УДК 004.89

ФУНКЦИЯ ОЦЕНКИ ИНФОРМАТИВНОСТИ ГИПОТЕЗ ДЛЯ АНАЛИЗА ТОНАЛЬНОСТИ ТЕКСТОВ НА ОСНОВЕ ДСМ-МЕТОДА

Котельников Е.В.

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

В статье рассматривается задача анализа тональности текстов, решаемая на основе ДСМ-метода автоматического порождения гипотез. Данный метод показал высокое качество анализа тональности в предыдущих работах. Однако существенной проблемой при этом является сложность обработки в процедуре аналогии огромного числа гипотез, порождаемых в процессе индуктивного вывода. В статье предлагается использовать для решения этой проблемы новую функцию оценки информативности гипотез. Отличия предлагаемой функции от существующих заключаются, во-первых, в учете расположения слов гипотезы в рассматриваемом тексте, во-вторых, во введении специального коэффициента, учитывающего присутствие в гипотезе оценочной лексики. Проведенные эксперименты с использованием текстовых коллекций отзывов о фильмах, книгах и цифровых фотокамерах семинаров РОМИП-2011 и РОМИП-2012 подтверждают эффективность разработанной функции оценки.

Ключевые слова: ДСМ-метод, процедура аналогии, анализ тональности текстов

ESTIMATE FUNCTION OF HYPOTHESIS INFORMATIVITY FOR TEXT SENTIMENT ANALYSIS BASED ON JSM-METHOD

Kotelnikov E.V.

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

The problem of text sentiment analysis solved on the basis of JSM-method for automatic hypotheses generation is considered in the article. This method showed high quality of sentiment analysis in previous works. However, a significant problem in this case in the procedure of analogy is the complexity of the processing of huge number of hypotheses generated during the inductive inference. A new function for estimating hypotheses' informativeness to solve this problem is proposed to be used in the article. The differences between the proposed function and the existing ones consist, firstly, in the considering the disposition of words of the hypothesis in the analyzed text, and secondly, in the introduction of a special factor which takes into account the presence of sentiment words in the hypothesis. The experiments in which we use the text collections of reviews of movies, books and digital cameras of seminars ROMIP-2011 and ROMIP-2012 confirm the effectiveness of the developed evaluation function.

Keywords: JSM-method, procedure of analogy, text sentiment analysis

Анализ тональности текстов представляет собой область компьютерной лингвистики, которая занимается автоматическим определением тональности в текстах на естественном языке [13]. Под тональностью при этом понимаются эмоции, мнения, оценки, выраженные в тексте по отношению к некоторым объектам.

В настоящее время можно выделить два основных подхода, используемых при анализе тональности [13]: машинное обучение с учителем (supervised machine learning) и машинное обучение без учителя (unsupervised machine learning). В первом из этих подходов строится классификатор, представляющий собой разделяющую функцию в пространстве признаков, на основе заранее подготовленной и размеченной обучающей текстовой коллекции. В качестве признаков, как правило, применяются отдельные слова и словосочетания. Во втором подходе используются такие лингвистические ресурсы, как словари оценочной лексики, содержащие отобранные экспертами в заданной предметной области термины, выражающие в тексте субъективное отношение.

Каждый подход обладает достоинствами и недостатками. Машинное обучение с учителем позволяет относительно быстро строить высокоточный классификатор, но требует наличия размеченной текстовой коллекции, а построенный классификатор сложно интерпретируется и его решения трудны для понимания пользователем. Машинное обучение без учителя не нуждается в обучающих коллекциях, процесс решения прозрачен для пользователя, однако создание словарей оценочной лексики и организация их применения для анализа весьма трудоемки.

В работах [3, 5] было предложено использовать для анализа тональности текстов ДСМ-метод автоматического порождения гипотез [1], позволяющий получать высокое качество классификации, организовывать учет контекста и при этом легко интерпретируемый. Однако серьезной проблемой применения ДСМ-метода является сложность обработки в процедуре аналогии огромного количества гипотез, порождае-

мых в процесс индуктивного вывода. В диссертации [7] осуществлялся отбор гипотез на основе специального генетического алгоритма. Однако для задач анализа текстовой информации, где количество гипотез может быть порядка 10^5–10^6, такой переборный подход не применим. В работе [2] предлагается генерировать ограниченное количество гипотез с использованием цепей Маркова. Но при этом нет гарантии, что полученное множество гипотез удовлетворяет требованиям пользователя, например хорошо объясняет обучающие примеры.

Целью данной статьи является разработка функции оценки информативности гипотез, на основе которой возможно эффективно обрабатывать большое число гипотез в ДСМ-методе.

