Разработка автоматизированной системы семантического анализа текстовой информации

О.А. Черненко[a], О.А. Гордеева[a]

[a] *Самарский национальный исследовательский университет имени академика С.П. Королева, 443086, Московское шоссе, 34,Самара, Россия*

Аннотация

В данной статье рассматриваются аспекты применения основных методов семантического анализа текстовой информации – стеммера Портера, частотно-семантического, латентно-семантического и синтаксико-семантического анализа. Разработанная автоматизированная система позволяет анализировать текст с использованием указанных методов. Рассмотрены характерные особенности методов, получены результаты их применения к текстам небольшой сложности. Проведенное исследование позволяет выявить особенности использования указанных методов в соответствии с целями анализа текста.

Ключевые слова: текст; стемминг Портера; анализ текста; синтаксико-семантический анализ; частотно-семантический анализ; латентно-семантический анализ; классификация текстов

1. Введение

В настоящее время сложно представить себе эффективную работу с текстовой информацией без использования компьютерной обработки. Одним из наиболее актуальных и постоянно развивающихся видов обработки текста является семантический анализ. В зависимости от поставленных в программной системе критериев, может быть выбран один из нескольких видов семантического анализа, удовлетворяющий этим критериям. Например, если речь идет о поисковом аудите сайта, то критериями выбора метода семантического анализа будут быстрота работы, минимальный объем словаря или его отсутствие. В случае подбора метода анализа для художественных произведений со сложными речевыми оборотами, главным критерием будет являться качество обработки. Соответственно, алгоритм семантического анализа должен выдавать результаты, максимально приближенные к человеческим, и такие параметры как быстрота и объем используемых библиотек не будут играть решающей роли.

2. Постановка задачи

Объект исследования представляет собой текст на русском языке, размером не более 20 предложений и однозначно трактуемой для человеческого понимания темой. Цель исследования – на основе разработанной системы семантического анализа текстовой информации проанализировать работу четырех выбранных методов анализа, сравнить такие характеристики методов, например, как эффективность и скорость анализа.

3. Методы семантического анализа текстов

Всю совокупность представленных на сегодняшний день методов анализа текста можно разделить на две группы:

- лингвистический анализ – основан на извлечении смысла текста по его семантической структуре;
- статистический анализ – основан на извлечении смысла текста по частотному распределению слов в тексте.

Деление на группы условное, так как в реальных задачах и при решении проблем всегда используется сочетание методов для достижения определенного результата.

В данной работе рассмотрены алгоритмы семантического анализа из обеих групп, наиболее часто применяемые на практике.

3.1. Частотно-семантический анализ

Метод частотно-семантического анализа (ЧСА) основан на подсчете частоты встречаемости слов в тексте. Для корректной работы алгоритма вводится несколько уточнений [1]:

- поскольку не всякое слово в тексте может являться темой или ядром текста, в качестве единиц подсчета будут учитываться только существительные;
- для определения в тексте существительных, необходимо использовать словарь.

Алгоритм работает таким образом: все слова текста сравниваются со словарем, совпавшие заносятся в массив, и далее сравниваются по числу вхождений. Слова с самым большим числом вхождений будет темой текста.

3.2. Алгоритм на основе стеммера Портера

Стемминг – отсечение от слова окончаний и суффиксов, чтобы оставшаяся часть являлась основой для всех грамматических форм слова. Стеммер Портера – алгоритм стемминга, в результате работы которого от исходного слова

находится основа. Стеммер может работать только с языками, которые реализуют словоизменение через аффиксы, примерами таких языков являются русский и английский. Основное преимущество данного алгоритма в отсутствии словаря.

Вначале вводятся несколько понятий о частях стемматизируемого слова:

- RV – область слова после первой гласной. Она может быть пустой, если гласные в слове отсутствуют;
- R1 – область слова после первого сочетания «гласная-согласная»;
- R2 – подобласть R1 после первого сочетания «гласная-согласная».

