Уфа : УГАТУ, 2009 Вестник УГАТУ • Управление, ВТ и И Т.12, № 1(30). С. 133–138

УДК 004.9:025.4

И. В. ЗАХАРОВА

ОБ ОДНОМ ПОДХОДЕ К РЕАЛИЗАЦИИ СЕМАНТИЧЕСКОГО ПОИСКА ДОКУМЕНТОВ В ЭЛЕКТРОННЫХ БИБЛИОТЕКАХ

В статье описаны методы и алгоритмы семантического полнотекстового поиска документов в электронных коллекциях. *Семантический поиск ; онтологии ; модели информационного поиска ; протокол Z35.50*

Основной задачей, возникающей при работе с распределенными полнотекстовыми коллекциями документов, 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 в поисковых алгоритмах. Современные поисковые системы (ПС) «научились» автоматически собирать информацию в Интернете, учитывать морфологические особенности и производить оценку значимости [1] найденных документов. В настоящее время в поисковых системах используется релевантная модель оценки соответствия исследуемого документа поисковому запросу. Данная модель практически не справляется с решением задач распознавания омонимов и многозначных слов.

Одним из возможных вариантов решения этой проблемы является семантический поиск. Семантический поиск позволяет находить ресурсы не только по заданным словам из запроса, но и по эквивалентным по смыслу.

Например, по запросу «маршрутизация в глобальных сетях» для полноты поиска его можно расширить следующими терминами: компьютерные сети, узел сети, модель сети, INTERNET, распределённые сетевые ресурсы, сервер DNS, маршрутизатор, протокол, межсетевое взаимодействие, CISCO, коммутатор.

Для эффективного семантического поиска необходима информация о предметной области, свойственных ей понятиях и отношениях между ними, а также ограничениях, свойственным этим отношениям. Такую информацию принято называть онтологией [2], и онтологическая модель может быть использована как для полнотекстового поиска, так и для классификации документов.

Онтология определяет термины, используемые для описания и представления знаний той или иной предметной области. Она необходима для людей, для приложений систем баз данных и различных других информационных систем, которые совместно используют специфическую информацию в какой-либо предметной области. Онтологии включают доступные для компьютерной обработки определения основных понятий предметной области и связи между ними [3].

МОДЕЛЬ ОНТОЛОГИИ, СПЕЦИАЛИЗИРОВАННАЯ ДЛЯ ЗАДАЧ СЕМАНТИЧЕСКОГО ПОИСКА И КЛАССИФИКАЦИИ

Формально определим онтологию как множество

$$O = (L, C, F_l, F_c, R_h),$$

где $L = \{(w_i, x_i)\}_{i=1,n}$ – словарь терминов предметной области, w_i – термин, возможно более одного слова; x_i – его рейтинг относительно других терминов в концепции.

Контактная информация: iren@csu.ru

C, $C = \{c_i\}_{i=1,m}$ – набор понятий (концепций), $F_l(L) \to C$ – функция интерпретации терминов сопоставляет набору терминов из словаря подмножество концепций; $F_c(C_i) \to L$ – функция интерпретации концепций; сопоставляет концепции набор терминов из словаря; R_h – отношения иерархии между концепциями [5].

В качестве функции интерпретации терминов возьмем $P(c_i \mid u)$ – вероятность выбора концепции при условии запроса *u*.

Применив формулы полной вероятности и формулы Байеса[4], получим

$$F_l(u) = \left\{ \begin{array}{l} c_i \mid P(c_i \mid u) = \\ = \max\limits_{c_j \in C} \left(\sum\limits_{w \in u} \left(\dfrac{x_w^j}{\sum\limits_{c_k \in C} x_w^k} \dfrac{\text{count}(w, L)}{\sum\limits_{w' \in u} \text{count}(w', L)} \right) \right) \end{array} \right\},$$

$$i = \overline{1,n}$$

Определим обратную 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

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

где $L_i = \bigcup\limits_j w_j^i$ – множество всех терминов, соответствующие концепции.

Формализуем модель поисковой системы с использованием полученной онтологии

Два варианта обработки поискового запроса:

1) $u = C_i$ – поисковый запрос совпадает с названием какой-либо концепции в онтологии.

Расширяем поисковый запрос, применяя функцию интерпретации концепций, т. е. дополняя запрос терминами из найденной концепции

$$U = u \bigcup F_c(c_i);$$

2) $w_i \subseteq L$, $w_i \in u$ – поисковый запрос или его часть совпадает с подмножеством словаря онтологии, применяем функцию интерпретации терминов к запросу *u*, получая множество наиболее релевантных концепций. К полученным концепциям применяем обратную функцию интерпретации, дополняя запрос терминами.

