Онтологии в решении задач компьютерной лингвистики

## Подходы к представлению знаний о предметной области

Одним из основных понятий инженерии знаний является понятие «представление знаний». Под этим термином может пониматься либо способ кодирования знаний в базе знаний, либо формальная система, которая используется для формализации знаний. Практика разработки систем, основанных на знаниях, что в каждой предметной области существует некоторая структура, занимающая промежуточное положение между представлением знаний, используемым в модели предметной области, и моделью предметной области (базой знаний), т.е. между структурой, определяющей, что может быть представлено, и тем, что существует в действительности.

Управление знаниями (Knowledge Management) - совокупность процессов и технологий, предназначенных для выявления, создания, распространения, обработки, хранения и предоставления для использования знаний.

Знания - совокупность сведений, отчетов, фактов, понятий, представлений о чем-либо, накопленных в результате обучения, опыта, в процессе деятельности.

## Определение онтологии

Одним из наиболее известных является определение онтологии как спецификации концептуализации. Данное определение обладает еще одной особенностью – это наиболее общее определение онтологии. С одной стороны, благодаря этому оно не противоречит ни одному из частных определений. С другой стороны, данное определение не учитывает специфику задачи, при решении которой используется онтология.

Другой подход заключается в определении онтологии через ее составляющие. Например: онтология – это словарь понятий предметной области и совокупность явным образом выраженных предположений относительно смысла этих понятий. Подобные определения акцентируются на внутреннем содержании онтологий, которое не является их определяющим признаком.

К определению онтологии существует еще подход на основе понятия информации. Согласно этому подходу, онтология есть явное описание смысла терминов, неявно определенных концептуализацией.

Далее в работе будем придерживаться определения онтологии как базы знаний специального типа, которая может "читаться" и пониматься, отчуждаться от разработчика и/или физически разделяться ее пользователями. Выбор данного определения объясняется тем, что при использовании онтологии для описания структуры и семантики документа системе необходимо понимать ее без участия человека.

## Формальное определение онтологии

Под *формальной моделью онтологии* O будем понимать упорядоченную тройку вида:

, где

 - X конечное множество концептов (понятий, терминов) предметной области, которую представляет онтология ;

 - K конечное множество отношений между концептами (понятиями, терминами) заданной предметной области;

 - Ф конечное множество функций интерпретации (аксиоматизация), заданных на концептах и/или отношениях онтологии.

В данном определении естественным ограничением, накладываемым на множество, является его конечность и непустота. От компонент K и Ф требуется только конечность, т.к. существуют вырожденные случаи, связанные с их пустотой. И если понятия предметной области специфичны в каждой онтологии, то отношения более универсальны. Обычно базовыми считают отношения “is\_a” («экземпляр – класс»), “part\_of” («часть – целое»), и некоторые другие.

В решении задач лингвистики онтология может уточняться:

O=(R, R, P, D, VD, TD, A, P, RAD, RDV), в которой

C={ }– конечное множество понятий,

R={r1,…, rm}, *R*⊆C×C *R=RC**RT**RA* – конечное множество бинарных отношений между понятиями,

*RC* ⊆C×C, *RC*⊆*R* – антисимметричное, транзитивное, нерефлексивное бинарное отношение наследования, являющееся отношением частичного порядка на множестве понятий C,

– бинарное отношение «часть-целое»,

*RA* ⊆C×C, *RA*⊆*R* – конечное множество ассоциативных отношений, *RRCCRAA*⊆×⊆,

– конечное множество доменов,

DV={dv1,…, dvu}= – конечное множество конкретных значений стандартного типа string, включенных в некоторый домен,

*DT*= *D*∪ {*string,* integer, *boolean*}– множество типов данных, включающее три стандартных типа string, boolean, integer и множество доменов,

* – конечное множество атрибутов, т.е. бинарных отношений или ,

– конечное множество конкретных свойств атрибута, включающее свойства {multiplicity, key, mandatory},

 – бинарное отношение инцидентности между множествами атрибутов A и свойств атрибутов AP,

 – бинарное отношение инцидентности между множествами доменов D и доменных значений DV.