Портер в своей статье [2] приводит алгоритм стемматизации слова, состоящий из отсечения приставок, окончаний и суффиксов:

- если в слове есть окончание деепричастия, то удалить его. Иначе, удаляем окончания «ся» или «сь», если они существуют. Далее ищем окончания прилагательных, глаголов и существительных, как только одно из них найдено – оно удаляется;
- ищем окончание «и», если найдено – удаляем его;
- ищем окончания «ост» или «ость», если одно из них найдено – оно удаляется;
- если слово оканчивается на «нн» – удаляем последнюю букву;
- если слово оканчивается на «ейш» или «ейше» – удаляем его и снова удаляем последнюю букву, если слово оканчивается на «нн»;
- если слово оканчивается на «ь» – удаляем его;

Для определения тематики текста с помощью алгоритма на основе стемминга Портера, необходимо провести стемматизацию всех слов анализируемого текста. В результате будет получен массив основ слов. Слова текста, являющиеся производными от основы с самым частым числом вхождений и будут являться тематикой текста.

3.3. Латентно-семантический анализ

Латентно-семантический анализ (ЛСА) – это метод обработки информации на естественном языке, анализирующий взаимосвязь между коллекцией документов и терминами, в них имеющимися, сопоставляющий некоторые факторы (тематики) всем документам и терминам. В основе метода латентно-семантического анализа лежат принципы факторного анализа. В качестве входной информации ЛСА использует матрицу термы-на-документы (термы – слова или словосочетания) [3]. Элементы этой матрицы содержат веса, учитывающие частоты использования каждого терма в каждом документе. Наиболее распространенный вариант ЛСА основан на использовании разложения диагональной матрицы по сингулярным значениям (SVD – Singular Value Decomposition). С помощью SVD-разложения любая матрица раскладывается во множество ортогональных матриц, линейная комбинация которых является достаточно точным приближением к исходной матрице.

Говоря более формально, согласно теореме о сингулярном разложении, любая вещественная прямоугольная матрица может быть разложена на произведение трех матриц:

$$A=\mathrm{USV}^{\mathrm{T}},$$

где матрицы U и V – ортогональные, а S – диагональная матрица, значения на диагонали которой называются сингулярными значениями матрицы A. Буква T в выражении VT означает транспонирование матрицы.

Такое разложение обладает примечательной особенностью: если в матрице S оставить только k наибольших сингулярных значений, а в матрицах U и V – только соответствующие этим значениям столбцы, то произведение получившихся матриц S, V и U будет наилучшим приближением исходной матрицы A к матрице Â ранга k:

$$\hat{A} \approx A=\mathrm{USV}^{\mathrm{T}}.$$

Основная идея латентно-семантического анализа состоит в том, что если в качестве матрицы A применялась матрица термы-на-документы, то матрица Â, содержащая лишь k первых линейно независимых компонент A, отображает основную структуру различных зависимостей, присутствующих в исходной матрице. Исходя из этого анализируется зависимость между термами и документами из разложения и определяется тематика текста.

3.4. Синтаксико-семантический анализ

Синтаксико-семантический анализ – метод обработки текстовой информации, который заключается в формировании шаблонов для сравнения со словами текста, в результате которого для каждого предложения создается список, состоящий из пар [4]:

- тип слова в предложении;
- позиция главного слова для данного зависящего.

Предполагается, что базовые шаблоны формируются из наиболее важных и часто используемых семантических отношений в тексте. Базовым семантическим шаблоном назовем правило, по которому в анализируемом тексте находится семантическое отношение. Базовый семантический шаблон состоит из 4 основных частей:

- последовательность слов или неделимых смысловых единиц, для которых указаны их морфологические признаки;

– название семантического отношения, которое должно быть сформировано в случае обнаружения в тексте последовательности, описанной в предыдущем пункте;
– последовательность чисел, определяющая позиции в последовательности, элементы которой должны быть добавлены в очередь с приоритетом, в соответствии с которой впоследствии будут удаляться слова из анализируемого предложения, подаваемого на вход семантическому анализатору;
– число, обозначающее значение приоритета, группы семантических зависимостей, к которой относится данное семантическое отношение.

С использованием базовых семантических шаблонов производится составление очереди с приоритетом. Очередь с приоритетом используется для хранения слов, являющихся правым аргументом некоторой семантической связи, найденной в анализируемом предложении.

Для определения тематики текста из каждого предложения, исходя из очереди с приоритетом, выбирается слово с наибольшим числом зависимостей и считается число его вхождений в текст. Слово с максимальным числом вхождений и есть тематика текста.

4. Работа системы

Для проведения исследований результатов применения описанных методов анализа текстов была разработана автоматизированная система. На начальном этапе работы система разбивает текст на слова или предложения, в зависимости от алгоритма, выбранного пользователем, и отправляет их на обработку.

