ВЫЯВЛЕНИЕ СЕМАНТИЧЕСКИХ ЭКВИВАЛЕНТОВ ПРИ АВТОМАТИЧЕСКОЙ ОБРАБОТКЕ ТЕКСТОВ

Петрасова С.В.
Национальный технический университет "Харьковский политехнический институт",
г. Харьков, ул. Пушкинская, 79/2, тел. 707–63–60,
e-mail: svetapetrasova@gmail.com

Целью работы является анализ существующих методов выявления семантических эквивалентов и выбор метода для решения задачи их автоматического определения в заданном текстовом ресурсе экономической направленности.

В настоящее время существует множество методик информационного поиска. Все они могут быть поделены на три большие группы: статистические методы поиска, методы поиска по семантическим сетям и комбинированные методы поиска.

Семантический поиск это метод поиска, в котором релевантность документа запросу определяется с использованием семантических, а не статистических методов, как происходит в подходах информационного поиска по ключевым словам.

Семантические эквиваленты – текстовые выражения, сопоставленные одному и тому же понятию. Семантическими эквивалентами являются синонимы и семантические близкие слова. Под семантически близкими словами подразумеваются слова с близким значением, встречающиеся в одном контексте.

В качестве базовых знаний для автоматической семантической обработки, обычно используются онтологии, которые представляют собой формальные описание терминов предметной области и отношений между ними, и тезаурус, как особая разновидность словарей общей или специальной лексики, в которых указаны семантические отношения (синонимы, антонимы, паронимы, гипонимы, гиперонимы и т.п.) между лексическими единицами. Таким образом, тезаурусы, особенно в электронном формате, являются одним из действенных инструментов для описания отдельных предметных областей. Благодаря использованию онтологий и тезаурусов удается строить образ достаточно релевантный запрашиваемому. Этот образ может использоваться для формирования более эффективных запросов для поисковой системы.

Из статистических моделей поиска остановимся на следующих. В "моделях векторных пространств" формируются векторные представления слов и других компонент текстов путем автоматического извлечения статистики их совместной встречаемости из больших массивов текстовой информации.

В основе метода латентно-семантического анализа лежит гипотеза о том, что между отдельными словами и обобщенным контекстом (предложениями, абзацами и целыми текстами), в которых они встречаются, существуют

неявные (латентные) взаимосвязи, обуславливающие совокупность взаимных ограничений.

Метод Клейнберга (HITS алгоритм) использует понятия авторитетного и хаб-документа и заключается в анализе ссылок, что позволяет ранжировать веб-страницы. Авторитетный документ – это документ, соответствующий запросу пользователя, имеющий больший удельный вес среди документов данной тематики. Хаб-документ – это документ, содержащий ссылки на авторитетные документы. Метод основан на вычислении собственного вектора матрицы, описывающей структуру ссылок в вебе.

Наиболее перспективной является группа методов, которые объединяет качественную статистическую модель поиска и учет семантических конструкций. К этой группе можно отнести метод расстояний, который положен в основу исследования. Он использует в качестве лингвистического ресурса толковый словарь, который позволяет дать количественную оценку семантической близости между терминами словаря. Суть метода расстояний заключается в том, что два слова считаются синонимами, если имеют общие слова в своих определениях (понятиях).

Для реализации метода осуществляется следующий алгоритм: выполняется предварительная лингвистическая обработка; создается лингвистическая база данных, состоящая из терминов и их определений; после введения пользователем запроса проводится попарное сравнение термина запроса с каждым термином словаря; степень семантической близости определяется расстоянием между двумя терминами, которое представлено в виде суммы количества компонент в определении первого термина, отсутствующих в определении второго, и количества компонент в определении второго термина, отсутствующих в определении первого. Величину расстояния нормализуем путем отношения полученной суммы необщих компонентов к сумме всех компонент первого и второго термина. В результате данного сравнения пользователь получает набор терминов, имеющих общие компоненты с термином запроса.

Предлагаемое данным методом решение задачи автоматического выявления семантических эквивалентов может использоваться в поисковых системах для расширения запроса, для автоматизированного построения онтологии по тексту, для расширения существующих и создания новых тезаурусов.