ИСПОЛЬЗОВАНИЕ АССОЦИАТИВНОГО ВЕКТОРНОГО ПРОСТРАНСТВА В СЕМАНТИЧЕСКОМ ПОИСКЕ

USING THE ASSOCIATIVE VECTOR SPACE IN A SEMANTIC SEARCH

V. Sachkov

Summary. The article investigates the possibility of using the properties of paradigmatic and syntagmatic associations used in associative vector spaces, for applying semantic search in solving problems, in modern information retrieval systems.

Keywords: Associations, paradigmatic, syntagmatic, search systems, associative vector space, EMW, WMD, semantics, semantic search.

Сачков Валерий Евгеньевич
Аспирант, ФГБОУ ВО «Московский технологический университет»
megawatto@mail.ru

Аннотация. В статье исследована возможность применения свойств парадигматических и синтагматических ассоциаций, использующиеся в ассоциативных векторных пространствах, для применения в решении задач семантическом поиска, в современных информационно поисковых системах.

Ключевые слова: Ассоциации, парадигматические, синтагматические, поисковые системы, ассоциативное векторное пространство, EMW, WMD, семантика, семантический поиск.

Введение

По сравнению с традиционными полнотекстовыми поисковыми системами, которые ориентированы на частоту появления слов, семантические поисковые системы более склонны пытаться понять значения, скрытые в полученных документах и пользовательских запросах, посредством добавления семантических тегов в тексты, чтобы структурировать и концептуализировать объекты в документах. Люди могут понять вопрос, основываясь на контексте и дать соответствующий ему релевантный ответ.

Ассоциации позволяют абстрагироваться от прямого значения слова, что позволяет заменить его на набор других слов. Этот эффект имеет и обратное действие, по набору слов (ассоциаций) человек способен восстановить искомое слово. Данное свойство позволяет человеку формировать поисковый запрос, не зная ключевых слов или терминов той предметной области, которой он не разбирается, но при этом получать нужный ему результат. В ассоциативном поиске совершенно не важен порядок слов и их количество, что позволяет человеку общаться с поисковым комплексом на естественном языке, не формируя поисковые фразы специальным образом, так как комплекс, сам обработает запрос и очистит от всего лишнего.

Данный подход к организации поиска кардинально отличается от организации существующих современных информационно поисковых систем и заслуживает внимания.

Поиск с помощью ассоциаций основывается на дистрибутивной семантике, которая позволяет интерпретировать семантическое поле как семантическое векторное пространство, которое позволяет вычислить семантическую близость между лингвистическими единицами.

Ассоциативное векторное пространство

Ассоциативное векторное пространство(АВП) — это многомерное векторное пространство, где каждый вектор содержит набор лингвистических единиц (слов или словосочетаний), соответствующих ассоциативному контексту документа на естественном языке.

Для формирования векторов можно использовать несколько подходов:

1. Оставить только уникальные ассоциации для текста
2. Оставить все ассоциации
3. Создать вектор частот ассоциаций
4. Создать вектор с применением метрики для оценки значимости слов (например, TF-IDF)

Поиск документов в АСВ происходит не по ключевым словам, а по смыслу, происходящий за счет сопоставления ассоциаций между текстом и поисковым запросом, что позволяет производить семантический поиск близких по смыслу текстов и документов на естественном языке.

Ассоциации позволяют абстрагироваться от прямого значения слова, что позволяет заменить его на набор других слов. Этот эффект имеет и обратное действие, по набору слов (ассоциаций) есть возможность восстановить искомое слово. В использование АСВ, применя-

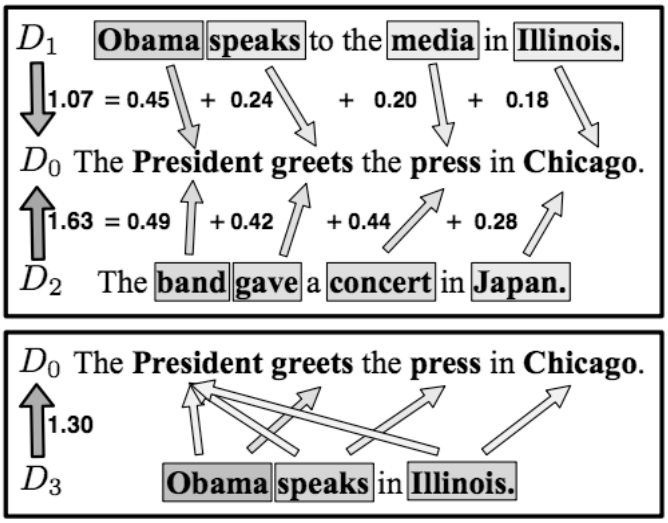

Рис. 1. Визуализация процесса вычисления расстояния WMD [5]

