УДК 004.89:004.93

М.Ю. Губин, В.В. Разин, А.Ф. Тузовский

Методы создания семантических метаописаний документов с применением семантических сетей, фреймовых моделей и частотных характеристик

Сформулирован метод создания семантических метаописаний документов с помощью семантических сетей, фреймовых моделей и частотных характеристик анализируемых документов.

Ключевые слова: онтология, семантическая сеть, фрейм, RDF.

Введение

Естественные человеческие языки обладают большой выразительностью и сложностью, существенное влияние на смысл текста в них оказывают контекст и эмоциональная составляющая. Понимание естественного языка включает куда больше, чем разбор предложений на индивидуальные части речи и поиск значений слов в словаре. Оно базируется на обширном фоновом знании о предмете, идиомах, используемых в этой области, а также на способности применять общее контекстуальное знание для понимания недомолвок и неясностей, присущих естественной человеческой речи. Поэтому системы, использующие натуральные языки с гибкостью и общностью, характерными для человеческой речи, лежат за пределами существующих методологий [1]. Однако для определённых условий (когда документ имеет достаточно строгую грамматическую структуру, а следовательно, и содержит достаточно информативную формальную составляющую) данная задача решаема с достаточно высоким качеством распознавания смысла [2]. В этой статье будут описаны условия, выполнение которых необходимо для успешного распознавания, и предлагаемый алгоритм.

Постановка задачи

Данный алгоритм решает задачу создания метаописаний документов для последующего семантического поиска по ним на данном множестве документов D_i, относящихся к одной предметной области. Под **документом** D_i в рамках данного исследования будем понимать фрагмент текста на естественном языке.

Для реализации семантического поиска по документам необходимо создать достаточно полные семантические метаописания документов T_i.

Семантическое метаописание документа строится согласно **онтологии предметной области** O, представляющей собой набор **понятий** C_i, связанных между собой **отношениями** R_i. Также в онтологию предметной области входят **экземпляры объектов** E_i. Понятия, отношения и экземпляры имеют одну или более текстовых меток T_i. **Текстовая метка** T_i элемента онтологии – слово либо словосочетание естественного русского языка, соответствующее некоторому элементу онтологии.

Для построения базового семантического метаописания на основе текста документа для каждого его предложения L_i формируется **семантическая сеть**, представляющая собой граф, состоящий из множества вершин W_i и соединяющих их рёбер L_i. Элементарная сеть представляет результат синтаксического анализа и дополнительных семантических трансформаций дерева синтаксических зависимостей между словами в отдельном предложении. **Вершинами** W_i семантической сети являются сущности, встречающиеся в предложении, а **рёбра** L_i представляют собой семантические отношения между сущностями. Семантические сети предполагается получать из результатов синтаксического разбора текстов на естественных языках. Задача синтаксического разбора текстов на данный момент в различной степени решена для русского [6, 7] и английского [3–5] языков. Также существуют работы по синтаксическому разбору текстов на французском, норвежском, корейском и греческом [4], а также испанском и японском [4, 5] языках. В данной работе рассматривается частный случай с русским языком.

Программный интерфейс большинства существующих семантических анализаторов позволяет получить для каждой сущности набор направленных связей, исходящих от нее к другим сущностям. Направление связи обычно соответствует направлению синтаксического подчинения (для равноправных однородных членов предложения пара одинаковых

направленных связей идет в обе стороны). Семантические сети, соответствующие описанным выше критериям, могут быть использованы в разрабатываемом алгоритме с незначительными преобразованиями.

Семантическое метаописание – это набор извлечённых из предложений документа RDF-триплетов T_i, представляющих собой кортежи вида $<S_i,P_i,O_i>$, где S_i включен в объединение C_i и E_i, P_i включен в R_i, а O_i включен в объединение C_i и E_i.

Также для ускорения актуализации метаданных алгоритмом генерируются частотные характеристики слов в документе – TF- и IDF-терминов [8].

Алгоритм формирования метаданных отдельного документа

На вход алгоритма поступает исходный текст файла, а также набор текстовых меток элементов онтологии.

Шаги алгоритма:

Производится семантический анализ текста. Выходом этого шага является программная структура, содержащая всю требуемую информацию о тексте – слова с номером их начальных символов, смысловые связи между словами, обнаруженные и преобразованные в RDF триплеты (части предложений, соответствующие одному из описанных выше фреймов). Эта программная структура приводится к семантической сети, пригодной для обработки алгоритмом.

Подсчитывается количество вхождений слов в текст. При этом не учитываются так называемые «стоп-слова». Стоп-словами являются предлоги, союзы и частицы. Остальные слова нормализуются и количество вхождений подсчитывается именно для нормы слова.

Составляется ранговое распределение слов в документе. Слова с одинаковым количеством вхождений объединяются в классы, которые затем нумеруются в порядке убывания количества вхождений слов-членов класса в тексте, начиная с 1 [8].

