7. Хохлова, М.В. Экспериментальная проверка методов выделения коллокаций [Текст] / М. В. Хохлова // Инструментарий русистики: корпусные подходы. — Slavica Helsingiensia: 2008. — № 34. — С. 343–357.
8. Захаров, В. П. Анализ эффективности статистических методов выявления коллокаций в текстах на русском языке [Текст] / В.П. Захаров, М.В. Хохлова // Труды международной конференции «Диалог-2006». — 2006. — С. 137–143.
9. Ахманова, О. С. Словарь лингвистических терминов [Текст] / О. С. Ахманова. — 2-е изд. — М.: Советская энциклопедия, 1969 — 607 с.
10. Браславский, П. Сравнение пяти методов извлечения терминов произвольной длины [Текст] / П. Браславский, Е. Соколов. // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 4–8 июня 2008). — Вып. 7 (14). — М.: РГГУ, 2008. — С. 67–74.
11. О программе mystem [Электронный ресурс] / Режим доступа : \www/ URL: http://company.yandex.ru/technology/mystem/ — 10.06.2011 г. — Загл. с экрана.
12. Энциклопедический Словарь Конституционного Права [Текст] / под ред. Р. А. Мандрик — Новосибирск, 2010. — 666 с., 61145

В даній статті надана формальна модель семантичного пошуку в спеціалізованій електронній бібліотеці. Надана схема побудови онтології. Сформульовано леми для функцій інтерпретації термів і концепцій

Ключові слова: семантичний пошук, пошук інформації, онтології

В данной статье представлена формальная модель семантического поиска в специализированной электронной библиотеке. Представлена схема построения онтологии. Сформулированы леммы для функций интерпретации термов и концепций

Ключевые слова: семантический поиск, поиск информации, онтологии

This article presents a formal model of semantic search in a specialized electronic library. A scheme for constructing an ontology is presented. The lemmas for the functions of interpretation of terms and concepts are formed

Key words: semantic search, information retrieval, ontology

УДК 519.767.6

ФОРМАЛЬНАЯ МОДЕЛЬ СЕМАНТИЧЕСКОГО ПОИСКА В ЭЛЕКТРОННОЙ БИБЛИОТЕКЕ

З.В. Дударь
Кандидат технических наук, профессор, директор Центра
Центр последипломного образования*
Контактный тел.: (057) 702-18-05, 702-14-46
E-mail: fpo@kture.kharkov.ua

В.А. Белоконь
Аспирант**
Контактный тел.: (057) 702-18-05, 702-14-46
E-mail: fpo@kture.kharkov.ua

В.Г. Хильский
Магистрант
Контактный тел. (0625) 27-62-20, 063-243-84-33
E-mail: xv1975@mail.ru
**Кафедра программного обеспечения ЭВМ
*Харьковский национальный университет радиоэлектроники
пр. Ленина, 14, г. Харьков, Украина, 61166

Введение

Развитие индустрии систем электронного документооборота, сопровождающееся ростом массивов обрабатываемых полнотекстовых документов, требует новых средств организации доступа к информации, многие из которых следует отнести к разряду систем искусственного интеллекта - систем обработки знаний. Основной задачей, возникающей при работе с полнотекстовыми базами данных, является задача поиска документов по их содержанию. Однако, ставшие традиционными средства контекстного поиска по вхождению слов в документ, представленные, в частности, поисковыми машинами в интернет, зачастую не обеспечивают адекватного выбора информации по запросу пользователя.

Первые информационно-поисковые системы (ИПС) появились более тридцати лет назад и с тех произошли существенные изменения, как в поисковых алгоритмах, так и в техническом оснащении. В настоящее время в поисковых системах используется релевантная модель оценки соответствия исследуе-

мого документа поисковому запросу.Одно из перспективных направлений развития информационно-поисковых систем - построение моделей «семантического», т.е. «смыслового» поиска - поиска ресурсов, наиболее релевантных запросу, а не просто содержащие слова из запроса [1]. В 1999-2002 годах, как зарубежными, так и российскими учеными было предложено использовать в модели семантического поиска онтологии предметных областей [2-4]. Последние несколько лет в работах [5-9] рассматриваются различные методы для автоматического формирования онтологий, для чего используется лексический и синтаксический анализ документов.

