УДК 339.138

Студ. Капелько И.В.
Науч. рук. доцент Д.В.Шиман
(кафедра информационных систем и технологий, БГТУ)

АНАЛИТИЧЕСКИЙ ОБЗОР МЕТОДОВ TEXTMINING

Большинство текстовых данных, встречающихся в повседневной жизни, неструктурированный, или простой текст. Это самый известный тип текста, встречающийся в повседневных источниках, таких как книги, газеты, электронные письма и веб-страницы. Под неструктурированным текстом подразумевается текст, который не был приведен к формату таблицы или реляционной базы данных. Часто в качестве данных используются поля электронных таблиц или поля таблиц баз данных. Этот тип текста называется полуструктурированным. Почти все данные в мире хранятся в неструктурированном или полуструктурированном формате, по последним оценкам от 75 до 80 процентов всех данных находятся в текстовом формате. В это легко поверить, учитывая сколько информация только в интернете хранится в виде текста [1].

Практическое применение технологии TextMining можно разделить на семь областей, основанных на уникальных характеристиках каждой области. Несмотря на то, что эти области очень взаимосвязаны, для типичного проекта, использующего TextMining, потребуются методы из нескольких областей.

LSA. Латентно-семантический анализ (ЛСА, LSA) – метод обработки информации на естественном языке, устанавливающий взаимосвязь между текстами (документами) и терминами в них встречающимися.

LSA используется для извлечения контекстно-зависимых значений лексических единиц на основе факторного анализа и статистической обработки больших корпусов текстов.

LSA запатентованв 1988г (Scott Deerwester, Susan Dumais, George Furnas, Richard Harshman, Thomas Landauer, Karen Lochbaum и Lynn Streeter).

LSA впервые применен для автоматического индексирования текстов, выявления семантической структуры текста и получения псевдодокументов.

LSA используется для:

1. для поиска информации (индексация документов) – области информационного поиска данный подход называют латентно-семантическим индексированием (ЛСИ);

2. классификации документов;
3. представления баз знаний.
4. построения когнитивных моделей (моделей понимания);

LSA представим (моделируем) трехслойной нейросетью:

1. множество слов (термов);
2. множество документов, соответствующих определенным ситуациям;
3. (латентный слой) множество узлов с различными весовыми коэффициентами, связывающих первый и второй слои.

В качестве исходной информации LSA использует матрицу термы-на-документы, описывающую набор данных, используемый для обучения системы. Элементы этой матрицы содержат, как правило, веса, учитывающие частоты использования каждого терма в каждом документе и участие терма во всех документах (TF-IDF).

Наиболее распространенный вариант LSA 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 исходным матрицам.

Достоинством метода можно считать его замечательную способность выявлять зависимости между словами, когда обычные статистические методы бессильны. ЛСА также может быть применен как с обучением (с предварительной тематической классификацией документов), так и без обучения (произвольное разбиение произвольного текста), что зависит от решаемой задачи [2].

Метод, основанный на терминах. Термин в документе – это слово, имеющее смысловое значение. В методе, основанном на терминах, документ анализируется на основе термина и имеет преимущества эффективной вычислительной производительности, а также уже состоявшиеся теории для установки весовых коэффициентов терминам. Эти методы возникли в течение последних нескольких десятилетий из области информационного поиска в сообществе машинного обучения. Методы, основанные на терминах, страдают от проблем многозначности и синонимии. Многозначность означает, что слово имеет несколько значений, а синонимия представляет собой несколько слов, имеющих одинаковое значение. Смысловое значение многих изученных терминов является неопределенным для ответа на то, что хочет пользователь.

Метод, основанный на фразах. Фраза несет в себе больше семантики как информация и она менее неоднозначна. В методе, осно-

ванном на фразах, документ анализируется на основе фразы, так как фразы наименее неоднозначны и более отчетливы, чем отдельные термины. Выделяют следующие причины высокой эффективности данного метода:

1. Фразы имеют подчиненные статистические свойства терминов;
2. Они имеют низкую частоту появления в тексте;
3. Среди них присутствует большое количество избыточных и «шумных» фраз.

Метод, основанный на концепциях или понятиях. В методе, основанном на концепциях или понятиях, термины анализируются на основе предложений и уровня документа. Методы интеллектуального анализа текста в основном базируются на статистическом анализе слова или фразы. Статистический анализ частоты появления термина фиксирует важность слова вне документа. Два термина могут иметь одинаковую частоту появления в одном и том же документе, но смысл в том, что один термин целесообразнее способствует пониманию документа, чем другой термин. Так как вводится новый интеллектуальный анализ, основанный на понятиях, то следует уделять большее внимание терминам, которые фиксируют семантику, то есть смысл, текста. Эта модель включает в себя три компонента. Первый компонент анализирует смысловую структуру предложений. Второй компонент создаёт концептуальный онтологический граф (conceptual ontological graph (COG)) для описания семантических структур. Последний компонент извлекает верхние понятия (концепции), основанные на первых двух компонентах, для того, чтобы построить векторы признаков или свойств, используя стандартную модель векторного пространства. Понятийно-ориентированная модель может эффективно делать различия между «неважными» терминами и терминами, имеющими значение, которые описывают смысл предложения. Понятийно-ориентированная модель, как правило, опирается на технологии обработки естественного языка. Выбор свойства применяется к понятиям запросов для того, чтобы оптимизировать представление и устранить помехи или «шум» и неоднозначность [2].

Метод шаблонной систематики. В методе шаблонной систематики документы анализируются на основе шаблона или образца. Образцы могут быть структурированы в систематику при помощи отношения наследования. Интеллектуальный анализ шаблонов широко изучается в сообществах интеллектуального анализа данных в течение многих лет. Шаблоны могут быть открыты с помощью таких технологий интеллектуального анализа данных, как правило ассоциации, «наи-

более частый элемент множества», последовательный интеллектуальный анализ шаблонов и закрытый анализ шаблонов. Использование обнаруженных знаний (шаблонов) в области интеллектуального анализа текста – это сложно и неэффективно, потому что некоторые полезные длинные шаблоны с высокой специфичностью не имеют поддержки (то есть это так называемая низкочастотная проблема). Не все часто встречающиеся короткие шаблоны полезны. Существует проблема неправильного истолкования шаблонов, которая приводит к неэффективной производительности анализа текста.

Метод, основанный на шаблонах, использует два процесса: приведение в действие шаблона и его развитие. Этот метод совершенствует обнаруженные шаблоны в текстовых документах. Экспериментальные результаты показывают, что модель, основанная на шаблонах, работает лучше, чем не только другие модели, основанные на интеллектуальном анализе данных, и понятийно-ориентированная модель, но и лучше модели на основе терминов

Заключение. В последнее время анализ текста привлекает всё больше внимания в различных областях, таких как безопасность, коммерция, наука. Непрерывное накопление текстовых данных привело к необходимости разработки методов интеллектуального анализа текстов для обеспечения эффективной работы с большими корпусами текстов.

В ближайшем будущем технология интеллектуального анализа текста станет доминирующей при анализе информации от клиентов в компаниях любого уровня, будь то телефонные центры поддержки, интернет-агентства или аналитические агентства.

ЛИТЕРАТУРА

1. Gary Miner. Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications. Academic Press, 2012 pp. 71–138.

2. Fayyad U., Piatetsky-Shapiro G., Smyth P. From Data Mining to Knowledge Discovery: an Overview // Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press, 1996. - pp. 1-34.