ПРИМЕНЕНИЕ МЕТОДALSA-АЛГОРИТМА ДЛЯ ОСУЩЕСТВЛЕНИЯ ИНТЕЛЛЕКТУАЛЬНОГО ПОИСКА КОМПЛЕКСА УПРАЖНЕНИЙ РАЗВИТИЯ КОМПЕТЕНЦИЙ

Хаперская А. В.
Научный руководитель Берестнева О.Г.
Томский политехнический университет
E-mail: khape@mail.ru

Семантический анализ не простая математическая задача, несмотря на то, что он пользуется спросом практически во всех областях жизни общества. Задача заключается в обучении компьютера распознавать и правильно интерпретировать образы, которые пользователь пытается передать. Ту Гонсалес считает, что человек является очень сложной информационной системой - в определенной степени это определяется чрезвычайно развитыми его способностями распознавать образы, которые представляют собой описание объекта. Естественный язык был сформирован во многом хаотично, в отличие от алгоритмических языков. По этой причине, существует целый ряд трудностей в понимании и распознавания текста [1]. Таким образом, чтобы разработать программу, которая позволит интеллектуальный поиск бизнес-кейсов, деловых игр, необходимо разработать математическую модель принятия решений, направленную на облегчение поиска заданий для развития индивидуальных компетенций.

Среди огромного числа алгоритмов, которые используются для поиска и анализа информации, особое место занимают те из них, целью которых является обнаружение скрытых закономерностей или неочевидных зависимостей

Используя их, мы можем сказать, например, что два текста похожи, даже если эта похожесть выражена косвенно:

Об одном из таких методов, который применяется для рекомендательных систем (коллаборативная фильтрация), информационного семантического поиска, разделения текстов по тематикам без обучения и многих других и пойдет речь далее. Метод этот называется латентно-семантическим анализом (LSA - Latentsemanticanalysis)[2].

Можно изобразить это графически на простом примере двух небольших текстов. Один текст про письменность, другой про неопределенность Гейзенберга. Стоп-слова удалены, а остальные приведены к основной форме (без окончаний). Каждая точка на графике - слово. На осях отложено, сколько раз слово встретилось в каждом документе. Т.е. если слово встретилось в тексте про неопределенность 3 раза, а в тексте про письменность 2 раза, то на рисунке 1 это слово изобразим точкой с координатами (3,2)

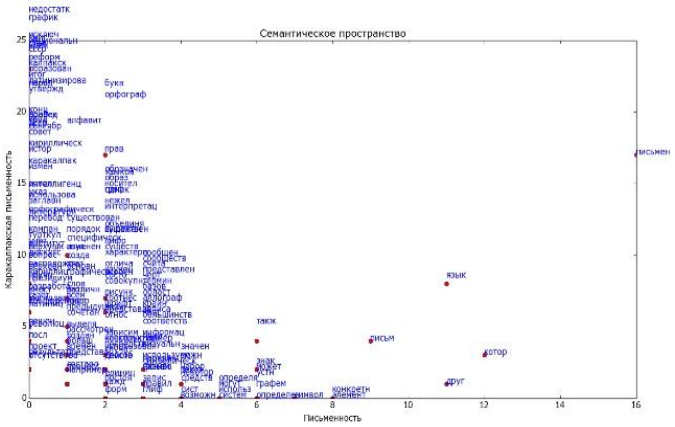

Рис. 1. Пример текста с семантическим пространством

Для сравнения документов можно подсчитать сумму векторов-слов, которые в них входят и опять же оценить расстояние между ними. В рассмотренном примере слова распределились хорошо, так как тематики существенно разные. А если тематики схожи, то может получиться такая картина (Рис.2):

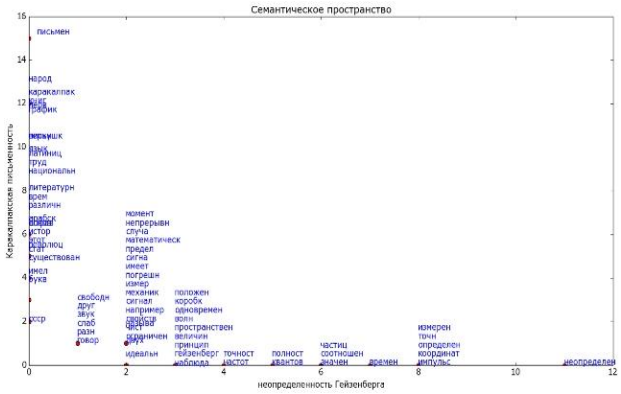

Рис. 2. Пример текста с семантическим пространством со схожими тематиками

По сравнению с предыдущей картинкой видно, что документы существенно похожи, и, кроме того, есть слова, которые характеризуют общую тематику для обоих текстов (например "язык" и "письмен"). Такие слова можно назвать ключевыми для данной темы. Т.е. напрашивается вывод, что имея такое представление текстов, теоретически можно сгруппировать документы по близости их содержимого, и таким образом построить тематическое разбиение коллекции

текстов. В частности может оказаться, что каждый документ - это отдельная тема. Также можно искать документы по запросу, при этом могут находиться документы, которые не содержат слов из запроса, но близки ему по теме.