## Классификации онтологий

В проектировании онтологий условно можно выделить два направления, до некоторого времени развивавшихся отдельно. Первое связано с представлением онтологии как формальной системы, основанной на математически точных аксиомах. Второе направление развивалось в рамках компьютерной лингвистики и когнитивной науки. Там онтология понималась, как система абстрактных понятий, существующих только в сознании человека, которая может быть выражена на естественном языке (или какой-то другой системой символов). При этом обычно не делается предположений о точности или непротиворечивости такой системы.

Таким образом, существует два альтернативных подхода к созданию и исследованию онтологий. Первый (*формальный*) основан на логике (предикатов первого порядка, дескриптивной, модальной и т.п.). Второй (*лингвистический*) основан на изучении естественного языка (в частности, семантики) и построении онтологий на больших текстовых массивах, так называемых *корпусах*.

В настоящее время данные подходы тесно взаимодействуют. Идет поиск связей, позволяющих комбинировать соответствующие методы. Поэтому иногда бывает сложно отделить лексические онтологии с элементами формальных аксиоматик от логических систем с включениями лингвистических знаний.

Независимо от различных подходов можно выделить 3 основных принципа классификации онтологий:

* По степени формальности
* По наполнению, содержимому
* По цели создания

В рамках этой классификации по цели создания выделяют 4 уровня: Онтологии представления, онтологии верхнего уровня, онтологии предметных областей и прикладные онтологии.

Цель создания онтологии представления - описать область представления знаний, создать язык для спецификации других онтологий более низких уровней. Пример: описание понятий языка OWL средствами RDF/RDFS.

Назначение онтологии верхнего уровня в создании единой “правильной онтологии”, фиксирующей знания общие для всех предметных областей и многократном использовании данной онтологии. Существует несколько серьезных проектов: SUMO, Sowa’s Ontology, Cyc. Но в целом попытки создать онтологию верхнего уровня на все случаи жизни пока не привели к ожидаемым результатам. Многие онтологии верхнего уровня похожи друг на друга. Они содержат одни и те же концепты: Сущность, Явление, Процесс, Объект, Роль и т.п.

Назначение онтологий предметных областей схоже с назначением онтологий верхнего уровня, но область интереса ограничена предметной областью (авиация, медицина, культура). Примеры: АвиаОнтология, CIDOC CRM, UMLS.

Назначение Прикладные онтологии этих онтологий в том, чтобы описать концептуальную модель конкретной задачи или приложения. Они содержат наиболее специфичную информацию. Примеры проектов: TOVE, Plinius.

**Теоретические подходы к автоматическому построению онтологий**

Постановку задачи можно сформулировать следующим образом. На входе имеется коллекция текстов, тематика которых принадлежит одной предметной области. Необходимо получить на основе данной коллекции множество понятий (концептов) предметной области и связи между ними. Формат входных данных и возможные связи между объектами могут варьироваться в зависимости от метода.

Большая часть встретившихся в ходе исследования методов автоматического построения онтологий включает в себя элементы лингвистического анализа. Иногда лингвистический анализ комбинируется с использованием статистики или методов искусственного интеллекта (например, генетических алгоритмов).

На основе исследованных методов автоматического построения онтологий можно выделить следующие группы:

- лингвистические методы;

- статистико-лингвистические методы;

- методы, опирающиеся на способ хранения информации;

- прочие методы.

Ниже будут рассмотрены методы, относящиеся к каждой из групп. Важно уточнить, что в данной статье приведены условные названия групп и самих методов.

**Лингвистические методы**

Данная группа содержит множество методов, опирающихся на лингвистический анализ текста.

Одним из таких методов является использование лексико-синтаксических шаблонов [7]. Подобные шаблоны разрабатываются экспертами и позволяют выявить отношения между терминами на основе части текста на естественном языке.

Разработанные Е.И. Большаковой и её коллективом шаблоны [8] применяются для анализа научно-технических документов. Существует язык LSPL, предназначенный для формализованной записи шаблонов. Однако стоит отметить, что данный язык не единственный: например, в работе [7] Е.А. Рабчевский описывает собственный язык лексико-синтаксических шаблонов, разработанный отдельно от группы Большаковой и названный также LSPL. По форме записи язык Рабчевского похож на языки HTML, XML. На основе полученных из текста на естественном языке семантических моделей можно построить онтологию.

