ПРИМЕНЕНИЕ СЕМАНТИЧЕСКОЙ СЕТИ ДЛЯ ПОИСКА ИНФОРМАЦИИ В ТЕКСТЕ

Потараев В.В., Серебряная Л.В.

Белорусский государственный университет информатики и радиоэлектроники, г. Минск, Беларусь, vic229@rambler.ru

Abstract. Text search is usually used by students when preparing to classes or an exam. It can be performed more effectively using search query semantics instead of simple textual match. The article describes usage of semantic network to expand search query and improve search.

При обучении студенты используют достаточно большое количество методических материалов и учебников, представленных в электронном виде. Во время подготовки к экзаменам, а также при решении различных задач обучаемые сталкиваются с необходимостью найти информацию в электронном документе по определённому вопросу. При этом поиск, как правило, осуществляется по полному совпадению текста, то есть нужно ввести искомый фрагмент именно в том виде, в каком он представлен в тексте. Синонимы, связанные по смыслу слова и различные формы слов не обнаруживаются. Но вероятной является ситуация, при которой студент не помнит точно все слова нужного фрагмента текста – например, помнит лишь несколько отдельных слов или хочет найти название некоторого понятия, определение которого он помнит.

Семантический анализ информации является довольно эффективным способом обработки данных, учитывающим их смысловую структуру.

Целью данной работы является разработка методов поиска информации в тексте, основанных на семантическом анализе. Поиск смысловой связи между словами запроса должен помочь вернуть наиболее соответствующий запросу фрагмент текста.

Одним из инструментов для осуществления семантического анализа данных являются базы знаний.

База знаний – это компонент экспертной системы, предназначенный для хранения долгосрочных данных, описывающих определенную предметную область, и правил, описывающих целесообразные преобразования данных этой области [1].

Существует четыре основных модели базы знаний:

1. Логическая модель (основанная на формулах).
2. Продукционная модель (основанная на правилах).
3. Фреймы (фрейм – это минимально возможное описание сущности объекта).
4. Семантическая сеть (ориентированный граф, отражающий понятия и их отношения) [2].

Рассмотрим возможность применения семантической сети для решения задачи поиска информации в тексте. Модель, основанная на семантической сети, наиболее соответствует современным представлениям об организации долговременной памяти человека [3]. Это её свойство может оказаться полезным при решении данной задачи.

Количество типов отношений в семантической сети определяется её создателем, исходя из конкретных целей. В реальном мире их число стремится к бесконечности. Каждое отношение является, по сути, предикатом (утверждением), простым или составным [4].

Выберем для рассмотрения модель семантической сети, в которой узлами являются слова текста. В качестве связей можно использовать следующие типы отношений: «действие», «объект действия», «принадлежность», «синонимичность», «признак предмета». Построение такой сети может быть автоматизировано. Например, синонимы могут быть добавлены в сеть при помощи использования словаря синонимов. Схожая модель семантической сети может быть использована для ответа на вопрос [5]. На каждом этапе работы с текстом имеет смысл учитывать некоторую начальную форму слова (полученную, например, с помощью некоторого алгоритма стемминга).

Слова запроса обычно связаны между собой по смыслу. В семантической сети связанные по смыслу слова в большинстве случаев будут связаны между собой через промежуточные узлы.

Поисковый запрос может не содержать слов текста, но содержать слова, связанные с ними семантически. В таком случае имеет смысл добавить в поисковый запрос слова, которые связаны по смыслу со словами текста. Добавление слов может быть реализовано с использованием семантической сети.

Предположим, что в тексте есть два предложения, разделённые другими предложениями:

«Маркс выделял стадии развития общества» и «каждая стадия отличается формами собственности».

Для текста, содержащего эти предложения, семантическая сеть будет содержать фрагмент, представленный на рисунке 1.

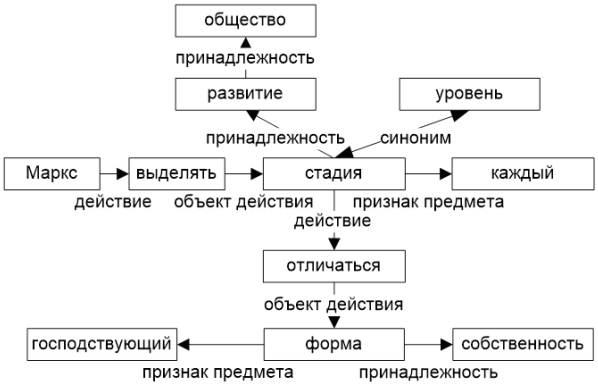

Рисунок 1 – Пример семантической сети

Предположим, пользователь ввёл для поиска слова «уровень собственности». Слово «собственность» встречается во втором предложении, в то время как «уровень» – ни в одном. То есть простой анализ наличия слов в предложениях позволит обнаружить только второе предложение как содержащее лишь одно слово запроса.