Расширяем запрос, применяя функцию интерпретации

$$U = u \bigcup \left(\bigcup_i \left(F_c(F_l(u)) \bigcup c_i \right) \right).$$

МЕТОД ПОСТРОЕНИЯ ОНТОЛОГИИ

Для реализации эффективного семантического поиска необходима онтология, которая по сути описывает не одну какую-либо предметную область, а классифицирует все виды сущностей и связи между ними. Создание подобной системы возможно как минимум двумя путями.

Специалисты в некоторой предметной области создают для собственных целей онтологию. Объединяя эти предметно-ориентированные онтологии и добавляя, возможно, при этом дополнительные связи, получаем «обобщенную онтологию». Метод, очевидно, долгий и требующий работы множества экспертов по многим предметным областям.

Второй способ – построить онтологию автоматически, используя для этого имеющиеся коллекции информационных ресурсов и библиографических баз данных, представленных в Интернет.

В 1962 г. в стране в качестве единой обязательной классификации принята Универсальная десятичная классификация (УДК), и введено обязательное индексирование всех публикаций, т. е. все информационные материалы в области естественных и технических наук издаются с индексами Универсальной десятичной классификации.

Пример дерева УДК для «ветки» 004.8 приведен на рис. 1.

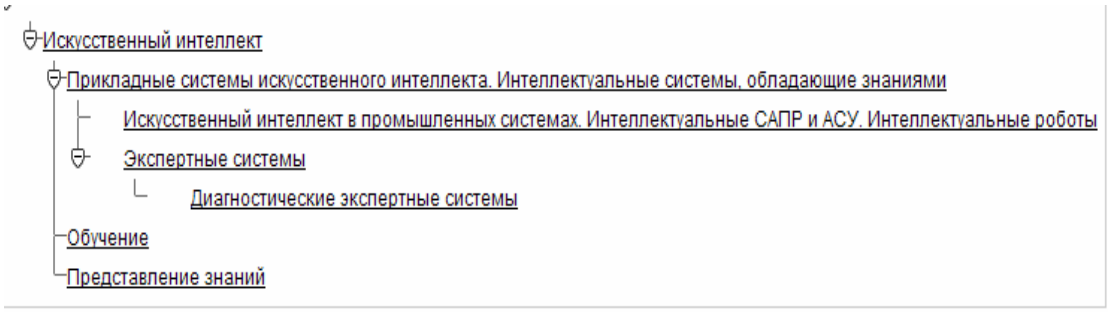

Рис. 1

В результате, мы имеем экспертную базу на многих языках, где для каждого классификационного кода определено подмножество различных публикаций, содержащих знания по данной теме. Наша задача – выделить эти знания и представить их в виде набора терминов, наиболее характерных для данной рубрики [6].

Но помимо электронных публикаций, в Интернете (и в каждой электронной библиотеке) присутствуют также так называемые библиографические базы данных, создаваемые в различных библиотеках мира.

Рассмотрим библиографическую запись об одной книге:

Ирбенек В. С. Алгоритмы проектирования топологии электрических соединений в САПР электронной аппаратуры// Зарубежная радиоэлектроника. Успехи современной радиоэлектроники.–2002.–N 7. - С. 71-79

Ключевые слова

автоматизация; автоматизированное проектирование; алгоритмы; деревья Краскала-Прима; деревья Штейнера; ортогональная метрика; проектирование автоматизированное; САПР; электроника; электронная аппаратура.

Код УДК

004.896

Все множество библиографических баз данных можно рассматривать как кортеж

$$G =< C, W, R_c, B, R_b >,$$

где $C = \{c_i\}_{i=1,m}$ – набор понятий (концепций УДК)

$W = \{w_i\}_{i=1,n}$ – множество ключевых терминов

$B = \{b_m\}_{m=1,M}$, $R_c : C \to C$ – множество библиографических записей,

R_b – отношения иерархии между концепциями

$R_b : B \to C \times W$ – отношения множества библиографических записей на множество концепций и терминов

Сопоставляет каждой библиографической записи присвоенную ей концепцию и набор терминов, определяемых экспертом.

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

N – количество ключевых терминов, заданных экспертом для этой книги.