Некоторые методы также используют языковые шаблоны для выявления связей между понятиями, но связи эти устанавливаются в ходе анализа корпуса текстов предметной области. В работе [9] предлагается информационная система для построения и самообучения онтологии. В процессе построения онтологии производится разметка корпуса синтаксическими тегами (несущими информацию о морфологии слова), определяются наиболее часто встречающиеся лингвистические структуры, на основе чего строятся лингвистические шаблоны. Далее из текста выделяются фрагменты, соответствующие полученным шаблонам, и после нормализации новое понятие вводится в базовую онтологию. Для оценки релевантности выделенных понятий необходимо мнение эксперта предметной области.

Ещё один лингвистический метод реализован в плагине OntoLT [10]. В нём используется XML-разметка текста, на основе которой по указанным правилам выделяются кандидаты на роли сущностей и связей. Правила вводятся пользователем до начала работы приложения на специальном языке, названном в документации «precondition language». Они тоже представляют собой своего рода лексико-семантические шаблоны.

**Статистико-лингвистические методы**

Следующая группа методов помимо лингвистических особенностей текста, учитывает различные статистические характеристики.

Их применение удобно рассмотреть на примере статьи Е.С. Мозжериной [11] об автоматическом построении онтологий на базе коллекции текстов. Перед вычислением собственно статистик производится предварительная обработка корпуса, включающая в себя приведение документов к единому формату, токенизацию, лемматизацию и исключение стоп-слов. Для выделения из коллекции текстов терминов автор вводит следующие эвристики:

- имя класса онтологии содержит хотя бы одно существительное;

- общеупотребительные слова по сравнению с терминами обладают большей частотой встречаемости, приблизительно равной в различных предметных областях;

- количество информации термина нескольких слов больше, чем количество информации отдельных слов, входящих в его состав.

С опорой на указанные эвристики выделяются термины. Также автор указывает на возможность использования существующих разработанных экспертами онтологий для работы с неспециализированными в предметной области документами.

Для случая двусложных терминов приведена следующая статистика, определяющая количество «взаимной информации» [11]:

 

где *x*, *y* – отдельные слова термина, *P(x)* – частота встречаемости *x*, *P(x,y)* – частота совместной встречаемости *x* и *у*.

Отношения между классами также определяются с помощью количества информации. В работе Мозжериной описаны отношения типов «is-a» и «synonym-of». По количеству информации можно распределить уровни иерархии между терминами. Также учитывается вхождение слова (более общего термина) в другие термины. Кроме этого, отношение «is-a» может устанавливаться из учёта контекста, как и отношение «synonym-of». В данном случае контекст термина понимается как множество слов, которые встречаются одновременно с данным.

Другой способ использования статистики показан в работе [12]. Этот способ удобен для викификации текстов (соотношения понятий и сущностей из текста со связанной страницей в Википедии). Подготовка корпуса текстов аналогична предыдущему методу. Далее из текста выделяются кандидаты в термины: все одно-, двух- и трёхсловные словосочетания, удовлетворяющие имеющимся лингвистическим шаблонам (например, «прилагательное\_существительное», «существительное\_существительное»). С помощью методов, реализованных в авторском фреймворке для обработки текстов Texterra, из имеющегося списка выделяются термины.

Texterra базируется на знаниях из Википедии и сохраняет все входящие и исходящие ссылки для статьи. Таким образом для каждого нового понятия определяются «соседи» по контексту, с учётом того, что термины, появляющиеся всхожих контекстах, обычно имеют схожее значение. При обработке термина выбираются термины из текстового «окна» (15 слов справа и слева от термина) и вычисляются случаи совместного употребления с данным термином. В качестве оценки возможности неслучайного «соседства» терминов используется статистика t-тест [13].

Ещё один статистический метод представлен в работе [14]. Извлечение терминов производится с помощью z-статистики (z-score), вычисляемого для значений частотности слов и «показателя странности» (weirdness index). Последняя характеристика вычисляется с использованием как специализированного корпуса, так и неспециализированного (e.g. British National Corpus). Отношения вида «is\_a» определяются между словами и содержащими их словосочетаниями (collocates), на основе чего строятся иерархии.

**Методы, опирающиеся на способ хранения информации**

Как видно из названия группы, следующие методы для построения онтологий используют особенности определённым образом структурированной информации.