Статья организована следующим образом. Во втором разделе описывается ДСМ-метод автоматического порождения гипотез и особенности его применения для анализа тональности текстов. В третьем разделе рассматриваются существующие функции оценки информативности гипотез. В четвертом разделе разрабатывается новая функция оценки. В пятом разделе приводятся описание и результаты экспериментов. В заключении изложены основные выводы и направления дальнейших исследований.

ДСМ-метод автоматического порождения гипотез

ДСМ-метод был разработан в конце 1970-х – начале 1980-х гг. В.К. Финном [6] и назван в честь английского ученого Д.С. Милля (1806–1873). В этом методе синтезируются три познавательные процедуры – индукция, аналогия и абдукция [1]. На вход метода поступают исходные обучающие данные – множество известных объектов, часть которых обладает интересующими исследователя целевыми свойствами, а часть – нет. Объекты представляются множеством признаков; при этом должна быть определена операция сходства объектов с учетом данных признаков. Кроме обучающих объектов ДСМ-метод может обрабатывать множество неопределенных объектов, для которых известно признаковое представление, но факт наличия или отсутствия интересующих свойств не определен.

В процедуре индукции на основе обучающих данных порождается множество гипотез, являющихся кандидатами в причины наличия или отсутствия целевых свойств. Гипотезой называется максимальное подмножество признаков, присутствующее в нескольких объектах, одинаково обладающих (или не обладающих) целевым свойством. В процедуре аналогии порожденные гипотезы применяются для установления факта наличия или отсутствия свойств у неопределенных объектов. Процедура абдукции служит для принятия гипотез, которые объясняют причины обладания объектами интересующих свойств. Таким образом, в ДСМ-методе решаются две основные задачи: во-первых, для неопределенных объектов устанавливается факт обладания ими целевых свойств; во-вторых, выявляются возможные причины наличия/отсутствия свойств у объектов.

В контексте задачи анализа тональности объектами служат тексты, признаками – слова и словосочетания, а свойствами являются позитивная и негативная тональности.

Функции оценки информативности гипотез

Множество гипотез, порождаемое в процедуре индукции ДСМ-метода, для большинства практических задач является весьма большим. Например, при анализе тональности отзывов о фильмах на основе коллекции, содержащей 150 текстов, и словаря из 850 слов было сгенерировано 220 000 гипотез [4]. Обработка такого огромного числа гипотез в процедуре аналогии требует аккуратного отбора, который может осуществляться на основе применения функций оценки информативности гипотез.

Существует несколько вариантов таких функций [10, p. 28–36; 11, p. 47–55; 12]. Для их описания введем следующие понятия и обозначения. Назовем положительным примером для некоторого целевого свойства объект, обладающий этим свойством; отрицательным примером – объект, не обладающий данным свойством. Будем считать, что пример распознается гипотезой, если все признаки, входящие в гипотезу, присутствуют в составе объекта. Обозначим: p – количество положительных примеров, распознаваемых гипотезой; n – количество отрицательных примеров, распознаваемых гипотезой; P – общее количество положительных примеров; N – общее количество отрицательных примеров.

Основные функции оценки информативности гипотез приведены в табл. 1.

Предварительные экспериментальные исследования представленных в табл. 1 функций показали, что наилучшие результаты демонстрирует функция RF (Relevance frequency) [12]. Однако ни одна из рассмотренных функций не учитывает особенности анализа текстов, поэтому было принято решение разработать новую функцию оценки информативности гипотез, позволяющую повысить качество классификации по сравнению с известными функциями.

Таблица 1

Основные функции оценки информативности гипотез

Название (русский язык)	Название (английский язык)	Функция
Правильность	Accuracy	$Accuracy = \frac{p + (N - n)}{P + N}$
Взвешенная относительная правильность	Weighted relative accuracy	$WRA = \frac{p}{P} - \frac{n}{N}$
Точность	Precision	$Precision = \frac{p}{p + n}$
Полнота	Recall	$Recall = \frac{p}{P}$
Информативность	Information content	$IC = -\log_2 \frac{p}{p + n}$
Индекс Джини	Gini index	$Gini = 1 - \left(\frac{p}{p + n}\right)^2 - \left(\frac{n}{p + n}\right)^2$
G-мера	G-measure	$G = \frac{p}{n + P}$
Оценка Лапласа	Laplace estimate	$LAP = \frac{p + 1}{p + n + 2}$
Релевантная частота	Relevance frequency	$RF = \log_2 \left(2 + \frac{p}{\max(1, n)}\right)$