Для преобразования кортежа G в кортеж O (онтологию), нам необходимо построить отображение $R_c : C \to L$, где $L = \{(w_i, x_i)\}_{i=1,n}$

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

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

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

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

Поскольку термины в разных записях могут повторятся, то вводим коэффициент повторения

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

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

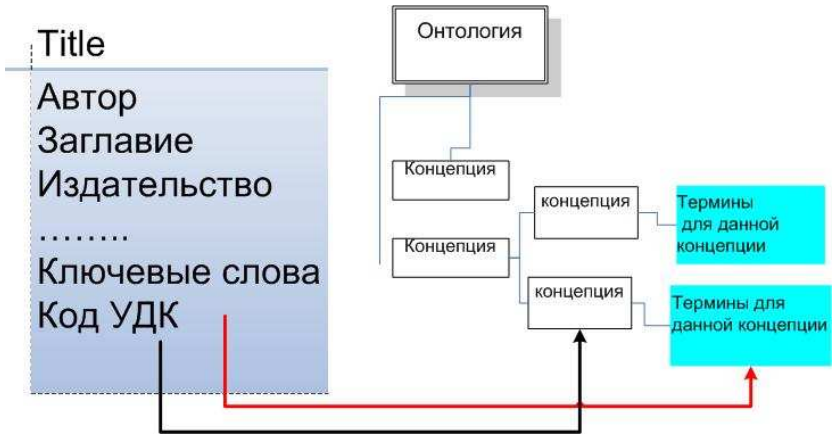

Рис. 2

Рис. 3

Итак, мы получили отображение

$$R_c(c_i) = \{(w_k^i, x_k^i)\}_{k=1,K_i}$$

т. е. $R_c : C \to L$, что соответствует функции интерпретации концепций в нашей модели онтологии.

Сам метод можно представить в виде схемы (рис. 2).

В настоящее время в России в библиотечном сообществе широко распространена идея создания сводных каталогов, объединяющих отдельные библиотечные каталоги участников либо в единый физический каталог (путем копирования данных на один сервер), либо в распределенный каталог (поиск и работа с которым осуществляется распределенно). Управление доступом к распределенным информационным ресурсам и взаимодействие электронных библиотек осуществляется по принципу распределенных систем на базе открытых стандартов обмена данными. Для реализации электронных библиотек используются в основном два протокола: Z39.50 и HTTP. В качестве подсистемы построения онтологии был выбран протокол Z39.50, изначально ориентированный на информационно-поисковые задачи именно в библиографических базах данных [8, 9].

Общая архитектура приведена на рис. 3.

С помощью программы были просканированы сводные и распределенные каталоги Ассоциации Региональных Библиотечных Консорциумов (АРБИКОН) и выделено 133 151 концепции, содержащие от 5 до 100 терминов для каждой концепции.

АЛГОРИТМ СЕМАНТИЧЕСКОГО ПОИСКА С ИСПОЛЬЗОВАНИЕМ ОНТОЛОГИИ

```
PROCEDURE
Семантический_поиск(запрос_исходный)

Запрос = STEM(запрос_исходный)
// обрезание окончаний и суффиксов
Номер_концепции =
Поиск_по_концепции(Запрос)

IF Номер_концепции <> 0  Then
    // нашлась концепция по названию
    Обход_по_дереву(Номер_концепции)
ELSE
// запрос не совпал не с одной концепцией

 Список_концепций =
Функция_интерпретации(Запрос)
 Номер_концепции =
концепция(Список_концепций)

 While Номер_концепции <> 0 LOOP
        Обход_по_дереву(Номер_концепции)
        Номер_концепции =
Следующая_концепция(Список_концепций)
  End LOOP

END IF
END PROCEDURE
```

Поскольку в основе онтологии лежит древовидная структура концепций, то определив наиболее релевантную концепцию получаем также список подчиненных.

Пользователю предоставляется возможность уточнить свой запрос, указав наиболее релевантную с его точки зрения концепцию из предложенных.

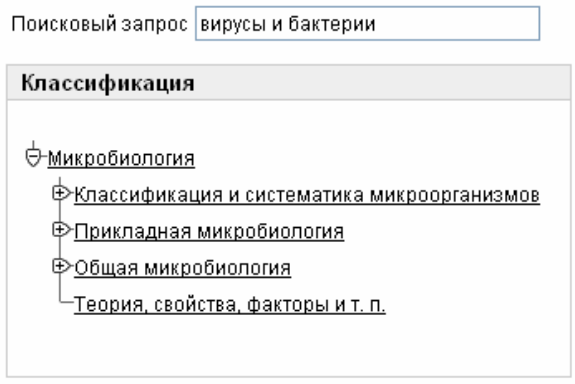

Рис. 4

Пример. Запрос «вирусы и бактерии» был расширен следующими терминами: микробиология, микроорганизмы, бактериальные клетки, микроб, возбудители болезней, инфекция, иммунология, вирусология и др.