Одним из таких методов является кластеризация [15]. Для построения онтологий с её помощью документы представляются в виде набора терминов и разбиваются на кластеры по тематике. Определяется частота встречаемости каждого термина (вес). Терминами онтологии выбираются термины с частотой выше средней, а термины с максимальным весом являются понятиями (концепциями). Представляя коллекцию документов в виде дерева, можно задать иерархические отношения в онтологии (таксономию).

В работе [15] также представлен способ построения онтологии на основе УДК и библиографических баз данных. Термины, использующиеся для построения онтологии, выбираются из расшифровки УДК, а также ключевых слов, указанных библиографами в описании статьи. На основе этих методов строятся классы онтологии и проводятся связи. Основная идея метода может быть представлена на схеме (рис. 2.1):

 

*Рисунок 2.1. Метод выделения терминов из ББД [*15*]*

Среди недостатков данного метода авторы отмечают проблемы, связанные с жёсткой структурой УДК и отсутствием кодов УДК в некоторых базах данных, а также с выделением ключевых слов для библиографических записей.

**Прочие методы**

Существуют методы, которые нельзя отнести ни к одной из вышеописанных групп. В частности, таким является один из методов искусственного интеллекта, описанный в монографии Л.В. Найхановой [16]. Здесь для автоматического построения онтологий применяется генетическое и автоматное программирование. Одним из особенностей и достоинств данного метода можно назвать возможность извлекать в ходе построения онтологии новую информацию.

**Сравнение методов автоматического построения онтологий**

Универсального алгоритма автоматического построения онтологий в настоящий момент не существует. Каждый из вышеописанных методов обладает достоинствами и недостатками, связанными с самой идеей алгоритма и требующимися для его осуществления данными.

К последним можно отнести такие данные, как набор лексико-семантических шаблонов (метод лексико-семантических шаблонов), правил определения онтологических структур (метод OntoLT [10]), список начальных эвристик для выделения терминов (метод, описанный в работе Мозжериной [11]). Кроме того, некоторые методы требуют особенной (не стандартной для корпуса текстов) предварительной разметки текста. Так, метод OntoLT базируется на работе с XML-разметкой.

Среди вышеописанных методов есть такие, которые требуют доступа к другим ресурсам. Например, к Википедии (Texterra [12]), к неспециализированному корпусу (метод Ahmad-Gilliam [14]), к библиографическим базам данных (метод построения онтологий по УДК [15]).

Большая часть методов автоматического построения онтологий может быть универсальной в плане работы с разными языками, но потребуют создания собственных или использования имеющихся наборов данных, относящихся к нужному естественному языку. Например, лексико-семантических шаблонов или стоп-слов.

Следует также отметить, что описанные методы позволяют выявлять различные виды онтологий. Наиболее широкий спектр возможностей в этом плане предоставляют методы лексико-семантических шаблонов, т.к. шаблоны составляются с учётом не только морфологии, но и семантики естественного языка.

**Оценка эффективности методов автоматического построения онтологий**

Качество построенных онтологий может различаться в зависимости от имеющихся данных и целей приложения, использующего онтологию. В работе [11] предложено оценивать онтологии по качеству работы систем семантического поиска, использующих онтологии.

Очевидным подходом является сравнение полученной онтологии предметной области с экспертным решением задачи. Однако работа экспертов требует большого труда и времени, поэтому особую ценность в деле оценки эффективности алгоритмов приобретают те общедоступные специализированные корпуса, для которых существует построенный тезаурус терминов [11].

Итак, существуют методы автоматического построения онтологий, использующие разные подходы для решения задачи. В зависимости от подходов было выделено четыре основных группы методов: лингвистические, статистико-лингвистические, опирающиеся на способ хранения информации и прочие методы. Основная идея лингвистических методов состоит в выявлении понятий и связей с помощью лексико-синтаксических шаблонов. Статистико-лингвистические методы при построении онтологии опираются на вычисление статистических функций, учитывая лексические особенности при выделении терминов. Методы, опирающиеся на способ хранения информации, работают с документами, имеющими чёткую структуру – например, из библиографических баз данных. Прочие методы включают в себя элементы генетического и автоматного программирования и др.