Разработка функции оценки информативности для анализа текстов

Новую функцию предлагается разрабатывать на основе функции релевантной частоты *RF*, показавшей лучшие результаты в ходе предварительных исследований. Информативность (вес) гипотезы должна зависеть не только от коллекции в целом (параметры *p*, *n*, *P*, *N*), но и от особенностей рассматриваемого в данный момент текста. Поэтому в новую функцию включаются, во-первых, компонент, уменьшающий вес гипотезы, если входящие в неё признаки (слова) расположены в тексте далеко друг от друга; во-вторых, компонент, изменяющий вес гипотезы, в которую входят слова, присутствующие в словаре оценочной лексики с учетом тональности и части речи данного слова в текущем тексте. Таким образом, новая функция оценки информативности *SAW* (Sentiment analysis weight) гипотезы *h* относительно текста *t* имеет следующий вид:

$$SAW = k_{sent} \cdot \frac{\log_2 \left(2 + \frac{p}{\max(1, n)}\right)}{\log_2 (Dist_{av} + 1)}, \quad (*)$$

где k_{sent} – коэффициент оценочной лексики, учитывающий наличие в гипотезе *h* слов из словаря оценочной лексики; $Dist_{av}$ – среднее расстояние между словами гипотезы *h* в текущем тексте *t*.

Коэффициент k_{sent} увеличивает вес гипотезы на некоторую величину (эмпирически установлено, что оптимальное значение этой величины равно 3) в случае если в позитивную гипотезу входит позитивное слово из словаря оценочной лексики и уменьшает вес в том случае, если в ту же гипотезу входит негативное слово. Среднее расстояние $Dist_{av}$ между словами гипотезы в тексте вычисляется на основе подсчета слов текста, разделяющих слова гипотезы. Логарифм среднего расстояния вводится для компенсирования больших значений.

Результаты экспериментов

Экспериментальное исследование разработанной функции оценки информативности гипотез *SAW* и её сравнение с существующими функциями проводилось с использованием текстовых коллекций Российских семинаров по оценке методов информационного поиска РОМИП-2011 [8] и РОМИП-2012 [9]. Указанные коллекции включают отзывы о фильмах, книгах и цифровых фотокамерах. Имеются обучающие и тестовые коллекции по всем трем предметным областям. В ходе исследования

ДСМ-метод выполнялся с целью определения тональности для всех тестовых коллекций с разными функциями оценки информативности гипотез. Качество анализа оценивалось на основе метрики *F1-measure* [8]. Результаты экспериментов приведены в табл. 2.

Таблица 2

Результаты экспериментального исследования функций оценки информативности на основе тестовых коллекций семинаров РОМИП (метрика *F1-measure*, %)

Функция	РОМИП-2011			РОМИП-2012			Среднее значение
	Фильмы	Книги	Фотокамеры	Фильмы	Книги	Фотокамеры	
SAW	**74,62**	69,66	82,21	**72,76**	**75,96**	**75,94**	**75,19**
RF	73,88	70,08	**84,46**	69,63	75,31	71,78	74,19
Laplace	74,55	70,95	83,21	69,34	73,37	72,32	73,96
Precision	65,81	**73,19**	81,79	68,13	73,60	71,51	72,34
WRA	60,77	57,32	71,39	61,30	69,51	58,46	63,13
G-measure	58,00	56,40	79,90	53,45	64,50	52,56	60,80
Recall	57,76	57,71	70,49	59,95	63,72	53,52	60,53
Accuracy	58,72	50,76	73,13	50,90	52,22	55,99	56,95
Gini index	55,07	52,10	58,25	50,02	57,95	51,93	54,22
IC	50,28	51,86	49,41	39,03	42,17	43,68	46,07

В табл. 2 жирным выделены максимальные значения по каждой коллекции. Из таблицы 2 видно, что разработанная функция оценки *SAW* превосходит остальные функции для четырех предметных областей: *фильмы – 2011*, *фильмы – 2012*, *книги – 2012* и *фотокамеры – 2012*, а также имеет наивысшее среднее значение метрики *F1-measure* (последний столбец табл. 2).

Заключение

Таким образом, в статье предложена новая функция оценки информативности гипотез *SAW*, позволяющая с высокой точностью вычислять вес гипотезы по отношению к анализируемому тексту. Отличия от существующих функций оценки заключаются, во-первых, в учете расположения слов гипотезы в рассматриваемом тексте, во-вторых, во введении специального коэффициента, учитывающего присутствие в гипотезе оценочной лексики. Проведенные эксперименты с использованием текстовых коллекций семинаров РОМИП-2011 и РОМИП-2012 подтвердили эффективность разработанной функции оценки.