Однако вопрос автоматического построения онтологий остается актуальным, так как релевантность полученных онтологий достаточно низкая.

Целью статьи является построение новой расширенной модели онтологии предметной области, в которой определены формальные функции интерпретации концепций и терминов; построение математической модели семантического поиска использующей расширенные функции интерпретации онтологии предметной области; разработка нового метода автоматического построения онтологии на основе информационных библиографических коллекций, распределенных в сети Интернет.

2. Формальные модели онтологий

Классическая модель онтологии [10] определяется как множество

$$O = <C, R, F>,$$

где C – конечное множество понятий предметной области;

R – конечное множество отношений между понятиями;

F – конечное множество функций интерпретации.

К заданным множествам предъявляются следующие требования:

C – непустое и конечное множество;

R и F – конечные множества.

Свойства онтологии:

1. если $R = 0$ и $F = 0$, то онтология трансформируется в простой словарь. Например, набор терминов, используемый в той или иной предметной области, без объяснений значений данных терминов. Простым словарем является любой орфографический словарь;
2. если $R = 0$ и $F \neq 0$, то онтология преобразуется в пассивный словарь (тезаурус). Например, толковый словарь — интерпретирование, уточнение, объяснение значения одних терминов на основе других, имеющихся в словаре;
3. если $R \neq 0$ и $F = 0$, то онтология является простой таксономией.

Таксономия - иерархически выстроенная система целей и результатов от простой к сложной системе. Математически таксономией является древообразная структура классификаций определенного набора объектов. Например, используемые в библиографии классификационные системы, которые задают отношения иерархии между понятиями. При этом не приводится интерпретации понятий.

Для решения конкретных задач в дальнейшем были введены и более сложные модели онтологий.

В работе [11] модель концептуализации предметной области определяется как множество

$$O = <U, R, F, L>,$$

где U – множество классов;

R – множество отношений;

F – множество функций;

L – множество констант.

Основное отличие от классической модели - множество понятий разделено на два разных множества: U (названное авторами классами) и L (константы). Это дало возможность уточнять понятия предметной области терминами из словаря, не являющимися сущностями, но семантически связанными с ними.

В работе [12] вводится ещё одна модель онтологии:

$$O = <L, C, F, H, Root>,$$

где L – словарь (набор терминов);

C – набор понятий (концепций);

F – функция интерпретации $F(L) \rightarrow C$. Отношение набора терминов к набору понятий, к которым они относятся;

H – таксономия. Концепции связаны направленным, нециклическим, рефлексивным отношением H;

Root – главная концепция.

Множество отношений R, на которое не было наложено никаких ограничений, заменяется на строго ограниченное множество H, допускающее только иерархические связи между концепциями. Это ограничение влечет за собой появление *параметра* Root, *который обеспечивает онтологическому* дереву хоть один «корень».

Данная модель может быть с успехом применена для классификации документов, но для задачи семантического поиска необходимо определение функции интерпретации концепций $F_c(C) \rightarrow L$, необходимой для операции расширения запроса семантически связанными терминами, определенными для данной концепции.

3. Модель онтологии, специализированной для задач полнотекстового поиска

Формально определим онтологию как кортеж

$$O = <L, C, Fi, Fc, Rh>,$$

где L – словарь терминов предметной области,

$L = \{(wi, xi)\}i=1,n$;

wi – термин, возможно более одного слова;

xi – вес термина в словаре;

C – набор понятий (концепций), $C= \{ci\}i=1,n$;

$Fi(L) \rightarrow C$ – функция интерпретации терминов, сопоставляет набору терминов из словаря подмножество концепций;

Fc – функция интерпретации концепций, $Fc(Ci) \rightarrow L$, сопоставляет концепции набор терминов из словаря;

Rh – отношения иерархии между концепциями.

4. Функция интерпретации терминов

В качестве функции интерпретации определим вероятностную функцию. Введем следующие обозначения:

u – поисковый запрос, состоящий из одного или нескольких слов;

$w \subset L$ –один термин из словаря.

Запрос и представим в виде множества терминов из словаря L, построенных на основе слов из запроса u:

$$u = \bigcup_m w_m.$$

Назовем априорной вероятностью вероятность события A – выбор концепции из множества C для запроса u.

Определим пространство гипотез. Событием B определим как термин $w \subset L$ присутствует в запросе u.

