Построение онтологических связей в области знаний на основании поиска и анализа текстовых ссылок

Д.С. Михеев

***Аннотация* – В статье рассматриваются основные аспекты контроля ссылочной связности множества электронных текстов на примере области знаний биомедицины. Анализируется сложность поиска вхождений одного термина в текст определения другого термина. Рассматриваются задачи по поиску вхождений потенциальных текстовых ссылок в текст определения того или иного термина. Указаны основные проблемы, с которыми предстоит столкнуться при поиске потенциальных текстовых ссылок: употребление терминов в различных синтаксических и грамматических формах (различные окончания, падежи, предлоги и т. д.), а также поиск терминов, наименование которых состоит из нескольких слов (поиск словосочетаний). Среди первоочередных задач выделяются следующие: определение основы слова в наименовании термина, поиск вхождений с учётом словосочетаний. Приводятся варианты решения данных задач при помощи современных методов обработки текстов: разбиение на n-граммы, стемминг, применение регулярных выражений. Предлагается идея создания программного средства, позволяющего обеспечить связывание конечного множества документов методом построения онтологических связей на основе анализа выявленных текстовых ссылок. Создание такого программного средства предоставит оператору инструмент способный повысить скорость обработки данных при разработке и ведении онтологий, обеспечит его необходимыми подсказками для принятия решений при построении онтологических связей.**

***Ключевые слова* – автоматизированные системы, информационные ресурсы, классификаторы, онтология, предметная область, моделирование предметной области, биомедицина, онтология предметной области, онтологические связи, текстовые ссылки, стемминг.**

I. ВВЕДЕНИЕ

В силу стремительного развития высокотехнологичных аналитических методов в последнее десятилетие, наблюдается стремительный рост объема информации во всех областях знаний по биологии и медицине. Одним из подходов, применяемых для интеграции данных и знаний, является использование онтологий. Онтологии призваны обеспечить эффективное использование накопленной и получаемой информации, что становится насущной проблемой современной науки. Современные онтологии, как правило, представляют собой словарь терминов-концептов, формирующих описание определенной области знаний. Кроме того, объекты в таких онтологиях связаны между собой различными связями. Например, наиболее известной и широко используемой биоонтологией является Генная онтология [1]. Каждый термин в этой онтологии связан с конкретными примерами – генами, определяющими ту или иную функцию, процесс или структуру. В результате, выбор определенного термина-концепта Генной Онтологии позволяет сразу же выйти на соответствующие гены и их продукты, обнаруженные в различных организмах. Онтологический анализ представляет собой высокоэффективную альтернативу стандартно применяемому поиску в различных поисковых системах, который зачастую приводит к избыточности информации.

Связывание объектов между собой – один из важнейших этапов создания онтологии. Применение методов онтологического анализа невозможно без достаточного связывания объектов онтологии между собой. В то же время этот процесс требует от специалиста широких знаний в предметной области разрабатываемой онтологии. Кроме того, с ростом количества объектов в онтологии неизбежно растёт время, необходимое для обеспечения связности. В данной статье предлагается метод предоставления помощи оператору в нахождении связей между объектами онтологии, основанный на поиске и анализе имеющихся в определениях терминов-концептов текстовых ссылок на другие термины-концепты. Термины в тексте могут быть употреблены в различном числе или падеже в зависимости от контекста, в составе словосочетаний и т.д. Очевидно, что необходимо осуществлять поиск не по точному наименованию термина, а с учётом различных языковых форм. Важные задачи, которые необходимо решить для обеспечения успешного процесса поиска текстовых ссылок:

– выделение неизменяемой части термина (основа слова, основы слов в словосочетании);

– поиск вхождений термина в текст статьи с учётом различных форм его изменяемой части (окончания).

Статья получена 30.10.2018 г.
Д.С. Михеев, МТУСИ (e-mail: shelbyterlingua1@gmail.com).

II. Методы выделения неизменяемой части слова

Основа слова – неизменяемая часть слова, которая выражает его лексическое значение. В изменяемых словах основа определяется как часть слова без окончания и формообразующего суффикса: здоров-ый, биолог-ия, медицин-ский, диагноз-ы, клеточн-ый, и т.д. В неизменяемых словах основа равна слову [2].

Одним из способов решения задачи выделения основы слова для последующего его использования при поиске терминов в тексте является применение регулярных выражений.

