УДК 004.89

АНАЛИЗ ВЛИЯНИЯ МОДЕЛЕЙ ПРЕДСТАВЛЕНИЯ ТЕКСТОВ НА КАЧЕСТВО КЛАССИФИКАЦИИ ОТЗЫВОВ ПО ТОНАЛЬНОСТИ

Вычегжанин С.В., Котельников Е.В.

ФГБОУ ВО «Вятский государственный гуманитарный университет», Киров, e-mail: vychegzhaninsv@gmail.com, kotelnikov.ev@gmail.com

В статье исследуется влияние векторной и графовой моделей представления текстов на качество классификации отзывов по тональности с использованием ДСМ-метода автоматического порождения гипотез. Представление текста в виде графа в отличие от векторного представления эффективно моделирует связи между терминами и позволяет отразить структуру документа. В работе эта особенность используется для получения более качественных гипотез, генерируемых в процедуре индукции ДСМ-метода. Приводится сравнительный анализ гипотез в векторной и графовой моделях представления текстов, отмечаются преимущества графовых гипотез перед векторными гипотезами, оценивается эффективность использования исследуемых моделей в задаче классификации текстов по тональности. Эксперименты по определению качества классификации проводятся с применением коллекции отзывов семинара РОМИП-2011.

Ключевые слова: ДСМ-метод, анализ тональности текстов, графовая модель, векторная модель

ANALYSIS OF THE EFFECT OF TEXT REPRESENTATION MODELS ON THE QUALITY OF REVIEW SENTIMENT CLASSIFICATION

Vychegzhanin S.V., Kotelnikov E.V.

Vyatka State Humanities University, Kirov, e-mail: vychegzhaninsv@gmail.com, kotelnikov.ev@gmail.com

The influence of vector space and graph-based text representation models on the quality of review sentiment classification with using of the JSM-method of automatic generation of hypotheses is researched in the article. The graph-based text representation unlike vector space one effectively models terms relations and allows to express structure of the document. In article this feature is used for receiving more qualitative hypotheses generated in the procedure of induction of the JSM-method. The comparative analysis of hypotheses in vector space and graph-based text representation models is provided, advantages of graph hypotheses as compared to vector ones are noted, efficiency of using information about structure of texts in task of sentiment classification is estimated. The experiments on estimation of the quality of classification are carried out with using of the text collection of reviews from seminar ROMIP-2011.

Keywords: JSM-method, text sentiment analysis, graph-based model, vector space model

В современном мире мнения, содержащиеся на таких веб-ресурсах, как социальные сети, блоги и форумы, оказывают сильное влияние на людей при выборе товаров и услуг, при формировании отношения к определенным явлениям, организациям или персонам. В связи с этим активно развивается одно из направлений компьютерной лингвистики – автоматический анализ мнений в текстах, который применяется при проведении маркетинговых, экономических, социальных и других видов исследований [7]. Одной из основных задач анализа мнений является классификация текстов по тональности, т.е. определение выраженной в тексте эмоциональной оценки по отношению к некоторому объекту.

В настоящей работе рассматривается бинарная шкала тональности, включающая два значения – позитивное и негативное. Классификация осуществляется на основе ДСМ-метода автоматического порождения гипотез, предложенного В.К. Финном в конце 1970-х гг. [1]. Целью работы является исследование влияния векторной и графовой моделей представления текстов на качество классификации отзывов по тональности.

Модели представления текста

Модель представления текстового документа является важной составляющей в процессе его компьютерной обработки. Выбор модели определяет эффективность выделения смыслового содержания и структуры документа. В настоящей работе исследуются две модели – векторная и графовая. Выбор моделей обусловлен существенным их различием в способности отражения структуры документа. Сравнительный анализ данных моделей позволит оценить эффективность использования информации о связях между терминами в задаче классификации по тональности.

Векторная модель – алгебраическая модель для представления текстовых документов в виде векторов в пространстве признаков. Предложена Дж. Солтоном в 1975 году [9].