Применим формулу полной вероятности [57]

$$P(A) = \sum_{w \in u} P(A \mid B_i) P(B_i), \qquad (1)$$

где P(A | Bi) – вероятность того, что будет выбрана концепция ci, если термин wi входит в запрос u.

Введем разбиение априорного события A как выбор одной из концепций ci из множества C и, применив формулу Байеса, получим

$$P(A_i|B) = \frac{P(B|A_i) \cdot P(A_i)}{\sum_{j=1,m} P(B|A_j) \cdot P(A_j)}. \qquad (2)$$

В нашей модели онтологии свойства концепции (родительская и подчиненные концепции, количество терминов, относящихся к концепции, и другие) никак не влияют на вероятность ее выбора. Вероятность выбора концепции зависит только от терминов, следовательно, вероятности выбора концепций равны, т.е.

$$\sum_{j=1,m} P(B|A_j) \cdot P(A_j) = P(A_i) \cdot \sum_{j=1,m} P(B|A_j).$$

Из чего следует

$$P(A_i|B) = \frac{P(B|A_i)}{\sum_{j=1,m} P(B|A_j)}. \qquad (3)$$

Оценим $P(B|A_i)$ – вероятность того, что если будет выбрана концепция c_i, то в ней будут термины из запроса u. Эта величина известна из модели нашей онтологии и имеет значение x_w^i - вес данного термина в словаре.

Оценим вероятность $P(B_i)$, т.е. вероятность того, что термин w присутствует в запросе и: $P(B_i) = \frac{count(w|L)}{\sum_{w' \in n} count(w', L)}$ – отношение количества вхождений термина w к общей сумме вхождений всех терминов из запроса в словарь.

Итоговая формула выглядит следующим образом

$$P(A) = \sum_{w \in n} \left(\frac{x_w^i}{\sum_{j=1,m} x_w^j} \cdot \frac{count(w|L)}{\sum_{w' \in n} count(w', L)} \right) \qquad (4)$$

Функция интерпретации терминов принимает вид:

$$F_i(u) = \left\{ c_i \mid P(c_i \mid u) = \max_{c_j \in C} \sum_{w \in n} \left(\frac{x_w^i}{\sum_{j=1,m} x_w^j} \cdot \frac{count(w|L)}{\sum_{w' \in n} count(w', L)} \right) \right\}, i = \overline{1}. \qquad (5)$$

Определение. Назовем запрос и корректным, если существует хотя бы одно w, такое что $w \in u \cap w \in L$.

Лемма 1. Для любого корректного непустого запроса u множество Fi(u) не пусто, т.е. будет найдена хотя бы одна концепция, соответствующая запросу.

Доказательство: $u = \{w_i\}_{i=1}^n$, существует jє[1,n], где $w_j \in L$, следовательно, для wj выполняются следующие два условия: $\frac{x_w^j}{\sum_{c_k \in C} x_w^k} > 0$ и $\frac{count(w^j, L)}{\sum_{w \in u} count(w, L)} > 0$

И по определению функции интерпретации (5), есть хоть одно ci, для которого выполняется P(ci | u) > 0.

5. Функция интерпретации концепций

Определим функцию интерпретации как множество терминов, относящихся к данной концепции с весом большим, чем средний вес всех терминов для данной концепции. Функцию интерпретации концепций определим как

$$F_c(c_i) = \left\{ w_j \mid x_w^j \geq \frac{\sum_{w \in L_i} x_w}{\sum_{w \in L_i} 1}, j = \overline{1,k} \right\},$$

где Li – множество всех терминов из L, соответствующие концепции ci.

Лемма 2. Для любой концепции ciєC, множество Fc(ci) не пусто, т.е. найдется хотя бы один термин, уточняющий данную концепцию.

Доказательство: исходя из неравенства о средних, $Max(x_1, ...x_k) \geq \frac{x_1 + ...x_k}{k}$, из чего следует, что существует хотя бы одно xj, которое больше либо равно среднему арифметическому. Т.е. множество Fc(ci) состоит хотя бы из одного элемента.

6. Математическая модель поисковой системы

Существует два варианта обработки поискового запроса:

1. u = ci – поисковый запрос совпадает с названием какой-либо концепции в онтологииwi
2. $w_i \subseteq L, w_i \in u$ – поисковый запрос или его часть совпадает с подмножеством словаря онтологии.

