространстве заключается в вычислении значения функции $\mathrm{sim}_{\mathrm{sem}}(mdv, msv)$ и $\mathrm{sim}_{\mathrm{sem}}\left(mdv, msv\right) \in [0,1]$.

Для вычисления $\mathrm{sim}(v_i, v_j)$ предлагается использовать два известных метода [2, 3], реализуемых на языках запросов SPARQL (simple protocol and RDF query language) или SERQL (sesame RDF query language).

Первый метод состоит в использовании следующей формулы:

$$\mathrm{sim}_c\left(c_i, c_j\right) = k_{st} \frac{\left|C_{ANC}(c_k) \cap C_{ANC}(c_l)\right|}{\left|C_{ANC}(c_k) \cup C_{ANC}(c_l)\right|}. \tag{3}$$

Здесь $C_{ANC}\left(c_i\right) = \left\{c_j \in C \mid T_C(c_i, c_j) \cup c_j = c_i\right\}$ – множество понятий, предшествующих понятию c_i, а также само понятие c_i в таксономической иерархии; запись $T_C(c_i, c_j)$ означает, что c_i предшествует c_j или c_i следует за c_j; параметр k_{st} определяется следующим образом:

$$k_{st} = \begin{cases} 1, \text{ если } C_{ANC}(c_k) \cup C_{ANC}(c_l) = C_{ANC}(c_l), \\ 0, \text{ иначе} \end{cases} \tag{4}$$

и означает, что измерение проводится только в том случае, если c_k предшествует c_l и $\mathrm{sim}_c\left(c_i, c_j\right) \in [0,1]$.

Второй метод состоит в использовании формулы

$$\mathrm{sim}_{WP}\left(c_1, c_2\right) = \frac{2depth(LSC)}{depth(c_1) + depth(c_2)}, \tag{5}$$

где $depth(c)$ – глубина понятия с подсчетом ребер; $depth(LCS)$ – глубина нижнего общего родителя; $\mathrm{sim}_{WP}\left(c_1, c_2\right) \in [0,1]$.

Для вычисления семантической близости могут быть использованы также другие методы, описание которых приведено в [4].

Решение задачи. Для решения данной задачи необходимо выполнить два основных шага: 1) фильтрация множества документов; 2) вычисление близости.

Фильтрация множества документов с помощью языка SERQL. Решение задачи рекомендации существенно осложняется в том случае, если для каждого документа в системе будет выполняться извлечение значимой части триплетов, а затем будут вычисляться их близости со значимыми частями триплетов шаблона. Для уменьшения объемов вычислений необходимо заранее фильтровать множество документов, каждый из которых имеет хотя бы одну значимую часть, совпадающую со значимой частью триплета шаблона, или хотя бы одну часть, являющуюся семантически близкой к значимым частям триплета шаблона. В данной работе эту проблему предлагается решать с помощью запроса к базе знаний (БЗ) на языке SERQL (sesame RDF query language).

Язык SERQL является простым и интуитивно понятным, подробное описание данного языка приведено в [5]. Ниже приведен запрос к БЗ, входными данными которого является только URI шаблона.

URI "**{semdl:lc}**" шаблона определяет все объекты "**{S}**" триплетов, описывающих шаблон. Объекты "**{S}**" могут быть конкретными объектами "**{S2}**" в некоторых онтологиях предметных областей.

```
SELECT * FROM
{semdl:lc} O {S},
{O}rdf:type
owl:OWL.OBJECTPROPERTY.getLocalName,
{S1} O1 {S}
WHERE S1 != semdl:lc AND namespace(S1) =
res: AND (namespace(O) = marcont:
OR namespace(O) = foaf:) UNION
USING NAMESPACE
marcont=<http://www.marcont.org/ontology/2.0,
res=MARCONT._RESOURCE,
foaf=<http://xmlns.com/foaf/0.1/>,
semdl=<ns>,
rdf=<http://www.w3.org/1999/02/22-rdf-syntax-
ns#>,
owl=<OWL.OBJECTPROPERTY.getNamespace()>;
skos = <http://www.w3.org/2004/02/skos/core#>,
```

```
SELECT * FROM {semdl:lc} O {S},
{O}rdf:type
owl:OWL.OBJECTPROPERTY.getLocalName,
{S} skos:narrowerTransitive {S2},
{S1} O1 {S2}
WHERE  S1 != semdl:lc AND namespace(S1)
= res: AND (namespace(O) = marcont:
OR namespace(O) = foaf:)
     lc,ns:переменные от URI шаблона;
     S – URI значимых частей триплетов
     шаблона;
     S2 – URIs значимых частей в таксо-
     номии;
     S1 – URIs фильтруемые документы.
```

Наконец, если существуют триплеты для субъекта "**{S1}**" URI документа, то данный субъект будет рекомендуемым документом, за исключением заданного шаблона. Выполнение данного запроса к БЗ при использовании логического вывода позволяет извлекать все документы, релевантные заданному шаблону. Некоторые документы в множестве R могут повторяться. Это означает, что они включают больше одной значимой части триплета во множестве триплетов шаблона. В дальнейшем такие повторения будут удаляться из множества R.

