Знания-Онтологии-Теории (ЗОНТ-09)

Способы автоматического построения онтологии для задач анализа текстов.

Захарова И.В. Тимченко М.С.

Челябинский Государственный университет, математический факультет.

iren@csu.ru , admike@csu.ru

Аннотация. В статье описаны методы автоматического построения онтологии для сложных задач классификации, аннотирования и поиска текстовых документов.

Ключевые слова: выявление знаний, онтология, естественные языки.

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[1].

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

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

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

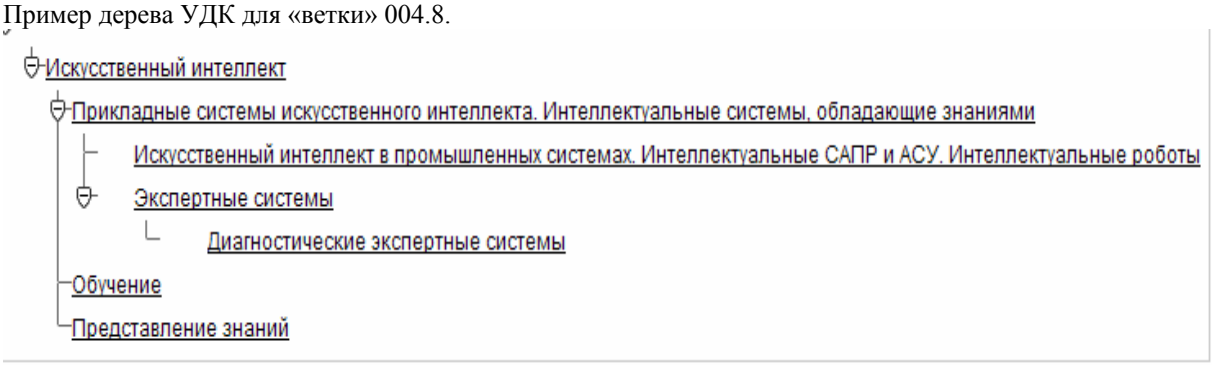
R_h - Отношения иерархии между концепциями [4].

3. Построение онтологии на основе УДК и библиографических баз данных.

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. Другой 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 [6].

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.–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

Сам метод выделения терминов из ББД можно представить в виде схемы

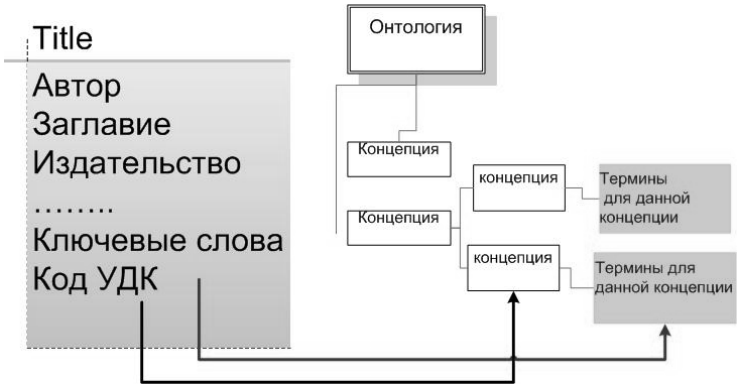


Рис. 1

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.

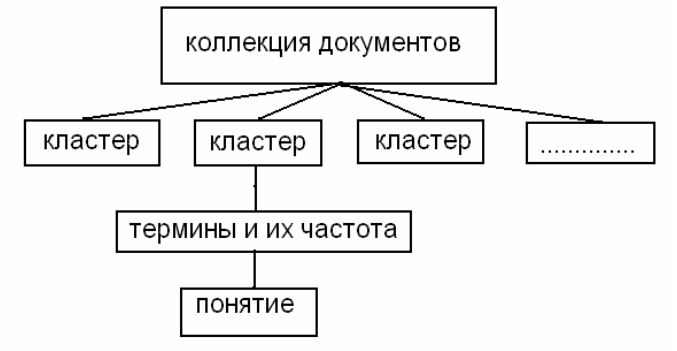
Проблемы, выявленные при анализе полученной онтологии:

- Жесткая и редко дополняемая структура УДК не позволяет отобразить информацию о современных разработках
- Во многих библиографических базах данных не используются коды УДК, в связи с этим они не учитываются при формировании онтологии
- Проблемы выделения ключевых слов в документах при заполнении библиографических записях (библиограф не является экспертом предметной области).

4. Построение онтологии с использование кластеризации.

Для решения этих проблем предлагается формировать онтологию предметной области на коллекции полнотекстовых документов с использованием кластеризации, что позволяет исправить недостатки, описанного выше метода.

Основная идея заключается в следующем: каждый документ представляется виде набора терминов, множество документов разбивается на подмножества документов близкой тематики (кластеры), в результате получаются группы терминов одной тематики. Это позволяет установить отношения между терминами и концепциями. Каждый термин характеризуется частотой встречаемости (весом). Термины с весом, больше среднего, задают термины онтологии, а в качестве концепции берутся термины с максимальным весом.



Кластеризация документов производится на основе иерархического алгоритма «минимальное покрывающее дерево»[8], позволяющего задать таксономию концепций в полученной онтологии. Коллекция документов представляется в виде дерева, разделение кластеров происходит в месте максимального расстояния между ближайшими документами внутри каждого кластера.

Для определения «близости» документов используется метрика Евклида, вычисляемая на основе частотных характеристик терминов, входящих в документ. Для оптимизации расчетов в качестве терминов берутся не слова, а устойчивые словосочетания, выделенные из текстов документов статистическим методом k-factor[7]. Изначально термины однословные, затем формируются многословные термины по следующему правилу: если более короткий термин-кандидат встречается лишь немногим чаще, чем более длинный термин-кандидат, в который он полностью входит, то «основным» считается более длинный вариант. Отбором управляет пороговое значение отношения частот терминов k.

Данный метод был протестирован на коллекции 4736 документов, состоящей из файлов формата txt, doc, pdf, html. Полученная онтология содержит 728 понятий и 51293 терминов. Для расширения онтологии необходимо использовать большее количество документов.

Проблемы, выявленные при анализе полученной онтологии:

- Наличие концепций, с небольшим количеством терминов – в кластере оказалось мало документов.
- Нерелевантные термины в некоторых концепциях – часть полученных терминов не являлись устойчивыми словосочетаниями.
- Существуют концепции с названиями, не соответствующими терминам – большая частота у терминов, не относящихся к тематике кластера.

5. Заключение.

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)
- Информационного поиска по коллекции документов

На данный момент реализованы алгоритмы семантического поиска с использованием полученных онтологий.

Список литературы.

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

2. T.R. Gruber. A translation approach to portable ontology specifications. Knowledge Acquisition, 5(2), 1993.

3. Соловьев В.Д., Добров Б.В., Иванов В.В., Лукашевич Н.В. Онтологии и тезаурусы. - Казань, Москва: Казанский государственный университет, МГУ им. М.В. Ломоносова, 2006. - 157 с.

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. Браславский П.И., Соколов Е.А. Сравнение пяти методов извлечения терминов произвольной длины// Компьютерная лингвистика и интеллектуальные технологии: труды Международной конференции Диалог’2008 – М.: Наука, 2008. – с. 67-75.

8. Мандель И.Д. Кластерный анализ. - М.: Финансы и статистика, 1988. - 176с