В первом случае, расширяем поисковый запрос, применяя функцию интерпретации концепций, т.е. дополняя запрос терминами из найденной концепции $U = u \cup F_c(c_i)$.

Во втором случае, применяем функцию интерпретации терминов, получая множество наиболее релевантных концепций. К полученным концепциям применяем функцию интерпретации терминов, дополняя запрос терминами, уточняющими данную концепцию. Расширяем запрос, применяя функцию интерпретации $U = u \cup \left(\bigcup_i (F_c(F_i(u)) \cup c_i \right)$. В результате алгоритм

расширения запроса сводится к заданию наиболее релевантных прямой и обратных функций интерпретации.

Теорема. Если u – корректно, то U \ u ≠ 0, т.е. и дополняется не пустым множеством.

Доказательство: рассмотрим случай, когда u = Ci. По лемме 2 множество Fc(ci) не пусто и следовательно множество U \ u также не пусто.

Рассмотрим случай, когда $w_i \subseteq L, w_i \in u$. По лемме 1 и 2 множество Fc(Fi(u)) ≠ 0 и, следовательно, множество $U \setminus u \neq O$.

7. Математическая модель библиографических баз данных

Существует два варианта обработки поискового запроса:

1. u = ci – поисковый запрос совпадает с названием какой-либо концепции в онтологии
2. $w_i \subseteq L, w_i \in u$ – поисковый запрос или его часть совпадает с подмножеством словаря онтологии.

В первом случае, расширяем поисковый запрос, применяя функцию интерпретации концепций, т.е. дополняя запрос терминами из найденной концепции $U = u \cup F_c(c_i)$.

Во втором случае, применяем функцию интерпретации терминов, получая множество наиболее релевантных концепций. К полученным концепциям применяем функцию интерпретации терминов, дополняя запрос терминами, уточняющими данную концепцию. Расширяем запрос, применяя функцию интерпретации $U = u \cup \left(\bigcup_i (F_c(F_i(u)) \cup c_i \right)$. В результате алгоритм

расширения запроса сводится к заданию наиболее релевантных прямой и обратных функций интерпретации.

Теорема. Если u – корректно, то U \ u ≠ 0, т.е. и дополняется не пустым множеством.

Доказательство: рассмотрим случай, когда $u = C_i$. По лемме 2 множество $F_c(c_i)$ не пусто и следовательно множество U \ u также не пусто.

Рассмотрим случай, когда $w_i \subseteq L$, $w_i \in u$. По лемме 1 и 2 множество $F_c(F_i(u)) \neq 0$ и, следовательно, множество $U \setminus u \neq O$.

8. Метод построения онтологии

Для преобразования кортежа G в кортеж O (онтологию), нам необходимо построить отображение R_c:

C → C, где $L=\{(w_i, x_i)\}_{i=1,n}$.

Определим отношение R_{bc}, выбрав множество библиографических записей, соответствующих конкретной концепции:

$$R_{bA}(b_{(i,m)}, c_i) = \bigcup_{n=1}^{N} w^n_{(i,m)}$$

.

Данное отношение означает, что для каждой библиографической записи и отнесенной к ней концепции существует свой набор терминов. Свернув множество отношений R_{bc} по всем библиографическим записям, получим

$$R_w(c_i) = \bigcup_{m=1}^{M} R_{bc}(b_{(i,m)}, c_i) = \bigcup_{k=1}^{K} w^i_k$$

Так как термины в разных записях могут повторяться, то введем коэффициент повторения

$$x^i_k = \underset{m=1,M}{count}(b_{(i,m)} | w^k_{(i,m)} \in b_{(i,m)})$$

Чем больше экспертов определили данный термин для соответствующего кода УДК, тем выше его вес x^i.

Итак, мы получили отображение

$$R_A(c_i) = \left\{ (w^i_k, x^i_k) \right\}_{k=1,K_i}$$

То есть R_c: C ® L, что соответствует функции интерпретации концепций в нашей модели онтологии. Сам метод можно представить в виде схемы на рис. 1.