Производится поиск класса, слова в котором являются значимыми для текста, с наибольшим номером. Все классы, идущие после него, отсеиваются и в дальнейшей работе алгоритма не участвуют. [8]

Выставляется первичное значение «веса» слов в документе. Оно равняется $N_{\max}/N_i$, где $N_{\max}$ – количество вхождений слов первого ранга, а N_i – количество вхождений слова t_i [8].

Производятся корректировки значений весов для упорядоченных пар слов, входящих в одни и те же триплеты либо предложения.

Из множества выделенных из текста RDF-триплетов выбираются:

– триплеты, каждая из позиций которых (субъект, предикат и объект) заняты в естественно-языковом представлении вхождением метки (соответственно, субъект и объект – метками понятия либо экземпляра, а предикат – меткой свойства);

– триплеты, одна из позиций которых занята вхождением ключевого слова, а две других – вхождением метки, так называемые триплеты-«кандидаты».

Выход алгоритма – метаописание документа, в которое входит набор записей вида $<E_i,S_i>$, где E_i – идентификатор элемента онтологии (так называемый URI – Universal Resource Identifier), а S_i – индекс значимости этого элемента для документа. При этом S_i имеет вид $S_i = <S_i\mathrm{TF},\ S_i\mathrm{IDF},S_i\mathrm{C}>$, где S_iTF – коэффициент значимости элемента с точки зрения документа (модифицированный коэффициент TF), S_iIDF – коэффициент значимости элемента с точки зрения набора документов (коэффициент IDF); S_iC – итоговое значение коэффициента значимости термина. В метаописание также входят все обнаруженные в тексте триплеты, все позиции которых заняты вхождениями меток элементов онтологии.

Кроме того, по завершении работы алгоритм генерирует набор вспомогательных записей, уменьшающих время возможной последующей повторной обработки документа.

Результаты работы алгоритма – семантические метаописания, которые позволяют реализовать семантический поиск и семантическую навигацию по обработанному множеству текстов. Качество распознавания находится на уровне примерно 60% от распознавания человеком, в зависимости от полноты онтологии предметной области и глубины анализа текста.

Литература

1. Люгер Д.Ф. Искусственный интеллект: стратегии и методы решения сложных проблем. – 4-е изд. – М.: Вильямс, 2003. – 864 с.

2. Хорошилов А.А. Компьютерная лингвистика и перспективные информационные технологии: теория и практика / А.А. Хорошилов, Г.Г. Белоногов, Ю.П. Калинин // НТИ. Сер. 2. Информ. процессы и системы / ВИНИТИ. – 2004. – № 8. – С. 30–43.

3. Poon H., Domingos P. Unsupervised semantic parsing. ACL Anthology. A Digital Archive of Research Papers in Computational Linguistics [Электронный ресурс]. – Режим доступа: www.aclweb.org/anthology/D/D09/D09-1001.pdf, свободный (дата обращения: 02.10.2010).

4. Deep linguistic processing with hpsg [Электронный ресурс]. – Режим доступа: http://www.delph-in.net, свободный (дата обращения: 02.10.2010).

5. Сайт лаборатории speech technology копрорации microsoft [Электронный ресурс]. – Режим доступа: http://research.microsoft.com/en-us/groups/srg/default.aspx, свободный (дата обращения: 02.10.2010).

6. Сайт рабочей группы «Автоматическая обработка текстов» [Электронный ресурс]. – Режим доступа: http://aot.ru/, свободный (дата обращения: 02.10.2010).

7. Сайт компании RCO [Электронный ресурс]. – Режим доступа: http://www.rco.ru, свободный (дата обращения: 02.10.2010).

8. Roelleke T. TF-IDF uncovered: a study of theories and probabilities / T. Roelleke, J. Wang // Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval (Singapore, July 20 – 24, 2008). – New York: SIGIR'08. ACM, 2008. – P. 435–442.

Губин Максим Юрьевич
Аспирант каф. оптимизации систем управления института кибернетики
Национального исследовательского Томского политехнического университета (НИТПУ)
Тел.: (382-2) 44-46-53, 8-906-950-73-14
Эл. почта: gubin.m.u@gmail.com

Разин Владимир Викторович
Аспирант каф. оптимизации систем управления института кибернетики НИТПУ
Тел.: (382-2) 46-52-23
Эл. почта: mooonbreeze@sibmail.com

Тузовский Анатолий Фёдорович
Д-р техн. наук, проф. каф. оптимизации систем управления Института кибернетики НИТПУ
Тел.: (382-2) 42-14-85
Эл. почта: tomo@tpu.ru

Gubin M.Y., Razin V.V., Tuzovsky A.F.

Method of creation of semantic document metadata using semantic networks, frame models and frequency characteristics

An approach to creation of semantic document metadata using semantic networks, frame models and frequency characteristics of documents to be analyzed is stated.
Keywords: ontology, semantic network, frame, RDF.