Но в жизни оказывается, что документов и слов очень много (гораздо больше чем тем) и возникают следующие проблемы:

- размерности (вычисление близости между векторами становится медленной процедурой);
- зашумленности (например, посторонние небольшие вставки текста не должны влиять на тематику);
- разряженности (большинство ячеек в таблице будут нулевыми).

В таких условиях довольно логично выглядит идея, вместо таблицы "слово-документ" использовать "слово-тема" и "тема-документ". Решение именно такой задачи предлагает LSA. Но, к сожалению, интерпретация полученных результатов может оказаться затруднительной.

Пусть имеются три документа, каждый - на свою тематику (первый про компетенции, второй про спорт и третий про компьютеры). Используя LSA, изобразим двумерное представление семантического пространства, и как в нем будут представлены слова (красным цветом), запросы (зеленым) и документы (синим). Напомню, что все слова в документах и запросах прошли процедуру лемматизации или стемминга.

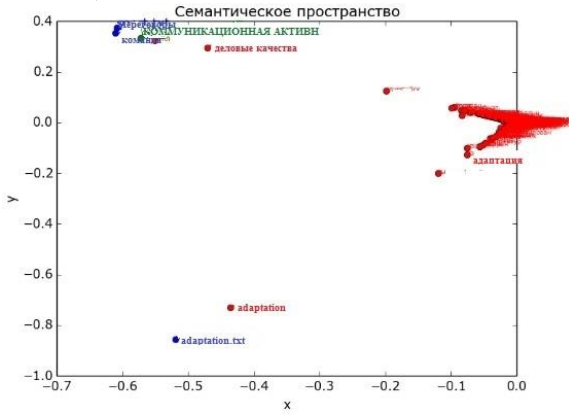

Рис. 3. Двумерное представление семантического пространства

Видно, что тема "адаптация" хорошо отделилась от двух других. А вот "переговоры" и "тимбилдинг" довольно близки друг другу. Для каждой темы проявились свои ключевые слова. Зеленым на рисунке изображен запрос "коммуникационная активность". Его релевантность к документам имеет следующий вид:

1. communication.txt' - 0.99990845
2. 'teambuilding.txt' - 0.99987185
3. 'adaptation.txt' - 0.031289458

Из-за близости тем "общение" и "тимбилдинг" довольно сложно точно определить, к какой теме он принадлежит. Но точно не к "адаптации". Если в системе, обученной на этих документах, попытаться определить релевантность к образовавшимся темам слова "рынок", то в ответ мы получим 0 (т.к. это слово в документах не встречалось ни разу).

Итак подведем итог:

1. LSA позволяет снизить размерность данных - не нужно хранить всю матрицу слово-документ, достаточно только сравнительно небольшого набора числовых значений для описания каждого слова и документа.
2. Получаем семантическое представление слов и документов - это позволяет находить неочевидные связи между словами и документами.

На рисунке 4 приведена матрица документов.

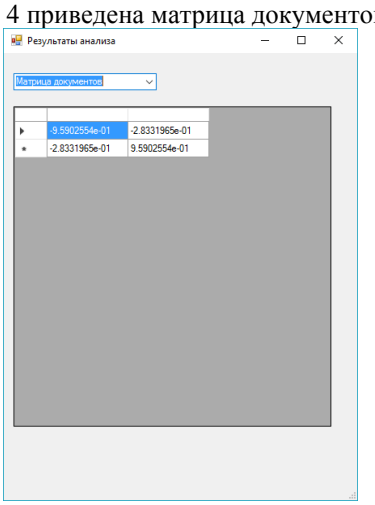

Рис. 4. Матрица документов

Итак, по разработанному алгоритму была создана система поиска заданий для развития компетенций, где результаты анализа выглядят в виде матрицы термов

Список использованных источников

1. Абакумова Н.Н., Малкова И.Ю. Компетентностный подход в образовании: организация и диагностика. - Томск: Томский государственный университет, 2007. - 368 с.
2. Осипов Г. С., Шелманов А. О. Метод повышения качества синтаксического анализа на основе взаимодействия синтаксических и семантических правил // Труды шестой международной конференции «Системный анализ и информационные технологии» (САИТ). — Т. 1. — 2015. — С. 229–240.