Об одном подходе к автоматическому построению онтологии для задач анализа текстов

An approach to automated ontology building in text analysis problems

Захарова И. В. (iren@csu.ru),
Городечный П. П. (petr.gorodechnyj@edu.csu.ru)

Челябинский государственный университет, математический факультет

В статье описан метод автоматического построения онтологии для сложных задач классификации, аннотирования и поиска текстовых документов.

1. 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. Одним из эффективных подходов к выявлению и обработке смысла текстовых документов является использование онтологий.

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[2].

2. 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.

$$C = \{c_i\}_{i=1,m}$$

C — набор понятий (концепций),

$F_l(L) -> C$ — Функция интерпретации терминов
Сопоставляет набору терминов из словаря подмножество концепций.

$F_c(C_i) -> L$ — Функция интерпретации концепций;
сопоставляет концепции набор терминов из словаря.

R_h — Отношения иерархии между концепциями [4].

$$P(c_i \mid u)$$

В качестве функции интерпретации терминов возьмем — вероятность выбора концепции при условии запроса *u*.

Применив формулы полной вероятности и формулы Байеса[3], получим

$$F_l(u) = \left\{ c_i \mid P(c_i \mid u) = \right.$$

$$\left. = \max_{c_j \in C} \left(\sum_{w \in u} \left(\frac{x_w^j}{\sum_{c_k \in C} x_w^k} \bullet \frac{count(w, L)}{\sum_{w' \in u} count(w', L)} \right) \right) \right\},$$

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

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

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

$L_i = \bigcup_j w_j^i$ — множество всех терминов, соответствующие концепции c_i .

3. Метод построения онтологии

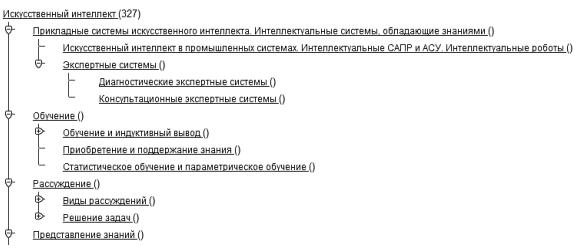
Для реализации эффективного семантического поиска необходима онтология, которая, по сути, описывает не одну какую-либо предметную область, а классифицирует все виды сущностей и связи между ними. Создание подобной системы возможно как минимум двумя путями.

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.



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[5].

Рассмотрим библиографическую запись об одной книге:

Ирбенек В. С. Алгоритмы проектирования топологии электрических соединений в САПР электронной аппаратуры // Зарубежная радиоэлектроника. Успехи современной радиоэлектроники. — 2002. — № 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

Сам метод выделения терминов из ББД можно представить в виде схемы

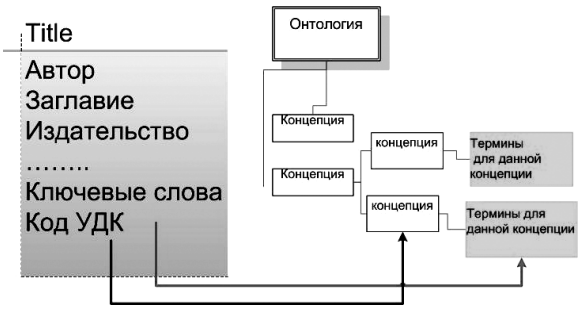


Рис. 1

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 [6,7].

Общая архитектура приведена на рисунке 2.

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.

Рис. 2

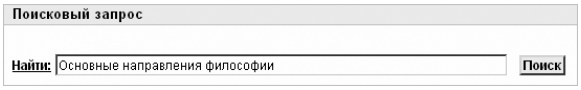
4. Применение онтологии

Полученную онтологию предполагается использовать в аналитической системе BIOAP (Basic Integrated Ontological Analytic Processor) 1.0 для:

- Классификация/рубрицирования (определения типа документа)
- Реферирования/аннотирования (извлечения краткого содержания из текста)
- Семантического поиска по коллекции документов