Вычисление близости понятий. Вычисления близости понятий по формулам (3), (5) имеют связи в таксономии предметной области. Глубина "depth(c)" одного понятия равна количеству понятий, предшествующих ему, и верны следующие выражения:

$$\left|C_{ANC}(c)\right| = depth(c) + 1,$$

$$\left|C_{ANC}(c_k) \bigcap C_{ANC}(c_l)\right| = \min\left(depth(c_k), depth(c_l)\right) + 1 = depth(c_k) + 1 \; \forall k_{st} = 1,$$

$$\left|C_{ANC}(c_k) \bigcup C_{ANC}(c_l)\right| = \max\left(depth(c_k), depth(c_l)\right) + 1 = depth(c_l) + 1 \; \forall k_{st} = 1,$$

$$depth(LSC) = \max_{c_i \in I}(depth(c_i)), \quad I = C_{ANC}(c_k) \bigcap C_{ANC}(c_l).$$

Следовательно, для вычисления всех компонентов (4) достаточно вычислить глубину двух понятий путем составления утверждения (триплета) для нахождения всех понятий, предшествующих понятию искомой глубины, и на основе двух значений глубины понятий вычислить остальные значения.

Алгоритм решения задачи. Алгоритм решения учитывает следующие условия: $Sim(c_i, c_j) = 0$, если $i = j$ или c_i, c_j принадлежат разным онтологиям предметной области (таксономиям), иначе для вычислений используется формула (4) или (6). Ниже приведены псевдокоды алгоритмов вычисления.

```
1:  Умножение(List<c> mdv, List<c> msv){
2:          цикл (c v_i : mdv){
3:            цикл (c v_j : msv){
4:              возврат += Sin(v_i, v_j);// count by (4) or (6)
5:            }
6:          }
7:  }
8:  БлизостьDS(List<c> mdv, List<c> msv)    {
9:          возврат
10:         Умножение(mdv,msv)/SQRT(Умножение(mdv,mdv)* Умножение(msv,msv));
11: }
12:
```

Функция *multipleDS*(*List<c> mdv, List<c> msdv*) вычисляет выражение $\sum_{v_i \in mdv}^{k} \sum_{v_j \in msv}^{h} sim(v_i, v_j)$, а функция *measureDS*(*List<c> mdv, List<c> msv*) – выражение (3).

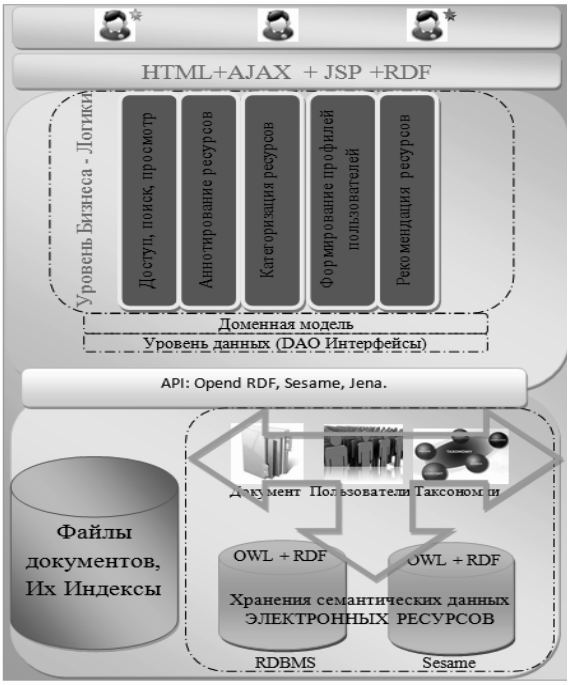

Рис. 2. Схема системы SemDL

Эксперименты. Эксперименты проводились с использованием разрабатываемой семантической электронной библиотеки SemDL. Программная модель данной системы и выполняемые ею функции показаны на рис. 2. Видно, что в качестве БЗ используется известная система Sesame [6], доступ к которой реализован в виде веб-сервиса, разработанного по технологии REST (с использованием JSP). Система Sesame позволяет создавать базы семантических данных на основе данных различных источников, а также поддерживает разные модули логических выводов. Для работы с системой Sesame разработан пакет API – opendrdf-sesame 2.6.4. В БЗ SemDL хранятся три вида онтологий для аннотирования ЭР (подробнее об этом см. [7]), а также три онтологии предметных областей: ACM (Ассоциация по вычислительной технике), UDC (Универсальная десятичная классификация), NSS (Специальности диссертаций).

Для решения задачи формирования рекомендаций в SemDL использовался описанный выше метод. С целью проверки качества предложенного метода и алгоритма вычисления близости проведены вычислительные эксперименты. В качестве тестов использовались следующие данные: авторы (an); ключевые слова (kn); онтология предметной области (NSSn); онтология предметной области (ACMn).