Покажем, что использование семантической сети позволяет найти больше предложений, соответствующих запросу. Найдём промежуточные узлы, связывающие данные понятия, и добавим их в поисковый запрос. Получим следующий запрос: «уровень, стадия, отличаться, форма, собственность».

В предложении «каждая стадия отличается формами собственности» содержатся четыре слова из пяти слов дополненного запроса. Значит, это предложение является подходящим. В предложении «Маркс выделял стадии развития общества» есть одно слово дополненного запроса. Оно также может быть отображено пользователю в ответ на запрос.

Таким образом, использование семантической сети в данном случае позволяет найти больше предложений, связанных по смыслу с запросом. Наличие связи «синоним» оказалось весьма полезным. Кроме того, возможна ситуация, при которой пользователь ввёл в запрос слова, отсутствующие в тексте. В таком случае семантическая сеть также способна найти соответствующий фрагмент текста, используя связи, которых нет в тексте.

Рассмотрим ещё один пример. Пользователь ввёл для поиска слова «развитие собственности». Каждое из этих слов по одному разу встречается в каждом из предложений. Простой анализ наличия слов в предложениях не позволит выделить более подходящее предложение текста. Найдём промежуточные узлы, связывающие данные понятия, и добавим их в поисковый запрос. Получим следующий запрос: «развитие, стадия, отличаться, форма, собственность».

В предложении «каждая стадия отличается формами собственности» больше слов, относящихся к дополненному запросу. Значит, оно является более подходящим.

Таким образом, использование семантической сети в данном случае позволяет выбрать предложение, содержащее больше слов, связанных с запросом семантически.

Для дополнения запроса можно использовать другие способы – например, добавлять слова из словаря синонимов, не используя сеть. Но семантическая сеть является моделью, которая собирает в себе различные виды отношений, полученные разными способами. Если для построения сети использовать дополнительные тексты, то она будет содержать больше смысловых связей и обладать более высокой способностью к поиску связанных по смыслу слов. Это может позволить сети находить предложения в тексте даже в случаях, когда ни одно из слов запроса не встречается в тексте.

Рассмотренный алгоритм обработки поискового запроса может быть представлен следующим образом:

1. Построить семантическую сеть для текста или множества текстов.

2. Найти фрагменты сети, соединяющие пары слов исходного запроса.

3. Добавить в поисковый запрос слова из найденных фрагментов сетей.

4. Выбрать предложения, содержащие наибольшее число слов дополненного запроса.

Данный алгоритм может быть реализован различными способами. Так, например, семантическая сеть может быть реализована с использованием различных наборов отношений.

Если в сети нет фрагментов, соединяющих пары слов исходного запроса, то он может быть дополнен словами, находящимися в сети на определённом удалении от слов запроса. Поиск пути между узлами сети может быть реализован с использованием алгоритма поиска в ширину, поиска в глубину или другими способами.

В поисковый запрос можно добавлять только те слова сети, которые расположены в ней на определённом отдалении от слов запроса. На этапе выбора предложений, соответствующих запросу, имеет смысл назначить различным словам различный вес. Например, слова, содержащиеся в исходном запросе, могут иметь наибольший вес. Если остальным словам присвоить нулевой вес, то получится результат, аналогичный обычному поиску без использования семантической сети и без дополнения запроса.

Итак, рассмотренный алгоритм поиска информации в тексте, основанный на семантической сети, учитывает смысл слов. Это позволяет находить больше фрагментов текста, связанных по смыслу с запросом, даже в случаях, когда фрагмент текста не содержит слов запроса, но содержит слова, связанные с ними по смыслу. Кроме того, добавление дополнительных слов в поисковый запрос позволяет более точно упорядочивать найденные фрагменты текста, используя большее количество искомых слов.

Литература

1. Базы знаний экспертных систем. [Электронный ресурс]. – Электронные данные. – Режим доступа: http://daxnow.narod.ru/index/0-18. Дата доступа: 17.11.2017.

2. Гаврилова, Т. А. Базы знаний интеллектуальных систем / Т. А. Гаврилова, В. Ф. Хорошевский – СПб. : Питер, 2000. – 384 с.

3. Масленникова, О. Е. Основы искусственного интеллекта [Электронный ресурс] : учеб. пособие / О. Е. Масленникова, И. В. Гаврилова. – М.: ФЛИНТА, 2013. – Режим доступа: http://search.rsl.ru/ru/record/01007574162.

4. Рахимова Д. Р. Построение семантических отношений в машинном переводе // Вестник КазНУ им. аль-Фараби. Серия «Математика, механика и информатика». – Алматы, 2014. – №1. – С.90-101.

5. Потараев, В.В. Алгоритм построения семантической сети и её применение / В.В. Потараев // Информационные технологии и системы 2017 (ИТС 2017): материалы международной научной конференции. Минск : БГУИР, 2017. – С.144-145.