Если был выбран частотно-семантический анализ, система сравнивает слова из текста со словами из словаря и находит среди них слова с максимальным числом вхождений в текст. Далее выводит результат нахождения ядра текста и список слов, не найденных в словаре, которые можно внести в словарь и запустить алгоритм заново.

Если был выбран алгоритм на основе стеммера Портера, система стемматизирует исходные слова и ищет среди них наиболее часто встречающиеся. Таким образом формируется ядро текста в данном алгоритме.

При выборе латентно-семантического анализа система составляет матрицу слов-на-предложения из предложений текста и производит с ней SVD преобразование. Далее используются только первые два столбца получившихся матриц. Из первых двух столбцов матрицы V^T, соответствующей предложениям, выбирается максимум и минимум, что соответствует максимальному и минимальному x и y на координатной плоскости. Таким образом обозначается область, вхождение в которую для точек из первых двух столбцов матрицы U, соответствующей словам, означает включение в ядро текста.

При выборе синтаксико-семантического анализа в каждом предложении слова проверяются на соответствие шаблонам, после чего каждому присваивается определенный вес, в зависимости от шаблона. Чем больше у слова зависимых слов, тем меньше вес и выше приоритет. Далее, в каждом предложении ищется слово с минимальным весом, самые часто встречающиеся формируют ядро текста.

На рисунке 1 представлена основная форма программы.

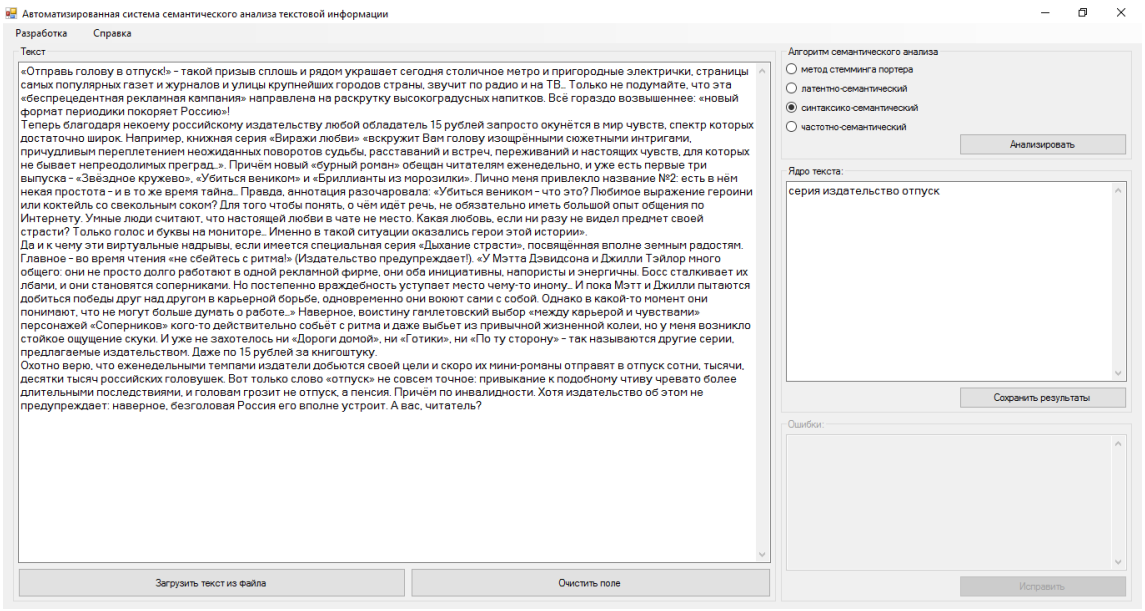

Рис. 1. Основная форма программы.

5. Результаты

В качестве объектов для исследования были выбраны тексты для сочинений ЕГЭ по русскому языку. Данные тексты были выбраны из-за своей простоты и небольшого размера, а также потому, что они рассчитаны на четкое определение тематики.

В таблицах 1-5 представлено соотношение методов анализа в виде результатов и времени их работы над определенным текстом.

Таблица 1. «Отправь голову в отпуск!» (П. Измайлов)

Метод анализа	Приблизительное время работы (сек.)	Ядро текста
Частотно-семантический	5	отпуск
Стемминга Портера	1	чувств чувствами друг другом другие серия серии голову головам отпуск
Латентно-семантический	210	голову чувств любви серия страсти время ритма рублей отпуск
Синтаксико-семантический	720	серия издательство отпуск

Тему первого текста можно определить как «влияние массовой литературы на интеллектуальное развитие человека». Ни один из методов не выдал похожих тем, но наиболее близки к ней латентно-семантический и стеммминга Портера.