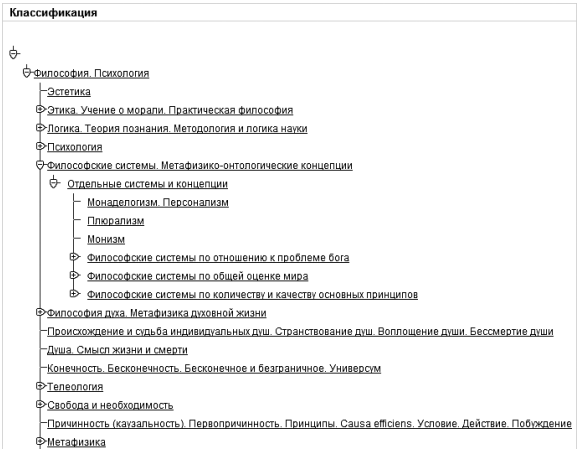
На данный момент реализованы алгоритмы семантического поиска.

Ввод запроса пользователем осуществляется как в любой поисковой системе:



Дальнейшая работа системы осуществляется по следующей схеме:

1. Запрос делится на термины
2. К терминам применяется функция, выделяющая корень слова (стемминг)
3. Для каждого термина рассчитывается его вес в контексте онтологии
4. Применяем функцию интерпретации терминов к запросу.
5. Получаем список концепций, наиболее релевантных запросу по смыслу. Например, для указанного выше запроса получим концепцию «*Философия. Психология*».
6. Для каждой концепции выполняем поиск подчиненных концепций и выводим их на экран в виде дерева. Например, для указанного запроса будет получено следующее дерево:



7. Пользователь уточняет запрос, указывая конкретно, какая тема его интересует.

В данном случае выбирается «Философские системы. Метафизико-онтологические концепции» и далее «отдельные системы и концепции», потом — «монизм»

8. Исходный запрос дополняется терминами из онтологии, семантически связанными с этой концепцией.

В данном случае запрос был дополнен следующими терминами: *Гуманизация образования, Тейяр Де Шарден, одаренность, формирование личности, всеединство, преджизнь, ноосфера*.

9. Расширенный запрос передается поисковой системе «Yandex Standard»
10. На экран выводится список найденных документов.

Например для указанной концепции «отдельные системы и концепции. Монизм» был получен следующий список документов :

1. Алешин А. И. Русская философия: Малый энциклопедический словарь.
2. Малахов В. С., Филатов В. П. Современная западная философия: Словарь.
3. Канке В. А. Основные философские направления и концепции науки. Итоги XX столетия.
4. Реале, Джованни. Западная философия от истоков до наших дней.
5. Суханов К. Н. Динамика тематической направленности философствования в XIX–XX столетиях.
6. Люсый А. К двумирной полноте — через терминологический монизм
7. Бородин Е. Т. Монизм и плюрализм в современной общественной науке.

5. Заключение

Предполагается дальнейшее использование онтологии для решения задач классификации и реферирования больших коллекций электронных полнотекстовых документов.

Литература

1. *Baeza-Yates R., Ribeiro-Neto B.* Modern Information Retrieval. ACM Press, 1999.
2. *Gruber T. R.* A translation approach to portable ontology specifications. Knowledge Acquisition, 5(2), 1993.
3. *Гнеденко Б. В.* Курс теории вероятностей. — М.: Наука, 1988.
4. *Zakharova I. V., Melnikov A. V., Vokhmitsev J. A.* «An approach to automated ontology building in text analysis problems». // Workshop on computer Science and Information Technologies CSIT'2006, Karlsruhe, Germany, 2006. P. 177–178.
5. *Melnikov A. V., Zakharova I. V.* «Method of automatic ontology creation based on bibliographic databases». // Workshop on computer Science and Information Technologies CSIT'2005, Ufa, Russia, 2005. P. 270–272.
6. *Глухов В. А., Голицына О. Л., Максимов Н. В.* Электронные библиотеки. Организация, технология и средства доступа // НТИ. — Сер. 1, — 2000, — №10.
7. *Жижимов О. Л.* Введение в Z39.50. — Новосибирск: Изд-во НГОНБ, 2000.