Для примера составим регулярное выражение, убирающее из слова некоторые из окончаний: -а, - ый, - ых, -ые. Регулярное выражение будет иметь следующий вид:

/(а|ые|ых|ый)\b/

Приведённое регулярное выражение осуществит поиск заданных окончаний в строке, после чего, в зависимости от задачи, их можно изменить, заменить или удалить для получения той или иной формы.

III. Стемминг

Стемминг представляет собой один из множества механизмов поискового алгоритма, который служит для выделения основы слова из его словоформ.

Благодаря стеммингу пользователь имеет возможность просматривать не только те документы, в тексте которых присутствует прямое вхождение ключевого слова (лемма), но и другие – в которых присутствует его словоформа.

Существует несколько типов алгоритмов стемминга каждый из которых имеет свою точность и производительность. Рассмотрим следующие типы стемминга: поиск по таблице флексий, алгоритмы усечения окончания.

A. Поиск по таблице флексий

Флексия – это совокупность морфем, выполняющих словообразование, например, окончаний. Так, можно в таблицы флексий хранить все возможные окончания слов и даже генерировать новые словоформы. В ходе выполнения данного алгоритма ведется простой поиск по таблице флексий.

К недостаткам можно отнести то, что все флективные формы должны быть явно перечислены в таблице: новые или незнакомые слова не будут обрабатываться, даже если они являются правильными (например, iPads ~ iPad), следовательно, такой подход применим не для всех языков.

B. Алгоритмы усечения окончаний

Алгоритмы усечения окончаний не имеют в своем распоряжении таблиц всех флективных форм, а часто используют для своей работы небольшой набор правил, вроде: если слово оканчивается на “ет”, то удалить “ет” и так далее.

Некоторые примеры правил применительно для английского языка выглядят следующим образом:

– если слово оканчивается на 'ed', удалить 'ed';

– если слово оканчивается на 'ing', удалить 'ing';

– если слово оканчивается на 'ly', удалить 'ly'.

Одним из недостатков алгоритма может быть то, что разработчик должен хорошо знать лингвистику языка.

Самым известным и распространенным на сегодняшний день является стеммер Портера. Основная идея стеммера Портера заключается в том, что существует ограниченное количество словообразующих суффиксов, и стемминг слова происходит без использования каких-либо баз основ: только множество существующих суффиксов и вручную заданные правила.

Алгоритм состоит из пяти шагов. На каждом шаге отсекается словообразующий суффикс, и оставшаяся часть проверяется на соответствие правилам (например, для русских слов основа должна содержать не менее одной гласной). Если полученное слово удовлетворяет правилам, происходит переход на следующий шаг. Если нет – алгоритм выбирает другой суффикс для отсечения. На первом шаге отсекается максимальный формообразующий суффикс, на втором – буква «и», на третьем – словообразующий суффикс, на четвертом – суффиксы превосходных форм, «ь» и одна из двух «н» [3].

Несмотря на свою распространенность, стеммер Портера не лишен недостатков. Данный алгоритм часто обрезает слово больше необходимого, что затрудняет получение правильной основы слова, например кровать->крова (при этом реально неизменяемая часть – кроват, но стеммер выбирает для удаления наиболее длинную морфему). Также стеммер Портера не справляется со всевозможными изменениями корня слова (например, выпадающие и беглые гласные).

C. Поиск вхождений слова в текст определения

Термины-концепты онтологии могут встречаться в тексте определений не только в различных словаформах, но и как словосочетания из двух или нескольких слов. В таком случае необходимо выделять такие словосочетания, последовательно разбивая текст на N-граммы.

N-грамма – последовательность из n элементов [4]. С семантической точки зрения, это может быть последовательность звуков, слогов, слов или букв. На практике чаще встречается N-грамма как ряд слов, устойчивые словосочетания называют коллокацией.

В области обработки естественного языка N-граммы используется в основном для предугадывания на основе вероятностных моделей. N-граммная модель рассчитывает вероятность последнего слова N-граммы, если известны все предыдущие. При использовании этого подхода для моделирования языка предполагается, что появление каждого слова зависит только от предыдущих слов [5].

Другим применением N-грамм является выявление плагиата либо близости двух текстов. Если разделить текст на несколько небольших фрагментов, представленных N-граммами, их легко сравнить друг с другом и таким образом получить степень сходства

анализируемых документов [6]. N-граммы часто успешно используются для категоризации текста и языка. Данное свойство N-Грамм также позволяет использовать их для помощи выбора оптимальной ссылки из двух на термины-синонимы.