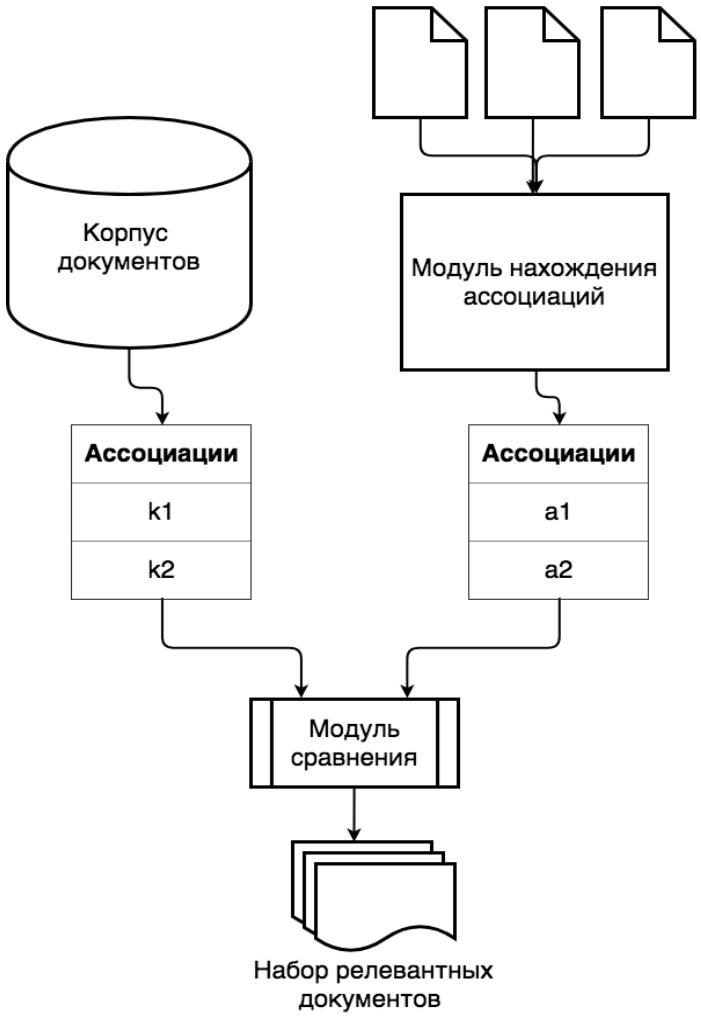

Рис. 2. Функциональная схема ассоциативно-семантического поиска в АВП

Таблица 1. Расшифровка результатов ассоциативно-семантического поиска в АВП

doc_num	EMD	text
485	113.226833	москв 28 — выяснен причин авиакатастроф 154 близ соч продолжа параллельн поисков операц задействова разнородн группировк корабл суд глубоководн аппарат самолет вертолет миноборон мчс фсб росс федеральн региональн структур поднят внов обнаружен останк погибш направл идентификац москв провед перви…
344	127.266138	петербург 4 — аэропорт храбров калининград ноч сред Airbus A320 компан аэрофлот выкат предел впп закр крайн мер 21 00 мск эвакуац поврежден воздушн судн продолжа сообщ пресс секретар аэропорт натал грицун первоначальн сообща инцидент обошел пострада ноч помощник глав север западн следствен управ…
332	138.888802	ноч сред калининградск аэропорт храбров 23 55 мск самолет аэрофлот Airbus A 320 рейс москв — калининград выкат посадк метр предел взлетн посадочн полос борт наход 167 пассажир пятер член экипаж происшеств самолет подлом передн стойк шасс пришл эвакуирова помощ надувн аварийн трап 9 55 10 55 мск …
347	139.834790	сообща 23 55 местн самолет аэрофлот Airbus A320 рейс москв калининград выкат посадк предел взлетн посадочн полос подлом передн стойк шасс пассажир эвакуирова посредств надувн аварийн трап дан север западн следствен управлен транспорт ск рф шестер пассажир обрат медицинск помощ факт чп возбужд уг…
510	141.854178	москв 27 — поисков операц район крушен самолет 154 близ соч начат воскресен продолжа круглосуточн режим вторник задействова 45 корабл суд 15 глубоководн аппарат 192 водолаз 12 самолет вертолет предполага район катастроф обследова 100 надводн поверхн бортов самописец поднят морск дна доставл расш…

ются все свойства парадигматических и синтагматических ассоциаций.

Ассоциации называются синтагматическими, если данная ассоциация по своей грамматической структуре отлична от слова — стимула, например, испытуемый пытается составить слово сочетание «*дом — большой, светлый*» Парадигматические ассоциации — это слова — реакции с той же грамматической структурой, что и слова — стимулы, например, испытуемый подбирает синонима или антонимы «*дом — шалаш, жилище*», или части «*дом — дверь, крыша*».

Используя АСВ можно сформировать поисковый запрос на естественном языке без единого колючего слова, не обращая внимания на порядок слов и их количество, без какой-либо дополнительной обработки текста и запроса. АСВ позволяет существенно улучшить и упростить создание интерактивных диалогов систем (например, чат бот), которые лучше понимают пользователя.