В векторной модели документ рассматривается как множество терминов, в качестве которых могут выступать отдельные слова или словосочетания. Каждая компонента вектора признаков соответствует отдельному термину. Числовое значение компоненты совпадает с весом термина, который характеризует важность данного термина для представления содержащего его документа. Если термин не встречается в документе, то его вес в этом документе равен нулю.

Векторная модель имеет ряд недостатков:

1) теряются взаимосвязи между словами;

2) если документы имеют одинаковый смысл, но состоят из разных слов, то сложно выявить сходство документов;

3) сложность представления больших документов.

Модель представления текста в виде графа позволяет устранить указанные недостатки. Такое представление эффективно моделирует связи между терминами и отражает информацию о структуре документа. При этом вершинам графа ставятся в соответствие отдельные термины, а ребрам – связи между ними. В настоящей работе каждый текст представлялся в виде графа совместно встречающихся слов. Множество вершин графа формировалось из уникальных слов, входящих в текст. Для расстановки ребер проводилось сканирование текста окном заданного размера. Ребро между двумя вершинами в графе устанавливалось в том случае, если соответствующие этим вершинам слова в тексте одновременно находились в пределах сканирующего окна. Данный подход является достаточно простым в реализации и основан на наблюдении из [8], что между двумя находящимися рядом словами часто существует семантическая связь.

Сравнительный анализ рассмотренных моделей осуществлялся в рамках задачи классификации текстов по тональности на основе ДСМ-метода.

ДСМ-метод

ДСМ-метод автоматического порождения гипотез является представителем логического подхода в интеллектуальном анализе данных. Преимуществом ДСМ-метода по сравнению со статистическими методами является прозрачность процесса логического вывода и хорошая интерпретируемость генерируемых гипотез [3]. ДСМ-метод реализует синтез трех познавательных процедур – эмпирической индукции, структурной аналогии и абдукции [1]. В настоящей работе рассматриваются две процедуры – индукции и аналогии.

Идея метода заключается в следующем [3]. Имеются две коллекции текстов: обучающая и тестовая. В процедуре индукции осуществляется поиск максимальных общих фрагментов текстов обучающей коллекции. Такие общие фрагменты называются гипотезами. Например, для текстов «*Один из лучших фильмов, которые видел в последнее время*» и «*Давно не видел настолько хорошего фильма*» гипотезой будет множество, состоящее из трех слов: <*фильм, видеть, хороший*>. Для каждого класса тональности формируется отдельное множество гипотез. В процедуре аналогии осуществляется поиск этих гипотез в текстах классифицируемой коллекции. В результате для одного и того же текста могут быть найдены как позитивные, так и негативные гипотезы. В этом случае возникает ситуация конфликта. Для ее разрешения гипотезы передаются в функцию разрешения конфликтов [4], которая присваивает тексту определенный класс тональности на основе некоторого критерия, например соотношения мощностей множеств гипотез каждого класса, найденных в классифицируемом тексте.

В работе осуществляется формирование гипотез с учетом векторной и графовой моделей представления текстов. При использовании графовой модели предполагается получение гипотез, отличных от гипотез в векторной модели, за счет наличия информации о связях между словами. Преимуществом графовых гипотез является более точная передача семантики текстов вследствие наличия близко расположенных терминов, образующих осмысленные словосочетания.

Методика проведения эксперимента

В ходе исследования использовались следующие материалы, инструменты и методы.

1. *Коллекции отзывов.*

Обучающая коллекция составлена из корпуса текстов семинара РОМИП-2011 [6], который содержит отзывы интернет-пользователей на фотокамеры, книги и фильмы. Отзывы собраны с сайтов imhonet.ru и market.yandex.ru. Каждый отзыв имеет оценку по шкале тональности: в категории «камеры» – от 1 до 5, в категории «книги» и «фильмы» – от 1 до 10. Для товаров категории «камеры» пятибалльная шкала приведена к двухбалльной по схеме: {1, 2} → «–», {4, 5} → «+»; для товаров категории «книги» и «фильмы» десятибалльная шкала приведена к двухбалльной по схеме: {1…4} → «–», {7…10} → «+». С целью учета близкого контекста отзывы поделены на предложения, и дальнейший анализ осуществлялся на уровне предложения. Характеристики обучающей коллекции представлены в табл. 1.