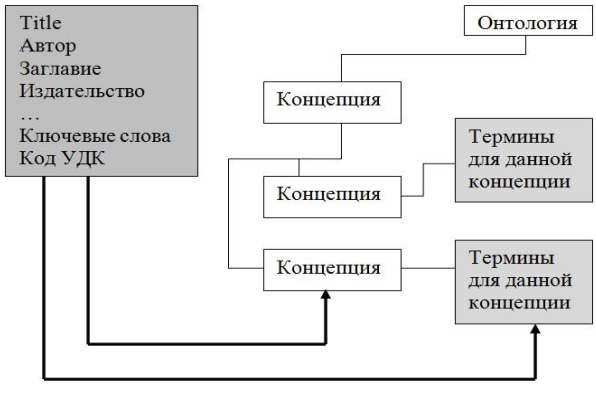

Рис. 1. Схема функции интерпретации концепций в модели онтологии

Выводы

В данной статье была рассмотрена формальная модель онтологии специализированной для полнотекстового поиска. На основе проведенных исследований были получены леммы для функций интерпретации термов и концепций.

К основным результатам следует отнести следующее:

разработана математическая модель семантического поиска, использующая онтологию предметной области, доказано существование непустого решения – семантической интерпретации запроса пользователя к ИПС;

разработана математическая модель онтологии, ориентированной на задачи информационного поиска, определены и математически обоснованы формальные функции интерпретации концепций и терминов;

предложен метод для автоматического создания онтологии на основе распределенных информационных библиографических коллекций, имеющихся в сети Интернет.

Литература

1. Ushold M. Ontologies: Principles, Methods and Applications [Текст] / M. Ushold, M. Gruninger // Knowledge Engineering Red view. – 1996. – V. 11, № 2. – P. 115-121.
2. Heflin J. Applying Ontology to the Web: A Case Study [Текст] / J. Heflin, J. Hendler, S. Luke // In Proc. IWANN. – 1999. – № 2. – P. 715-724.
3. Лукашевич Н. Тезаурус русского языка для автоматической обработки больших текстовых коллекций [Текст] / Н. Лукашевич, Б. Добров // Компьютерная лингвистика и интеллектуальные технологии: Труды Международного семинара «Диалог 2002». – 2002. – Т. 2. – С. 338-346.
4. Гаврилова Т. А. Базы знаний интеллектуальных систем. Учебник для вузов [Текст] / Т.А. Гаврилова, В.Ф. Хорошевский. – СПб.: Питер, 2000. – 384 с.
5. Jones K. A Statistical Interpretation of Term Specificity and Its Application in Retrieval [Текст] / K. Jones // Journal of Documentation. – 1972. – V. 28. – P. 11-21.
6. Браславский П. Сравнение пяти методов извлечения терминов произвольной длины [Текст] / П. Браславский, Е. Соколов // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной международной конференции «Диалог». – 2008. – № 7 (14). – С. 67-75.
7. Ермаков А. Автоматизация онтологического инжиниринга в системах извлечения знаний из текста [Текст] / А. Ермаков // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной международной конференции «Диалог». – 2008. – № 7 (14). – С. 154-159.
8. Лукашевич Н. Отбор словосочетаний для словаря системы автоматической обработки текстов [Текст] / Н. Лукашевич, Б. Добров, Д. Чуйко // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной международной конференции «Диалог». – 2008. – № 7 (14). – С. 339-345.
9. Сидорова Е. Подход к извлечению фактов из текста на основе онтологии [Текст] / Е. Сидорова, И. Кононенко // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог 2009». – 2009. – № 8 (15). – С. 451-458.
10. Gruber T. Ontolingua: A Mechanism to Support Portable Ontologies [Текст] / T. Gruber // Technical Report KSL-91-66 Stanford, Stanford University, Knowledge Systems Laboratory. – 1992. – P. 61-69.
11. Weiss S. Model-Based Method for Computer-Aided Medical Decision Making [Текст] / S. Weiss, C. Kulikovski, S. Amarel, A. Safir // Reading in Medical Artificial Intelligence. – 1984, the First Decade. – P. 160-189.
12. Hotho A. Ontology-based Text Clustering [Текст] / A. Hotho, A. Maedche, S. Staab // In: Proc. of the Workshop «Text Learning: Beyond Supervision» at IJCAI 2001. – 2001, August 6. – P. 225-230.
13. Боровков А. А. Теория вероятностей [Текст] / А. А. Боровков. – М.: Эдиториал УРСС, 1999. – 472 с.