УДК 004.82

МЕТОД ИЗВЛЕЧЕНИЯ ИЗ ВИКИПЕДИИ ПРЕДМЕТНО-ОРИЕНТИРОВАННЫХ ЛИНГВИСТИЧЕСКИХ ОНТОЛОГИЙ

Кравцов Д.В.

ФГБОУ ВПО «Брянский государственный технический университет», Брянск, e-mail: dkrbox@gmail.com

В рамках инжиниринга лингвистических онтологий на основе Википедии востребованной является задача автоматизированного извлечения из полной мультипредметной Википедии онтологий конкретных предметных областей, решению которой посвящена статья. Для этого был разработан универсальный алгоритм, не зависящий от предметной области, основанный на структуре категорий Википедии и оценке семантической близости понятий предметной области. Алгоритм работает с ранее построенным взвешенным графом отношений Википедии, суть его сводится к тому, что он постепенно пополняет выходной граф понятий предметной области понятиями, семантически близкими к уже имеющимся в нем. В ходе обзора литературы для оценки семантической близости была выбрана взвешенная мера Дайса. Для оценки алгоритма была подготовлена небольшая тестовая коллекция на тему «Всемирная паутина» на основе одноименной категории Википедии. Было изучено поведение алгоритма при разных уровнях порога отсечения понятий по величине семантической близости, и получено его оптимальное значение, при котором максимальна F-мера.

Ключевые слова: онтология предметной области, лингвистическая онтология, автоматическое построение онтологий, Википедия, семантическая близость, мера Дайса, извлечение понятий предметной области, семантические отношения

METHOD OF DOMAIN SPECIFIC LINGUISTIC ONTOLOGY EXTRACTION FROM WIKIPEDIA

Kravtsov D.V.

FGBOU VPO «Bryansk State Technical University», Bryansk, e-mail: dkrbox@gmail.com

Within linguistic ontology engineering based on Wikipedia, the important task is an automated retrieval of domain ontologies from the full multi-domain Wikipedia, and this is subject of this article. For this purpose, a universal algorithm independent of the specific domain was developed, based on Wikipedia category structure and the assessment of the semantic relatedness of domain concepts. The algorithm works with a previously built weighted relations graph of Wikipedia, it gradually adds to the output graph of domain concepts new concepts that are semantically related to those already existing in it. After literature review the weighted Dice measure was chosen to calculate semantic relatedness. For the evaluation of the algorithm small test collection named «World Wide Web» was prepared based on the corresponding Wikipedia category. We have studied the behavior of the algorithm with different levels of semantic relatedness cut-off threshold and obtained its optimal value at which the F-measure is maximal.

Keywords: domain ontology, linguistic ontology, ontology learning, Wikipedia, semantic relatedness, Dice measure, domain concepts extraction, semantic relationships

Практика отечественных [4, 10] и зарубежных [8, 7] ученых показывает, что базы знаний онтологического типа, в частности т.н. лингвистические (лексические) онтологии, ЛО (подробно описаны в книге [6]), могут весьма успешно применяться в самых разных задачах обработки естественного языка и информационного поиска. Они позволяют перейти от статистической обработки слов к обработке понятий и их смысловых связей, что существенно улучшает качество решения этих задач. Однако создание таких ресурсов вручную с нуля – дело очень трудозатратное и долгое, поэтому многие исследователи разрабатывают методы автоматизации их создания. Одним из популярных источников знаний для этого служит Википедия – самая большая в мире электронная энциклопедия. Среди проектов по созданию семантических баз знаний на основе Википедии (и других источников) можно выделить DBpedia, YAGO версий 1, 2 и 3, BabelNet, Текстерра [10]. Работа с онтологиями таких размеров требует больших вычислительных мощностей. С точки зрения использования в практических приложениях, больший интерес представляют онтологии конкретных предметных областей, которые на порядок меньше, но обеспечивают решение многих задач с сопоставимым качеством. Поэтому актуальна задача их автоматизированного извлечения из мультипредметных онтологий, построенных из Википедии, которой и посвящена данная работа. Подобную задачу решали авторы работы [9] с использованием инфраструктуры системы анализа текстов Текстерра. Их подход основан на представлении категорий как векторов в пространстве понятий, где компоненты вектора соответствуют степени принадлежности понятия категории.