Таблица 1

Характеристики текстовых коллекций

Параметр		Предметная область		
		Камеры	Книги	Фильмы
Количество отзывов	позитивных	3 387	3 000	1 500
	негативных	3 951	2 060	1 500
Количество предложений	позитивных	14 384	11 149	12 847
	негативных	13 524	11 076	12 716
Среднее количество слов в отзыве	позитивном	37	32	125
	негативном	30	50	103
Среднее количество слов в предложении	позитивном	9	9	15
	негативном	9	10	12

При классификации тестовой коллекции оценивалось отдельно каждое предложение отзыва. Для получения оценки отзыва в целом производилось суммирование оценок его предложений: отзыву присваивался класс тональности, которому принадлежало наибольшее количество предложений отзыва. В случае равенства количества предложений в каждом классе отзыву присваивался класс последнего предложения. Если ни одному из предложений не был присвоен класс тональности, то отзыву присваивалась позитивная тональность.

2. *Словари и морфологическая обработка.*

Использовались два словаря: статистический и экспертный. При составлении статистического словаря каждое слово текстов обучающей коллекции взвешивалось с помощью функции отношения шансов *OR* (Odds Ratio) [2]. Слова сортировались по убыванию весов. В словарь добавлялось по 30 % слов с наибольшим весом из каждого класса тональности. В качестве экспертного словаря использовался словарь оценочной лексики, отдельный для каждой предметной области [5]. Информация о размерах словарей представлена в табл. 2. На этапе предварительной обработки каждое слово отзыва преобразовывалось к начальной форме при помощи морфологического анализатора *Mystem* от компании Яндекс.

Таблица 2

Размер словарей, слов

Словарь	Предметная область		
	Камеры	Книги	Фильмы
Статистический	5 888	9 565	15 413
Экспертный	1 224	2 124	2 384

3. *Построение моделей текста.*

При реализации векторной модели каждый текст представлялся в виде бинарного вектора признаков, в качестве которых использовались уникальные слова. При реализации графовой модели каждый текст представлялся в виде неориентированного графа. Сканирование текста осуществлялось окном размером пять слов. Данный размер окна оказался оптимальным по качеству классификации и времени работы программы при проведении предварительных экспериментов.

Качество классификации оценивалось с помощью F1-меры [6]. Для получения объективных оценок применялась процедура пятикратного скользящего контроля, которая заключается в разбиении обучающей коллекции на пять равных частей и объявлении поочередно каждой из частей в качестве тестовой, остальных – в качестве обучающих.

4. *Функции разрешения конфликтов.*

Разрешение конфликтов гипотез осуществлялось с помощью функций из [2]. Наилучшие результаты по F1-мере получены с помощью следующих функций:

1) отношение шансов (Odds Ratio, OR):

$$OR = \frac{a \cdot d}{b \cdot c}, \qquad (1)$$

где a – количество документов позитивного класса, содержащих гипотезу h; b – количество документов позитивного класса, не содержащих гипотезу h; c – количество документов негативного класса, содержащих гипотезу h; d – количество документов негативного класса, не содержащих гипотезу h;

2) релевантная частота (Relevance Frequency, RF):

$$RF = \log\left(2 + \frac{a}{c}\right); \qquad (2)$$

3) произведение количества терминов в гипотезе t на значение функции OR ($t \cdot$OR).

В качестве терминов в исследовании рассматривались отдельные слова.

Результаты экспериментов

Эксперименты проводились при объединении статистического и экспертного словарей, что обеспечивает совместное использование статистической информации о текстах и экспертных знаний в предметной области. Результаты по количеству гипотез представлены в табл. 3, по качеству классификации – в табл. 4.