Целесообразность использования того или иного из вышеописанных методов зависит от множества факторов. Среди них: наличие или отсутствие необходимых данных, в том числе учитывающих специфику естественного языка; доступа к ресурсам, которые используются в методе; наличие возможности структурировать имеющиеся данные определённым образом; типа онтологий, необходимых для решения пользовательских задач.

Возможно также повышение эффективности алгоритмов за счёт использования дополнительных лингвистических шаблонов, статистических характеристик, кластеризации коллекции документов.

Таким образом, подобрав и реализовав оптимальный метод автоматического построения онтологий, можно расширить функционал приложения, предназначенного для анализа текстовых документов. Также при построении онтологии будут полезны метаданные о документе и его семантический индекс, которые могут быть получены с использованием многоаспектной онтологии электронных документов.

Литература:

1. Национальный корпус русского языка [Офиц. сайт]. // URL: http://www.ruscorpora.ru/ (дата обращения: 01.10.2014).
2. Software, Tools, Lists, Recourses [Электр. ресурс] // URL: http://www.uow.edu.au/~dlee/software.htm (дата обращения: 01.10.2014).
3. *Лукашевич Н.В.* Тезаурусы в задачах информационного поиска. // М.: Издательство Московского университета, 2011. – С. 100-103.
4. *Murray T.* Special purpose ontologies and the representation of pedagogical knowledge // ICLS '96 Proceedings of the 1996 international conference on Learning sciences. – p. 235-242
5. W3C Semantic Web Activity Homepage [Офиц. cайт]. // URL: http://www.w3.org/2001/sw/ (дата обращения: 01.10.2014).
6. *Рабчевский Е.А.* Автоматическое построение онтологий на основе лексико-синтаксических шаблонов для информационного поиска. // Труды 11-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL’2009. – Петрозаводск, 2009. – С. 69-77.
7. *Большакова Е.И., Васильева Н.Э., Морозов С.С.* Лексико-синтаксические шаблоны для автоматического анализа научно-технических текстов // Десятая Национальная конференция по искусственному интеллекту с международным участием КИИ-2006. Труды конференции в 3-х томах. М.: Физматлит, 2006. – Т.2. – С. 506-524.
8. *Оробинская Е.А.* Метод автоматического построения онтологии предметной области на основе анализа лингвистических характеристик текстового корпуса [Электронный ресурс] // URL: arxiv.org/pdf/1405.1346 (дата обращения: 01.10.2014).
9. *Buitelaar, P., Olejnik, D., Sintek,* *M.* A protégé plug-in for ontology extraction from text based on linguistic analysis. // Bussler, C.J., Davies, J., Fensel, D., Studer, R. (eds.) ESWS 2004. LNCS – T. 3053 – C.31-44.
10. *Мозжерина Е. С.* Автоматическое построение онтологии по коллекции текстовых документов // Электронные библиотеки: Перспективные Методы и Технологии, Электронные коллекции – RCDL 2011 – Воронеж, 2011 – С. 293-298.
11. *Astrakhantsev N. A., Fedorenko D. G., Turdakov D. Y.* AUTOMATIC ENRICHMENT OF INFORMAL ONTOLOGY BY ANALYZING A DOMAIN-SPECIFIC TEXT COLLECTION // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 4 — 8 июня 2014 г.). Вып. 13 (20). — М.: Изд-во РГГУ, 2014. – C. 29-42.
12. *Manning C.D.* Foundations of statistical natural language processing. // H. Schütze (Ed.). MIT press, 1999.
13. *Ahmad K, Gillam L.* Automatic Ontology Extraction from Unstructured Texts // In (Eds.) R. Meersman and Z. Tari. On the Move to Meaningful Internet Systems -OTM, Confederated Int. Conf., CoopIS, DOA, and ODBASE 2005, Agia Napa – Ч. II – С. 1330-1346.
14. *Захарова И.В., Тимченко М.С.* Способы автоматического построения онтологии для задач анализа текстов // Знания—Онтологии—Теории: труды Всеросийской конференции ЗОHT-09. — Новосибирск, 2009. — Т. 2. — С. 164—167.
15. *Найханова Л.В.* Технология создания методов автоматического построения онтологий с применением генетического и автоматного программирования // Л.В.Найханова. - Улан-Удэ: Издательство БНЦ СО РАН, 2008. – 244 c.