Цель работы

Данная работа выполняется в рамках проекта, посвященного разработке методов

автоматизированного построения лингвистических онтологий на основе Википедии и других вики-систем для применения в разнообразных задачах обработки естественного языка. Та часть работы, которой посвящена данная статья, имеет своей целью разработку метода извлечения из полной онтологии, построенной по Википедии, онтологии произвольной заданной предметной области. Для достижения этой цели необходимо решить следующие задачи:

1) разработать алгоритм выделения понятий предметной области,

2) написать программную реализацию алгоритма,

3) подготовить тестовую коллекцию,

4) провести экспериментальную оценку алгоритма.

Материалы и методы исследования

Очевидное решение задачи – выбрать всё от заданной вики-категории вниз по структуре категорий, часто дает неудовлетворительный результат: в выборку попадает очень много совершенно непертинентных категорий и понятий. Например, предметной области «Телекоммуникационные технологии» наилучшим образом соответствует категория Википедии «Телекоммуникации», но оказывается, что в нее среди прочего попадает через категорию «Телевидение» более 500 названий телефильмов и 2300 телесериалов, которые, очевидно, к предметной области не относятся и, если не считать категорийных ссылок, имеют мало общего (как по ссылкам, так и по тексту статей) с понятиями, релевантными предметной области. Поэтому для фильтрации понятий был разработан алгоритм, который учитывает семантическую близость (СБ, semantic relatedness) понятий, относящихся к одной предметной области.

Для расчета семантической близости понятий мы используем заранее построенный по всей Википедии взвешенный *граф отношений* W [5]. Данный граф имеет 2 типа узлов – p-узлы, которым соответствуют обычные страницы (т.е. понятия), k-узлы, которым соответствуют страницы-категории, и 2 типа ребер – взвешенные p`-ребра (обычные ссылки между страницами) и k`-ребра (ссылки на вышестоящие категории). Парсер вики-текста страниц разбивает контент на странице на ряд информационных блоков: «шапка» страницы, раздел «История», инфоблок, основной текст, текст ссылок типа «Основная статья», блок «См. также», навигационный шаблон. Каждому типу блока назначен весовой коэффициент, выражающий его относительную значимость при определении семантической близости к другим понятиям (страницам), на которые ссылается текущая страница. Наибольший вес имеет «шапка» страницы – текст вначале статьи до блока «Содержание», который обычно содержит определение и самые основные сведения о предмете статьи, наименьший вес – основной текст. Алгоритм, вычисляющий вес p`-ребра от узла *a* к узлу *b*, учитывает такие параметры, как количество ссылок со страницы *a* на страницу *b,* их расположение по блокам на странице *a*, размер блоков, общее количество ссылок в блоке и в целом на странице.

К настоящему времени предложено большое количество мер (и алгоритмов их вычисления) похожести вершин графов вообще и семантической близости понятий Википедии, в частности. Хороший обзор таких мер с упором на анализ их вычислительной сложности сделан в работах [3] и [2]. Меры, основанные на ссылках между статьями Википедии, можно классифицировать на три основные группы:

- меры парного случайного блуждания (SimRank, мера близости Ньюмана),
- меры случайного блуждания (мера Грина, локальный PageRank, PageSim и др.),
- нерекурсивные меры (косинус, меры Дайса, Жаккара, Кульчинского и др.).

Популярная рекурсивная мера парного случайного блуждания SimRank вычисляется по следующей итерационной формуле:

$$S_{ij} = \frac{C}{k_i k_j} \sum_{uv} A_{iu} A_{vj} S_{uv},$$

где S_{ij} – элемент матрицы подобия вершин, A_{ij} – элемент матрицы смежности, k_i – степень *i*-й вершины, C – коэффициент затухания. Вычислительная сложность этой меры как и меры Ньюмана очень высока – $O(n^3)$, n – количество ребер графа, и, как отмечает автор [3], из-за очень маленького диаметра графа Википедии обе меры вычисляют полную матрицу семантической близости, а потому практически не вычислимы.

Меры случайного блуждания в плане вычислительной сложности существенно превосходят меры парного случайного блуждания – O(n), и все же в больших онтологиях могут оказаться недостаточно эффективными.

Среди традиционных нерекурсивных мер интерес вызывает мера Дайса, здесь *N(a)* – множество вершин, соседних с вершиной *a*:

$$Dice(a,b) = \frac{\left|N(a) \cap N(b)\right|}{\left|N(a)\right| + \left|N(b)\right|}.$$

Несмотря на простую интерпретацию – отношение количества общих соседей к сумме количеств соседей каждой из вершин, согласно экспериментальным данным мера Дайса показывает очень хорошие результаты. Так, в работе [1] показано, что при решении задачи разрешения лексической многозначности по методу системы Текстерра мера Дайса показывает самые хорошие результаты на всех четырех использовавшихся тестовых наборах данных по сравнению с различными вариациями мер на основе поиска кратчайших путей. Мы используем взвешенную меру Дайса, которая для пары понятий *a* и *b* рассчитывается как

$$Dice(a,b) = \frac{\sum_{i \in N(a) \cap N(b)} \left(w_{a,i} + w_{b,i}\right)}{\left(\sum_{j \in N(a)} w_{a,j} + \sum_{k \in N(b)} w_{b,k}\right)},$$

где $w_{a,i}$ – вес ребра между вершинами *a* и *i*. Интерпретация также проста – отношение суммы весов связей с общими соседями к сумме весов связей со всеми соседями каждой из двух вершин *a* и *b*.