Использование модели представления текстов в виде графа по сравнению с векторной моделью привело к сокращению количества гипотез от 25 до 43%. Это объясняется уменьшением количества допустимых связей между словами за счет использования окна фиксированного размера, что приводит к уменьшению количества возможных комбинаций слов в гипотезах. Качество классификации текстов для наилучшей функции *OR* изменилось от –1,2 до 0,4%. Наличие информации в виде связей между словами привело к сокращению количества гипотез, обеспечив сохранение качества классификации на приемлемом уровне.

Анализ множеств гипотез для каждой из рассматриваемых моделей представления текстов показал, что данные множества различны и одно из них не является подмножеством другого: имеются как одинаковые гипотезы, так и разные. Многие гипотезы в векторной модели состоят из слов, не имеющих логической связи между собою. Напротив, многие гипотезы в графовой модели представляют собой словосочетания из семантически связанных слов. Среднее расстояние между двумя наиболее удаленными друг от друга словами гипотезы, полученной пересечением двух и более текстов, для векторной модели составляет 8,3 слов, для графовой – 2,1 слова. В табл. 5 приведен пример гипотез, состоящих из семантически связанных и семантически не связанных слов для конкретного отзыва.

Таблица 3

Количество гипотез

Гипотезы	Камеры		Книги		Фильмы	
	Вектор. модель	Граф. модель	Вектор. модель	Граф. модель	Вектор. модель	Граф. модель
Позитивные	37 499	20 318	15 915	11 837	35 394	19 781
Негативные	28 402	17 218	15 836	11 836	26 647	17 184

Таблица 4

Значения F1-меры, %

Функция разрешения конфликтов	Камеры		Книги		Фильмы		Среднее	
	Вектор. модель	Граф. модель	Вектор. модель	Граф. модель	Вектор. модель	Граф. модель	Вектор. модель	Граф. модель
OR	80,2	80,2	72,1	72,5	64,1	62,9	72,1	71,9
RF	78,5	78,8	70,0	70,1	58,7	58,7	69,1	69,2
t·OR	79,6	77,9	71,1	69,0	63,5	61,1	71,4	69,3

Таблица 5

Примеры гипотез из семантически связанных и семантически не связанных слов

Отзыв	Модель представления текста	
	векторная	графовая
«Мне фильм абсолютно не понравился, напомнил картину «Код да Винчи»: когда герои каждую секунду открывают потайные двери, обнаруживают загадки, и на месте их сразу же отгадывают»	*Семантически связанные:* фильм, не понравиться фильм, напоминать картина, напоминать герой, секунда картина, фильм дверь, герой *Семантически не связанные:* место, обнаруживать место, фильм, герой место, секунда загадка, фильм картина, место дверь, фильм герой, напоминать обнаруживать, фильм	*Семантически связанные:* фильм, не понравиться фильм, напоминать картина, напоминать герой, секунда *Семантически не связанные:* место, обнаруживать

Вследствие того, что мощность множества гипотез в векторной модели больше мощности множества гипотез в графовой модели, в первом случае в процессе выполнения процедуры аналогии ДСМ-метод находит большее количество гипотез в классифицируемых текстах, чем во втором случае. Однако достаточно часто метод в обоих случаях присваивает тексту одинаковый класс тональности. Среднее количество одинаково классифицированных текстов для функции *OR* изменяется от 92 до 96 % в зависимости от предметной области.

Сравнительный анализ использования векторных и графовых моделей в задаче классификации текстов по тональности на основе ДСМ-метода выявил в случае графовой модели существенное уменьшение количества гипотез и, соответственно, времени их обработки, при сохранении качества классификации на приемлемом уровне. Данная модель позволила исключить многие гипотезы, состоящие из слов, взятых из разного контекста.