ОЦЕНКА ЭФФЕКТИВНОСТИ ПОИСКА

Введем следующие обозначения

N – общее число документов в коллекции, релевантных запросу

R – количество релевантных документов из найденных по запросу

R' – количество нерелевантных документов из найденных по запросу

Для оценки эффективности работы ИПС мы использовали следующие параметры:

$r(u)$ – точность выдачи

$r(u) = \frac{R}{R+R'}$ – отношение числа выданных релевантных документов к сумме числа выданных релевантных и числа выданных нерелевантных документов.

$P(u)$ – полнота выдачи

$P(u) = \frac{R}{N}$ – отношение числа выданных релевантных документов к сумме числа выданных релевантных и числа не выданных релевантных документов.

Чтобы избежать сравнения пар полнота, точность используются однозначные оценки. Одной из таких оценок является E-мера, позволяющая избежать сравнения пар полнота, точность за счет введения отношения их значимости.

$$E(u) = \frac{1+b^2}{\frac{b^2}{r(u)} + \frac{1}{P(u)}},$$

где b – отношение значимости полноты и точности. Возьмем $b = 1$ – среднее гармоническое (компромисс между точностью и полнотой)

Оценка эффективности производилась в сравнении с ИПС Indexing Services, использующим обычную булевскую модель поиска:

- среднее значение $E(u)$ для SemanticSearch = 0.41
- среднее значение $E(u)$ для Windows IS = 0.34 [7]

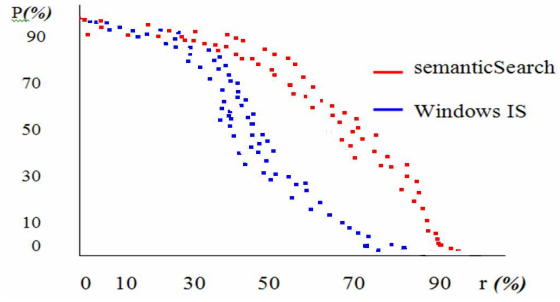

Рис. 5

Как видно из приведенных результатов, средний показатель эффективности работы ИПС с учетом предпочтения пользователя (коэффициент *b*) для ИПС с семантическим поиском выше, несмотря на то, что некоторые запросы были с большим «шумом», т. е. большим количеством нерелевантных документов. При увеличении коэффициента *b* (т.е. предпочтение отдается полноте) мы соответственно получим и большее значение E-меры для семантического поиска, поскольку сам алгоритм разработан для поиска максимального количества релевантных документов.

СПИСОК ЛИТЕРАТУРЫ

1. **Baeza-Yates, R.** Modern Information Retrieval / R. Baeza-Yates, B. Ribeiro-Neto. ACM Press, 1999.

2. **Gruber, T. R.** A translation approach to portable ontology specifications / T. R. Gruber // Knowledge Acquisition. 1993. № 5(2).

3. **Овдей, О. М.** Обзор инструментов инженерии онтологий / О. М. Овдей, Г. Ю. Проскудина // Электронные библиотеки. 2004. Т.7, Вып. 4.

4. **Гнеденко, Б. В.** Курс теории вероятностей / Б. В. Гнеденко. М. : Наука, 1988.

5. **Zakharova, I. V.** An approach to automated ontology building in text analysis problems / I. V. Zakharova, A. V. Melnikov, J. A. Vokhmitsev // Workshop on Computer Science and Information Technologies, 2006. P. 177–178.

6. **Melnikov, A. V.** Method of automatic ontology creation based on bibliographic databases / A. V. Melnikov, I. V. Zakharova //Workshop on Computer Science and Information Technologies, 2005. P. 270–272.

7. **Zakharova, I. V.** Evaluating effectiveness of semantic search / I. V. Zakharova, A. V. Melnikov, M. S. Timchenko // Workshop on Computer Science and Information Technologies, 2008.

8. **Глухов, В. А.** Электронные библиотеки. Организация, технология и средства доступа / В. А. Глухов, О.Л. Голицына, Н. В. Максимов // НТИ. Сер. 1, 2000. № 10.

9. **Жижимов, О. Л.** Введение в Z39.50 / О. Л. Жижимов. Новосибирск : НГОНБ, 2000.

ОБ АВТОРЕ

Захарова Ирина Викторовна, ст. преп. каф. выч. механики и инф. техн. ЧелГУ. Дипл. математик (ЧелГУ, 1992). Иссл. в обл. разработки современных средств доступа к ресурсам электронных библиотек.