Предметная область, лингвистическую онтологию которой необходимо построить, задается с помощью набора *K* категорий Википедии, соответствующих наиболее общим, высокоуровневым понятиям предметной области («потолок» предметной области) и набора C статей Википедии, соответствующих

ключевым понятиям предметной области, основным ветвям ее таксономии («ядро» предметной области). Итак, предлагаемый алгоритм выделения понятий предметной области на входе принимает следующие данные:

1) граф отношений W;

2) набор категорий K;

3) набор понятий C (p-узлов графа);

4) порог семантической близости **r** для определения релевантных понятий;

5) стоп-лист S для исключения заведомо нерелевантных понятий (можно задавать с помощью регулярных выражений).

Алгоритм использует функцию семантической близости rel(A, B), где аргументами могут быть одиночные понятия или наборы понятий. В общем случае она вычисляется как среднее значение взвешенной меры Дайса для всех пар из A×B. На выходе алгоритм должен построить граф D, являющийся максимальным подграфом графа W, состоящим только (в идеальном случае) из понятий, относящихся к заданной предметной области. Алгоритм состоит из следующей последовательности шагов.

1. Граф D инициализируется узлами из множеств C и K. Здесь и далее при добавлении пары узлов добавляются и соответствующие им ребра.

2. Для каждой пары узлов $p_i \in C$ и $k_j \in K$ в графе W находятся все пути от понятия p_i к категории k_j по ребрам k`-типа. Все k-узлы путей, а также p-узлы, соответствующие «Основной статье» категорий добавляются в граф D. Здесь и далее узлы перед включением в выходной граф проверяются на отсутствие в стоп-листе S.

3. Для каждой добавленной категории k_i проверяются непосредственно входящие в нее понятия на предмет соответствия предметной области. Для каждого p_j, инцидентного k_i последовательно проверяются следующие условия (по мере увеличения вычислительной сложности):

1) если p_j соответствует стоп-листу, то переходим к следующему понятию – p_{j+1};

2) все категории, с которыми связано понятие p_j, уже включены в граф D;

3) понятие p_j достаточно близко другим понятиям текущей категории, включенным в граф D, т.е. $rel(p_j, P_D(k_i)) \geq r$, где $P_D(k_i)$ – множество дочерних p-узлов узла k_i графа D;

4) понятие p_j достаточно близко понятиям текущей категории и ее подкатегорий, включенным в граф D, т.е. $rel(p_j, T_D(k_i, q)) \geq r$, где $T_D(k_i, q)$ – множество всех p-узлов поддерева графа D, построенного от узла k_i вниз по k`-ребрам на глубину q (по умолчанию q = 2);

Если какое-либо из условий 2–4 выполняется, добавляем p_j в D и переходим к p_{j+1}.

4. По аналогии с шагом 3 для каждой добавленной категории k_i проверяются непосредственно входящие в нее подкатегории k_j на предмет соответствия предметной области с той разницей, что семантическая близость вычисляется не для одного понятия p_j а для набора понятий, непосредственно принадлежащих категории k_j , т. е. $P_W(k_j)$.

5. Для каждой подкатегории, добавленной на шаге 4, рекурсивно повторяются шаги 3 и 4. Алгоритм выполняется до тех пор, пока не будет обойдено всё поддерево категорий вниз полностью, либо на заданную глубину.

Суть алгоритма заключается в том, что он постепенно пополняет выходной граф понятий предметной области понятиями, семантически близкими к уже имеющимся в нем. Мы запускаем алгоритм итерационно, т.к. после каждого прохода результирующее множество узлов меняется, а с ним и оценка семантической близости понятий.