Заключение

В работе проведен сравнительный анализ векторной и графовой моделей представления текстов в задаче классификации отзывов по тональности с использованием ДСМ-метода. Графовое представление текстов позволило получить более качественные по смысловому содержанию гипотезы при выполнении процедуры индукции за счет наличия информации о структуре текста. При этом меньшего количества графовых гипотез оказалось достаточным для достижения качества классификации отзывов, сопоставимого с качеством классификации в случае использования векторных гипотез.

Работа выполнена в рамках государственного задания Минобрнауки РФ, проект № 586.

Список литературы

1. Автоматическое порождение гипотез в интеллектуальных системах / под ред. В.К. Финна. – М.: Либроком, 2009. – 528 с.

2. Вычегжанин С.В., Котельников Е.В. Исследование влияния способов взвешивания терминов на качество анализа тональности текстов с использованием ДСМ-метода // Информатика: проблемы, методология, технологии: материалы XV Международной научно-методической конференции, Воронеж, 12–13 февраля 2015 г.: в 4-х томах. – Воронеж: Издательский дом ВГУ, 2015. – Т. 3. – С. 236–241.

3. Котельников Е. В. Структура интеллектуальной ДСМ-системы для анализа тональности текстов // Научно-технический вестник Поволжья. – 2013. – № 6. – С. 344–346.

4. Котельников Е.В. Функция оценки информативности гипотез для анализа тональности текстов на основе ДСМ-метода // Фундаментальные исследования. – 2014. – № 11(10). – С. 2150–2154.

5. Котельников Е. В., Клековкина М. В. Определение весов оценочных слов на основе генетического алгоритма в задаче анализа тональности текстов // Программные продукты и системы. – 2013. – № 4. – С. 296–300.

6. Chetviorkin I., Braslavskiy P., Loukachevitch N. Sentiment Analysis Track at ROMIP 2011 // Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference «Dialogue». – 2012. – № 11(18). – Vol. 2. – P. 1–14.

7. Liu B. Sentiment Analysis and Opinion Mining // Synthesis Lectures on Human Language Technologies. – 2012. – Vol. 5(1).

8. Mihalcea R., Tarau P. TextRank: Bringing order into texts // Proceedings of the Conference on Empirical Methods in Natural Language Processing. – Spain, 2004. – P. 404–411.

9. Salton G.A., Wong A., Yang C.S. Vector Space Model for Automatic Indexing // Communications of the ACM. – 1975. – Vol. 18, № 11. – P. 613–620.

References

1. Avtomaticheskoe porozhdenie gipotez v intellektualnyh sistemah [Automatic generation of hypotheses in intellectual systems]. Moscow, Librokom, 2009. 528 p.

2. Vychegzhanin S.V. *Informatika: problemy, metodologija, tehnologii* [Informatics: problems, methodology, technology]. Voronezh, Izdatelskij dom VGU, 2015, Vol. 3, pp. 236–241.

3. Kotelnilov E.B. *Nauchno-tehnicheskij vestnik Povolzhja*, 2013, no. 6, pp. 344–346.

4. Kotelnikov E.V. *Fundamentalnye issledovanija*, 2014, no. 11(10), pp. 2150–2154.

5. Kotelnilov E.B., Klekovkina *Programmnye produkty i sistemy* [Software products and systems], 2013, no. 4, pp. 296–300.

6. Chetviorkin I., Braslavskiy P., Loukachevitch N. *Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference «Dialogue»*, 2012, no. 11(18), Vol. 2, pp. 1–14.

7. Liu B. *Synthesis Lectures on Human Language Technologies*, 2012, no. 5(1).

8. Mihalcea R., Tarau P. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Spain, 2004, pp. 404–411.

9. Salton G.A., Wong A., Yang C.S. *Communications of the ACM*, 1975, no. 18(11), pp. 613–620.

Рецензенты:

Страбыкин Д.А., д.т.н., профессор, заведующий кафедрой электронных вычислительных машин, Вятский государственный университет, г. Киров;

Прозоров Д.Е., д.т.н., профессор кафедры радиоэлектронных средств, Вятский государственный университет, г. Киров.