Для ассоциативно-семантического поиска в АСВ, необходима математическая модель, которая позволит оценивать семантическую близость двух документов по их ассоциациям. В качестве метрики расстояния будет использоваться «Earth mover's distance» (EMD). EMD это метод оценки несходства между двумя многомерными распределениями в каком-то пространстве признаков, где дана дистанционная мера между одиночными признаками [1].

Метрика EMD вычисляет минимальную стоимость изменений или работы необходимой, для преобразования одного документа в другой. Вычисление EMD базируется на решении транспортной задачи [2] линейного программирования, для решения которой существуют эффективные алгоритмы [3].

Так как вычисления будут производиться над словами, то наилучшим алгоритмом для вычисления дистанции, с использованием EMD, будет модифицированный метод вычисления Word Mover's Distance (WMD). WMD это метод, который позволяет оценивать «расстояние» между двумя документами, даже если у них нет общих слов [4]. Визуально процесс вычисления расстояния представлен на Рис. 1

В WMD текстовые документы представляются, как облако точек встроенных слов. Расстояние между двумя текстовых документами A и B является минимальное совокупное расстояние, необходимое для слов из документа A чтобы переместиться в облако точек документа B. В качестве примера рассмотрим вычисления расстояния между двумя предложениями на английском языке [5]:

- Obama speaks to the media in Illinois
- The President greets the press in Chicago

Метод ассоциативно-семантического поиска в ассоциативном векторном пространстве

Проверка семантического поиска в ассоциативном векторном пространстве, будем проведена поиском, отбором и ранжированием тематических схожих документов на основе собранных корпусов документов. Для этой

цели, специально, был подготовлен большой корпус документов с новостного портала «РИА НОВОСТИ» (https://ria.ru), с помощью веб-скраппера (web-scraping), был собран корпус текстов, состоящий из 250 тысяч документов, по основным разделам сайта, за 3 года новостных публикаций портала.

Пользователем будет задана тема в виде небольшой коллекции документов, описывающая интересующею его тематику.

На основе анализа представленной коллекции автоматизировано формируется поисковый запрос, а результаты поиска фильтруется и ранжируется в соответствии с метрикой расстояния в АВП.

Тема тестового поиска «Авиационные происшествия», была собрана коллекция из 10 новостных документов с разными инцидентами связанные с авиацией. Номера документов в корпусе: 119, 121, 332, 342, 347, 355, 504, 559, 661, 948. Функциональная схема ассоциативно-семантического поиска в АВП представлена на Рис. 2.

Принцип работы алгоритма заключается в следующих шагах:

1. На вход подается набор подготовленных документов, по которым будет производиться поиск.
2. Модуль нахождения ассоциаций создает список ассоциаций
3. Набор найденных ассоциаций подается на вход модуля сравнения, который создает АВП и ищет совпадения по разным наборам алгоритмов с ассоциациями, содержащимися в соответствующем по тематике в корпусе документов.

Расшифровка результатов работы алгоритма представлена в таблице 1.

Как показало тестирование большая часть найденных документов соответствуют тематики поиска. В найденных документах могут полностью отсутствовать ключевые слова из поискового запроса, но содержание документа соответствует тематики поиска.

Недостатки, которые были получены при экспериментах показали следующие, если поисковый запрос менее пяти слов, то необходимо вводить новые методы расчета расстояния WMD. Также если поисковый запрос имеет значительный размер, состоящий из нескольких десятков слов и более, то будет больше семантическое расстояние в ассоциативном векторном пространстве, и тем менее будут схожи по смыслу найденные документы.

Выводы

Рассмотрена проблема полноты результатов семантического поиска. Сделана попытка решения данной задачи с помощью ассоциативного векторного пространства, данный метод позволил обеспечить полноту более высокую, относительно обычных поисковых систем. Использование парадигматических и синтагматических ассоциаций может применяться в разработке семантических поисковых систем, способных выявить контекстную информацию на заданном уровне.

ЛИТЕРАТУРА

1. Y. Rubner, C. Tomasi, and L. J. Guibas. A metric for distributions with applications to image databases. // IEEE International Conference on Computer Vision, pages 59–66, January 1998.
2. F. L. Hitchcock. The distribution of a product from several sources to numerous localities. J. Math. Phys., 20:224–230, 1941
3. Труды Международной научно-технической конференции. Т. 2 / под ред. С. А. Прохорова.— Самара: Издательство Самарского научного центра РАН. 2015.— с. 37–41
4. Finding similar documents with Word2Vec and WMD [электронный ресурс] URL: https://markroxor.github.io/gensim/static/notebooks/WMD_tutorial.html (дата обращения 10.12.2017)
5. Matt Kusner, Yu Sun, Nicholas Kolkin, Kilian Weinberger From Word Embeddings To Document Distances // Proceedings of the 32nd International Conference on Machine Learning, PMLR37:957–966, 2015.

Журнал «Современная наука: актуальные проблемы теории и практики»