Результаты исследования и их обсуждение

Для оценки качества работы предложенного алгоритма выделения предметной области из Википедии была вручную (с применением приемов автоматизации) подготовлена тестовая коллекция понятий Википедии, относящихся к относительно небольшой предметной области. Для этого в качестве корневой была выбрана категория «Всемирная паутина» и рекурсивно ее подкатегории и статьи, исключая те из них, которые слабо относятся к основной теме (например, в каком-то частном аспекте). В ходе эксперимента варьировалось значение порога семантической близости r и оценивалась получаемая выборка понятий относительно тестовой коллекции по стандартным метрикам, приведенным в таблице. r изменялось в диапазоне 0..1 с шагом 0,1 и затем в диапазоне 0,2..0,3 с шагом 0,01, оптимальное значение (для данной коллекции), при котором F_1-мера максимальна, составило 0,22.

Результаты тестирования алгоритма извлечения понятий

Порог сем. близости r	Точность	Полнота	F_1-мера
0	0,18	1	0,31
0,1	0,52	0,93	0,67
0,22	0,74	0,89	0,81
0,3	0,79	0,46	0,58
0,9	0,95	0,12	0,21

При r = 0 выбираются все понятия из поддерева «Всемирная паутина», при этом, как видно, только 18% из них включены в тестовую коллекцию. Объясняется это тем, что в дереве категорий Википедии присутствуют, а из тестовой коллекции исключены такие категории, как например «Сайты», в которую входит список из более чем 1100 описанных в Википедии сайтов произвольной тематики. С увеличением r большая часть таких нерелевантных понятий отсекается и точность резко возрастает. При значениях порога значительно больше оптимального резко падает полнота, т.к. выбираются только понятия, сильно связанные с понятиями «ядра» предметной области. При оптимальных же значениях r алгоритм демонстрирует хорошие показатели, особенно если учитывать, что качество те-

стовой коллекции было не очень высоким. Алгоритм показал высокую производительность, т.к. использована «легкая» мера семантической близости.

Заключение

В ходе проделанной работы был разработан универсальный метод извлечения из Википедии множества понятий заданной предметной области и построения ее лингвистической онтологии, основанный на структуре категорий Википедии и оценке семантической близости понятий предметной области. Качество его работы сильно зависит от задаваемого исходного набора понятий и порога семантической близости. На тестовом примере был подобран порог, для которого показатели качества оказались высокими. Однако пока остается открытым вопрос, насколько значение порога, подобранное для одной предметной области, будет оптимальным или хотя бы приемлемым для другой, и как лучше его определять. Это задача для дальнейшей работы. Также планируются такие улучшения алгоритма, как извлечение связанных понятий, не попадающих в заданную структуру категорий, определение и отсечение паразитных категорий, не несущих значимой семантической связи, например, таких как «Категории по алфавиту» или «Родившиеся 3 июля».

Работа выполнена при поддержке РФФИ (проект № 14-07-31261 мол_а).

Список литературы

1. Варламов М.И., Коршунов А.В. Расчет семантической близости концептов на основе кратчайших путей в графе ссылок Википедии // Machine Learning. – 2014. – Т. 1.

2. Варламов М.И. Расчет семантической близости концепций с использованием связей в графе ссылок Википедии: Дипл. работа. – Москва, 2014. – 49 с.

3. Велихов П.Е. Меры семантической близости статей Википедии и их применение к обработке текстов // Информационные технологии и вычислительные системы. – 2009. – № 1. – С. 23–37.

4. Добров Б.В., Лукашевич Н.В. Лингвистическая онтология по естественным наукам и технологиям для приложений в сфере информационного-поиска // Ученые записки Казанского Государственного Университета. Серия «Физико-математические науки». – 2007. – Т. 149, Книга 2. – С. 49–72.

5. Кравцов Д.В. Формализация построения лингвистических онтологий на основе Википедии с использованием нечетких семантических отношений // Труды XVII Международной конференции DAMDID/RCDL'2015. – Обнинск, 2015.

6. Лукашевич Н.В. Тезаурусы в задачах информационного поиска. – М.: Издательство Московского университета, 2011. – 512 с.

7. Abulaish M., Dey L. Information extraction and imprecise query answering from web documents //Web Intelligence and Agent Systems. – 2006. – Т. 4, № 4.

8. Janik M., Kochut K. Training-less ontology-based text categorization : Diss. University of Georgia, 2008.

9. Korshunov A.V., Turdakov D.Yu. A category-driven approach to deriving domain specific subset of Wikipedia // Труды Института системного программирования РАН. – 2011. – Т. 21. – С. 323–347.

10. Texterra: инфраструктура для анализа текстов / Турдаков Д., Астраханцев Н., Недумов Я., Сысоев А., Андрианов И., Майоров В., Федоренко Д., Коршунов А., Кузнецов С. // Труды Института системного программирования РАН. – 2014. – Т. 26, № 1. – С. 421–438.