Таблица 2. «Вещи и книги, книги и вещи...» (Л. Лиходеев)

Метод анализа	Приблизительное время работы (сек.)	Ядро текста
Частотно-семантический	5	паровоз свет книгах вещей время собеседника
Стемминга Портера	2	книгах книгой книгу
Латентно-семантический	240	мыслей мысль мысли собеседник собеседника свет свете книгах книгой вещи вещах вещей вещами времени время другой друг друга
Синтаксико-семантический	840	мысли вещи время собеседника

Тему второго текста можно определить как «взаимоотношения книги и времени». Латентно-семантический анализ выдал результат, наиболее удовлетворяющий теме.

Таблица 3. «Земля — космическое тело, а мы — космонавты...» (В. Солоухин)

Метод анализа	Приблизительное время работы (сек.)	Ядро текста
Частотно-семантический	5	систему жизнеобеспечения корабле космонавты земли реки возможность общения стороны человека болезнь
Стемминга Портера	1	жизнеобеспечения космическое космическом космонавты космонавтов корабле корабля человека человек
Латентно-семантический	120	Солнца Солнцем жизнеобеспечения космическое космическом космонавты космонавтов маленьком маленького корабле корабля земли Земля реки природы природой внешним внешний миром мир человека человек духовного духовному болезнь
Синтаксико-семантический	540	космонавты корабле человек

Тему третьего текста можно определить как «взаимоотношения человека и природы». Латентно-семантический анализ выдал результат, наиболее удовлетворяющий теме.

Таблица 4. «Книги...» (А. Етоев)

Метод анализа	Приблизительное время работы (сек.)	Ядро текста
Частотно-семантический	5	жизнь людей человека люди детстве книги друг
Стемминга Портера	1	душа души душе
Латентно-семантический	120	меряют меряет мере встречи человек человека люди людей одинаково одинакова книги книга способна способны
Синтаксико-семантический	540	человек пространство жизнь население люди книга

Тему четвертого текста можно определить как «роль книги в жизни человека». Синтаксико-семантический анализ выдал результат, наиболее удовлетворяющий теме.

Таблица 5. «О душе» (М. Пришвин)

Метод анализа	Приблизительное время работы (сек.)	Ядро текста
Частотно-семантический	5	душа плащ
Стемминга Портера	1	душа душе души
Латентно-семантический	90	душа душе души плащ плащом
Синтаксико-семантический	600	душа год плащ

Тему пятого текста можно определить как «душа человека». Все алгоритмы выдали удовлетворительный результат, наиболее точный из них у метода стемминга Портера.

6. Заключение

В статье были рассмотрены методы классификации текстов, такие как стемминг Портера, синтаксико-семантический, частотно-семантический и латентно-семантический анализы. Были приведены результаты анализа текстов небольшой сложности. Из них можно сделать вывод о том, что применение методов определения темы текста зависит от сложности самого текста: чем сложнее текст, тем точнее должен быть анализ. То же относится и к тривиальным текстам: использование на них сложных методов приводит к лишней трате времени и ресурсов, а результат получается избыточным по сравнению с простыми алгоритмами. Таким образом, проведенный анализ показал, что самым эффективным оказался латентно-семантический анализ, наиболее быстрым – метод стемминга Портера. Также стоит отметить целесообразность применения комбинированных методов анализа текста: например, совмещение метода стемминга Портера и частотно-семантического анализа.

Литература

[1] Понимание и синтез текста компьютером [Электронный ресурс]. – Режим доступа: http://compuling.narod.ru/index2.html (25.10.16)

[2] Russian stemming algorithm [Электронный ресурс]. – Режим доступа: http://snowball.tartarus.org/algorithms/russian/stemmer.html (11.12.16)

[3] Заболеева-Зотова, А.В. Латентный семантический анализ: новые решения в Internet / А.В. Заболеева-Зотова. - Москва: Информационные технологии, 2001. - 22 с.

[4] Рабчевский, Е.А. Автоматическое построение онтологий на основе лексико-синтаксических шаблонов для информационного поиска / Е.А. Рабчевский - Петрозаводск, 2009. - 107 с.