ПОСТРОЕНИЕ ТЕЗАУРУСА ПРИ СЕМАНТИЧЕСКОМ АНАЛИЗЕ ТЕКСТОВОЙ ИНФОРМАЦИИ

Михайлова А.С., аспирант факультета прикладной математики-процессов управления СПбГУ

Аннотация

В статье рассматривается метод автоматизированного построения тезауруса, который основан на онтологической модели представления текста, имеющей количественную характеристику связей между элементами.

Введение

Семантический анализ текстовой информации дает возможность сопоставить этой информации предельно допустимое количество слов, которые могут кратко определить смысл содержания. Данные слова называют <терминами>, <метками>, <ключевыми словами>, <словами – определениями>. В информационной архитектуре – это <метаданные>.

В связи с тем, что ручной метод структурирования информации является достаточно трудоемким использование автоматизированного установления семантических связей между словами (терминами) значительно упрощает структуризацию текста, при этом связанные определенным отношением термины образуют тематические блоки. Таким образом, частично структурированный текст преобразуется в тезаурус.

Применение тезаурусов является классическим методом в задачах информационного поиска [1].

Целью работы является разработка метода построения тезауруса, основанного на онтологической модели представления текста, имеющей количественную характеристику связей между элементами. Новизна модели онтологического представления текстовой информации заключается в том, что числовые значения связей между единицами текста вычисляются с помощью коэффициентов корреляции.

Метод построения тезауруса

Тезаурус [2] - это:

1. Словарь, в котором максимально полно представлены все слова языка с исчерпывающим перечнем примеров их употребления в текстах.

2. Идеологический словарь, в котором показаны семантические отношения (родовидовые, синонимические и др.) между лексическими единицами.

Моделью тезауруса служит семантическая сеть.

Семантическая сеть – информационная модель предметной области, имеющая вид ориентированного графа, вершины которого соответствуют объектам предметной области, а дуги (рёбра) задают отношения между ними.

Связи между элементами в тезаурусе отображаются в виде семантической сети (ориентированного графа), в котором все слова и устойчивые словосочетания связаны определенной связью, имеют иерархию и вес.

Исходными данными для формирования тезауруса являются тексты. Текст – в общем плане связная и полная последовательность знаков.

С понятием тезаурус связано понятие онтология. Онтология - это описание перехода от неформального представления знаний о предметной области к формальному представлению.

Онтологическая модель представления текстовой информации, имеющая количественную характеристику связей между элементами, представляет собой тройку:

$$MST = \{W, Df, R\} \quad (1)$$

где W – множество слов и устойчивых словосочетаний текста,

Df – определение элемента из множества W ,

R – множество бинарных отношений между элементами множества W, имеющие количественную характеристику близости элементов в отрезке [0;1].

Для того чтобы получить первый параметр W в (1), нужно:

- во-первых, произвести первоначальную обработку текста, разбив текст на множество слов;
- во-вторых, удалить <ненужные> части речи: имена числительные, местоимения, союзы, предлоги, частицы и междометия;
- в-третьих, произвести морфологический анализ этих слов.

Целью и результатом морфологического анализа является нахождение морфологических характеристик слов и их основных словоформ

(канонических или нормальных). Перечень всех морфологических характеристик слов и допустимых значений зависят от языка.

Так, например, в русском языке начальными словоформами являются: для имен существительных – именительный падеж, единственное число; для имен прилагательных – именительный падеж, единственное число, мужской род; для глаголов, причастий и деепричастий – глагол в форме инфинитива [3].

Для установления связей между словами и устойчивыми словосочетаниями текстовой информации используется корреляционный анализ. Вычисление коэффициентов корреляции для каждой пары слов производится по формуле [4]:

$$r = \frac{\sum (x_{1,i} - \bar{x}_1) \cdot (x_{2,i} - \bar{x}_2)}{\sqrt{\sum (x_{1,i} - \bar{x}_1)} \cdot \sqrt{\sum (x_{2,i} - \bar{x}_2)}}$$

где $\bar{x}_1 = \frac{\sum x_{1,i}}{n}$, $\bar{x}_2 = \frac{\sum x_{2,i}}{n}$ – средние значения для каждого параметра массива из точек.

Коэффициент корреляции r отображает степень статистической зависимости между двумя числовыми переменными (в данном случае между словами), $r \in [-1; 1]$.

При $r = 1$ корреляция считается положительной, при $r = -1$ корреляция считается отрицательной. При $r=0$ слова независимы друг от друга.

В Таблице 1 представлены виды корреляции в соответствии со значением коэффициента корреляции

Значение	**Интерпретация**
до 0,2	Очень слабая корреляция
до 0,5	Слабая корреляция
до 0,7	Средняя корреляция
до 0,9	Высокая корреляция
свыше 0,9	Очень высокая корреляция

Таблица 1: виды корреляции в соответствии со значением r

После подсчета всех коэффициентов корреляции строится график,

ранжируются связи терминов по заданному значению коэффициента корреляции, и удаляется информационный шум. Из оставшихся значимых пар слов формируется семантическая сеть.

Добавление в семантическую сеть определений терминов преобразует её в тезаурус текстовой информации [5].

Таким образом, алгоритм построения тезауруса выглядит следующим образом:

1. Первоначальная обработка текста и выделение множества слов.
2. Удаление из полученного множества слов имен числительных, местоимений, союзов, предлогов, частиц и междометий.
3. Морфологический анализ каждого элемента из множества слов.
4. Вычисление коэффициентов корреляции для всех пар слов.
5. Ранжирование связей по заданному значению коэффициента корреляции и выделение устойчивых словосочетаний. В итоге сформированы массивы терминов и связей.
6. На основе полученных данных формируется семантическая сеть.
7. Сформированные массивы терминов и связей дополняются определениями терминов, и семантическая сеть преобразуется в тезаурус.

Заключение

Таким образом, построение тезауруса при семантическом анализе текстовой информации основано на онтологической модели представления текста и использует корреляционный анализ для выявления связей между словами и устойчивыми словосочетаниями текста.

Литература

1. Башмаков А. И., Башмаков И. А. Интеллектуальные информационные технологии: учебное пособие. М.: Изд-во МГТУ им. Н.Э. Баумана, 2005. 304 с.
2. Шайкевич А. Я. Введение в лингвистику. М.: Академия, 2005. 400 с.
3. Тузов В. А. Компьютерная семантика русского языка. СПб.: Изд-во СПбГУ, 2004. 400 с.
4. Гмурман В. Е. Теория вероятностей и математическая статистика: Учебное пособие для вузов. 10-е издание, стереотипное. М.: Высшая школа, 2004. 479 с.
5. Лагутина Н. С., Лагутина К. В., Адрианов А. С., Парамонов И. В. Русскоязычные тезаурусы: автоматизированное построение и

применение в задачах обработки текстов на естественном языке // Моделирование и анализ информационных систем. Т. 25. No 4. 2018, С. 435–458.