IV. Программное средство создания онтологических связей на основе анализа текстовых ссылок между терминами-концептами

Итак, применяя описанные выше подходы и алгоритмы, предлагается создать программное средство, предоставляющее помощь оператору (автору онтологии) в создании онтологических связей между терминами-концептами. Алгоритм работы программы можно разделить на этапы, перечисленные ниже.

Этап первый: приведение всех терминов-концептов к неизменяемой форме. Применяя метод стемминга или регулярные выражения, производится выявление основ слов в терминах, создаётся совокупность «урезанных» терминов. На первый взгляд результат представляется похожим на таблицу флексий, но это верно лишь от части. Если в таблице флексий должны храниться все словоформы, до в данной совокупности будут храниться только по одной форме для каждого термина-концепта.

Этап второй: разбиение текста на N-граммы. Обрабатываемый оператором термин последовательно разбивается на N-граммы разной длинны (от большой длины N-граммы, которая может быть задана оператором, до N-грамм из одного слова). После этого каждая N-грамма также подвергается выделению основных частей слов, входящих в неё.

Этап третий: поиск и расстановка ссылок. Осуществляется сравнение каждой из полученных на предыдущем этапе N-грамм с совокупностью полученных на первом этапе «урезанных» терминов. Если поиск даёт результат, программа предлагает оператору вставить ссылку на термин-концепт в соответствующем месте текста и создать онтологическую связь между двумя объектами.

Таким образом, при работе с биомедицинской онтологией, оператор получает инструмент, который предоставляет помощь в выявлении связей между терминами-концептами. Подсказки в виде текстовых ссылок, получаемые оператором в ходе работы с данным программным средством, предназначены для помощи при принятии решения о необходимости создания той или иной онтологической связи.

Заключение

Предлагаемое программное средство осуществляет поиск вхождений терминов с учетом различных словоформ и словосочетаний. Оператор получает инструмент, позволяющий осуществлять построение онтологических связей в той или иной области знаний на основании поиска и анализа текстовых ссылок.

Автоматизированный поиск ссылок имеет широкую область применения и может быть использован в автоматизированных системах, работающих с онтологическими моделями данных. На его основе можно реализовать механизмы поиска, обмена и управления ресурсами в системе. Предлагаемый метод поиска потенциальных онтологических связей полезен для применения в системах разработки информационного обеспечения, позволяя, за счёт автоматизации процесса, обеспечивать оператора необходимыми подсказками и ускорять его работу

Библиография

[1] Gene-Ontology-Consortium. Creating the Gene Ontology Resource: Design and Implementation // Genome Res. - 2001. - V. 11. - P. 1425-1433.

[2] Огекян И. Н., Волчек Н. М., Высоцкая Е. В. и др. «Большой справочник: Весь русский язык. Вся русская литература» – Мн.: Изд-во Современный литератор, 2003. – 992 с.

[3] Russian stemming algorythm, URL: http://snowball.tartarus.org/algorithms/russian/stemmer.html (дата обращения: 25.08.18).

[4] Proceedings of the 7th Annual Conference ZNALOSTI 2008, Bratislava, Slovakia, pp. 54-65, February 2008. ISBN 978-80-227-2827-0.

[5] Jurafsky, D. and Martin, J.H. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. – Pearson Prentice Hall, 2009. – 988 p. – ISBN 9780131873216.

Building ontological knowledge links based on searching and analyzing text links

D.S. Miheev

Abstract **– The article discusses the main aspects of controlling the referential connectivity of a set of electronic texts on the example of the field of knowledge of biomedicine. The complexity of finding the occurrences of one term in the text of the definition of another term is analyzed. Considered the problem of finding the occurrences of potential text links in the text of the definition of a particular term. The main problems are indicated: the use of terms in various syntactic and grammatical forms (various endings, cases, prepositions), as well as the search for terms whose name consists of several words (search for phrases). Among the priorities are the following: the definition of the basis of the word in the name of the term, the search for occurrences, taking into account phrases. There are options for solving these problems with the help of modern text processing methods: splitting into n-grams, stemming, the use of regular expressions. The author proposes the idea of creating a software tool that allows for the binding of a finite set of documents by the method of building ontological links based on the analysis of the identified text links.**

Keywords **– automated system, information resource, classifier, ontology, domain knowledge, knowledge representation model.**