В дальнейшем планируется исследовать применимость новой функции для выявления закономерностей в исходных данных.

Список литературы

1. Автоматическое порождение гипотез в интеллектуальных системах / Сост. Е.С. Панкратова, В.К. Финн; под общ. ред. В. К. Финна. – М.: Книжный дом «ЛИБРОКОМ», 2009. – 528 с.

2. Виноградов Д. В. Вероятностное порождение гипотез в ДСМ-методе с помощью простейших цепей Маркова // НТИ. Сер. 2. Информационные процессы и системы. – 2012. – № 9. – С. 20–27.

3. Котельников Е.В. Классификация отзывов о фильмах с использованием ДСМ метода // В мире научных открытий. – 2013. – № 6.1 (42). – С. 225–242.

4. Котельников Е.В. Повышение быстродействия ДСМ-метода в задачах обработки текстовой информации // Труды Четырнадцатой национальной конференции по искусственному интеллекту с международным участием КИИ-2014 (24–27 октября 2014 года, г. Казань). – Т.2. – Казань: Изд-во РИЦ «Школа», 2014. – С. 274–282.

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

6. Финн В.К. О возможностях формализации правдоподобных рассуждений средствами многозначных логик // Всесоюз. симпозиум по логике и методологии науки. – Киев: Наукова думка, 1976. – С. 82–83.

7. Шашкин Л. О. Приближенные средства установления сходств для ДСМ-метода автоматического порождения гипотез: автореф. дис. ... канд. техн. наук. – М., 2000. – 26 с.

8. 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). – P. 739–746.

9. Chetviorkin I., Loukachevitch N. Sentiment Analysis Track at ROMIP 2012 // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference «Dialog 2013». – Bekasovo, 2013. – Vol. 2. – P. 40–50.

10. Fürnkranz J. Separate-and-Conquer Rule Learning // Artificial Intelligence Review. – 1999. – Vol. 13. – P. 3–54.

11. Fürnkranz J., Flach P. A. ROC ‘n’ Rule Learning-Towards a Better Understanding of Covering Algorithms // Machine Learning. – 2005. – Vol. 58(1). – P. 39–77.

12. Lan M., Tan C. L., Su J., Lu Y. Supervised and Traditional Term Weighting Methods for Automatic Text Categorization // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 2009. – Vol. 31. – № 4. – P. 721–735.

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

References

1. *Avtomaticheskoe porozhdenie gipotez v intellektual'nyh sistemah* [Automatic hypothesis generation in intelligence systems] Eds. E.S. Pankratova, V.K. Finn. Moscow, LIBROKOM, 2009, 528 p.

2. Vinogradov D.V. *Nauchno-tehnicheskaja informacija. Serija 2. Informacionnye processy i sistemy*, 2012, no. 9, pp. 20–27.

3. Kotelnikov E.V. *V mire nauchnyh otkrytij*, 2013, no. 6.1 (42), pp. 225–242.

4. Kotelnikov E.V. *Trudy Chetyrnadcatoj nacional'noj konferencii po iskusstvennomu intellektu s mezhdunarodnym uchastiem KII-2014* [Proceedings of the 14th national conference on artificial intelligence with international participation]. Kazan, School, 2014, pp. 274–282.

5. Kotelnikov E.V. *Nauchno-tehnicheskij vestnik Povolzh'ja*, 2013, no. 6, pp. 344–346.

6. Finn V.K. *Vsesojuznyj simpozium po logike i metodologii nauki* [all-USSR symposium on logic and science methodology]. Kiev, Scientific thought, 1976, pp. 82–83.

7. Shashkin L.O. *Priblizhennye sredstva ustanovlenija shodstv dlja DSM-metoda avtomaticheskogo porozhdenija gipotez* [Approximate means of establishing similarities to DSM-method of automatic hypothesis generation], Moscow, 2000, 26 p.

8. Chetviorkin I., Braslavskiy P., Loukachevitch N. *Annual International Conference «Dialogue»*, 2012, no. 11(18), pp. 739–746.

9. Chetviorkin I., Loukachevitch N. *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference «Dialog 2013»*, 2013, vol. 2, pp. 40–50.

10. Fürnkranz J. *Artificial Intelligence Review*, 1999, Vol. 13, pp. 3–54.

11. Fürnkranz J., Flach P.A. *Machine Learning*, 2005, Vol. 58(1), pp. 39–77.

12. Lan M., Tan C. L., Su J., Lu Y. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2009, Vol. 31, no. 4, pp. 721–735.

13. Liu B. *Synthesis Lectures on Human Language Technologies*, 2012, Vol. 5(1).

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.

Работа поступила в редакцию 18.11.2014.