Данные и результаты тестов представлены в таблице (*S* – шаблон одного документа, Dn – релевантные документы в системе, значение 1 соответствует наличию в этом документе данного, интересующего пользователя). Фрагменты таксономии NSS, ACM и график близости документов показаны на рис. 3.

Данные экспериментов

Data	a1	a2	a3	k1	k2	k3	k4	k5	k6	nss1	nss 2	nss3	nss4	acm1	acm2	Sim
S	1	1	1	1	-	-	-	-	-	-	-	-	-	1	-	1,00000
D1	1	1	1	-	-	-	-	-	1	-	-	-	-	-	1	0,89520
D2	-	-	1	-	-	1	-	-	-	-	-	-	-	-	1	0,57154
D3	1	-	1	-	1	-	-	-	-	1	-	-	-	-	-	0,54772
D4	-	1	1	-	-	-	-	1	1	-	1	-	-	-	-	0,50000
D5	-	-	1	-	-	-	1	-	-	-	-	1	-	-	-	0,40824
D6	-	-	1	-	-	-	-	-	-	-	-	-	1	-	-	0,40824

а *б*

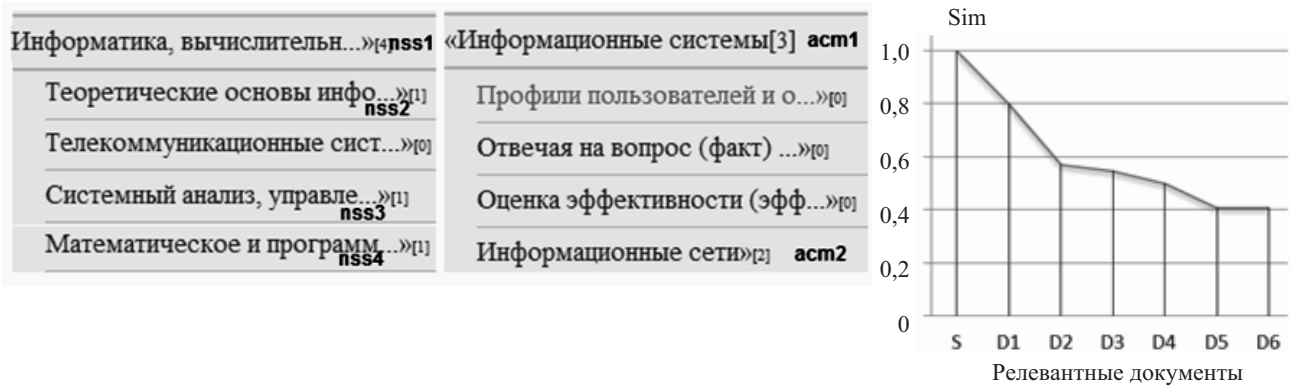

Рис. 3. Фрагменты таксономии (*а*) и график результата теста, упорядоченного по релевантности Dn (*б*)

Заключение. Результат эксперимента подтвердил эффективность предложенного метода и построенного на его основе алгоритма. В отфильтрованном документе *R* каждый документ имеет по крайней мере один объект триплета его метаописания (см. таблицу), сходный с объектом шаблона или семантически связанный с ним.

Список литературы

1. Ле Х. Х. Разработка электронных библиотек на основе семантических технологий // Науч.-техн. вестн. Поволжья. 2012. № 3. С. 138–145.

2. Черний А. В., Тузовский А. Ф. Развитие информационной системы организации с использованием семантических технологий // Материалы Всерос. конф. с междунар. участием "Знания – Онтологии – Теория", Новосибирск, 20–22 окт. 2009 г. Новосибирск: ЗАО "РИЦ Прайс-Курьер", 2009. Т. 2. С. 52−59.

3. Wu Z., Palmer M. Verb semantics and lexical selection // 32nd Annual meet. of the Assoc. for comput. linguist, 27–30 Jun. 1994. San Francisco: Morgan Kaufmann Publ., 1994. P. 133–138.

4. Semantic relatedness applied to all words sense dissambiguation. [Electron. resource]. http://www.scribd.com/doc/44296031/Jason-Thesis.

5. The SERQL query language (revision 3.1). [Electron. resource]. http://www.openrdf.org/doc/sesame2/users/ch09.html.

6. Home of Sesame. Б. м., 2012. [Electron. resource]. www.opendrdf.org.

7. Ле Х. Х., Тузовский А. Ф. Использование онтологии в электронных библиотеках // Изв. Том. политехн. ун-та. 2012. Т. 320, № 5. С. 36−42.

Ле Хоай – асп. Института кибернетики Томского политехнического университета; тел.: 8-9131-080-144; *e-mail: lehotomsk@yahoo.com;*

Тузовский Анатолий Федорович – д-р техн. наук, проф. Института кибернетики Томского политехнического университета; e-mail: tuzovskyaf@tpu.ru

Дата поступления – 22.09.12 г.