ЖУРНАЛ РАДИОЭЛЕКТРОНИКИ, N 10, 2010

УДК 004.93'1

ПРИМЕНЕНИЕ КОНТЕКСТНЫХ ВЕКТОРОВ В КЛАССИФИКАЦИИ ТЕКСТОВЫХ ДОКУМЕНТОВ

А.С. Епрев

Омский государственный университет им. Ф.М. Достоевского

Получена 27 сентября 2010 г.

Аннотация. В работе представлен подход к решению задачи классификации текстовых документов на базе классификатора k–ближайших соседей с использованием информации о значениях слов, полученных из словаря WordNet методом контекстных векторов. Полученные в ходе экспериментов результаты подтверждают эффективность предложенного подхода.

Ключевые слова: классификация текстовых документов, разрешение лексической многозначности, контекстный вектор, WordNet.

Введение

Классификация текстовых документов является задачей автоматического определения документа в одну или несколько категорий на основании его содержания. Классификатор автоматически создается в процессе обучения, при котором просматривается множество документов с заранее определенными категориями. Существуют различные методы классификации текстов — деревья решений, метод наименьших квадратов, адаптивные линейные классификаторы, метод ближайших соседей, метод опорных векторов и другие [1].

Последние несколько лет большой интерес представляет интеграция различных баз знаний в методы классификации текстовых документов [2–4]. От части это обусловлено, тем что такие источники данных становятся доступными в

электронной форме. Широкую популярность получил ресурс WordNet.

Словарь WordNet

WordNet это семантический словарь английского языка, базовой словарной единицей которого является синонимический ряд, так называемый «синсет», объединяющий слова со схожим значением. Синсеты связаны между собой различными семантическими отношениями. WordNet содержит приблизительно 155 тысяч различных лексем и словосочетаний, организованных в 117 тысяч синсетов, разбитых по частям речи: существительные, глаголы, прилагательные и наречия.

Чтобы воспользоваться информацией WordNet в классификаторе, необходимо решить задачу устранения лексической многозначности слов. Разрешение лексической многозначности (Word Sense Disambiguation, WSD) — это задача выбора значения (концепта) многозначного слова или фразы из множества их значений (концептов) в зависимости от контекста, в котором данное слово находится. Одним из эффективных методов устранения лексической многозначности на базе WordNet является метод, основанный на оценке семантической близости концептов WordNet с помощью контекстных векторов второго порядка [5].

Контекстные векторы

В определении значений слов существенную роль играет контекст. Одно и тоже значение слова, как правило, употребляется в одинаковом контексте. Контекстные векторы широко используются в информационном поиске и в задачах обработки естественного языка. Контекстный вектор $\vec{w}$ указывает на все слова вместе с которыми слово w встречается в тексте. Векторы, сформированные из контекстных векторов (контекстные векторы второго порядка), можно использовать для представления значений слов [6].

Чтобы построить контекстные векторы второго порядка (векторы дефиниций) для синсетов WordNet, необходимо определить пространство слов W . Оно обычно представляется матрицей, строки которой являются контекстными векторами первого порядка. Значения на пересечениях строк и столбцов указывают на частоты совместной встречаемости двух слов в тексте. Определив пространство слов, контекст можно представить как сумму контекстных векторов первого порядка слов, определяющих этот контекст.

Итак, пространство слов W определяется множеством контекстных векторов первого порядка. Чтобы построить контекстный вектор первого порядка для слова w, необходимо последовательно выполнить следующие действия:

1. Инициализировать контекстный вектор первого порядка $\vec{w}$ нулевыми значениями.
2. Найти каждое вхождение слова w в тексте.
3. Для каждого вхождения слова w увеличить значения вектора $\vec{w}$ в позициях соответствующих словам из пространства слов, которые находятся на заданном расстоянии от слова w в тексте.

Таким образом, контекстный вектор первого порядка $\vec{w}$ содержит информацию о совместной встречаемости слова w.

В качестве корпуса текстов для построения контекстных векторов первого порядка используются дефиниции синсетов WordNet. Такой корпус содержит приблизительно 1,4 миллиона слов, а размерность пространства слов составляет порядка 20 тысяч (без учета редко встречающихся и стоп-слов).

Классификация текстов с использованием контекстных векторов

Чтобы определить влияние на эффективность классификации механизма разрешения лексической многозначности слов на базе контекстных векторов второго порядка, необходимо построить два классификатора одного типа, таких что, пространство признаков первого составляли базовые словоформы, а второго

— синсеты WordNet. За основу классификаторов был взят метод k–ближайших соседей [7]. Он показывает высокие результаты классификации и сравнительно прост в реализации.

На этапе индексирования документа d_j происходит выделение термов с использованием морфологического анализа. Для каждого терма t_i документа d_j вычисляется весовой коэффициент по формуле:

$$\omega_{i,j} = \frac{tf_{ij} \cdot idf_i}{\sqrt{\sum_k \left(tf_{kj} \cdot idf_k\right)^2}},$$

где ω_{ij} — вес i-го терма в документе d_j, tf_{ij} — частота встречаемости i-го терма в рассматриваемом документе, $idf_i = \log\left((1+N)/(1+n)\right)$ — логарифм отношения количества документов в коллекции к количеству документов, в которых встречается i-ый терм. Веса нормализованы таким образом, что сумма квадратов весов каждого документа равна единице.

Для того чтобы найти категории соответствующие документу d_j, классификатор выполняет следующие действия:

1. Документ d_j сравнивается со всеми документами d_z из обучающей коллекции и вычисляется расстояние между документами — значение косинуса угла между векторами $\vec{d}_j$ и $\vec{d}_z$.
2. Выбираются k ближайших к d_j документов.
3. Определение категорий документа d_j осуществляется выбором наиболее встречающихся категорий среди k ближайших к d_j документов.

В построенных классификаторах значение k равняется 30.

Во втором классификаторе в качестве признаков документов необходимо использовать значения термов, представленных синсетами WordNet. Определение

значения терма t_i осуществляется следующим образом:

1. Вычисляется контекст для терма t_i. Контекст определяется суммой контекстных векторов первого порядка слов, находящихся на расстоянии в пять позиций слева и справа от терма t_i в документе.
2. Производится оценка семантической близости всех возможных концептов t_i. Для каждого концепта вычисляется косинус угла между вектором его дефиниции и контекстом.
3. Самый близкий концепт s_i выбирается в качестве значения терма t_i.

Результаты экспериментов

В табл. 1 приведены результаты первого эксперимента. Построение классификаторов и оценка их эффективности проводилась с использованием разбиения «ModApte» коллекции документов «Reuters–21578» [8]. Это разбиение задает 90 категорий, 9603 документа содержатся в обучающем наборе и 3299 документов в тестирующем.

Табл. 1. Эффективность классификаторов на коллекции «Reuters–12578».

Классификатор	Микро *p*	Микро *r*	Макро *p*	Макро *r*
№1	**.8340**	**.7727**	**.8939**	**.2993**
№2	.8240	.7650	.8917	.2950

Как видно, использование в классификаторе механизма разрешения лексической многозначности привело к незначительным потерям эффективности. Это объясняется тем, что корпус текстов «Reuters–21578» содержит в большинстве своем тексты экономического характера, т.е. термины, встречающиеся в документах, употребляются, как правило, в одних и тех же значениях.

Для повторного эксперимента было решено использовать корпус текстов

«Reuters Corpus Volume 1» (RCV1) [9]. В отличии от коллекции «Reuters–21578», для RCV1 не определены стандартные разбиения на обучающие и тестирующие множества. Для эксперимента были выбраны 10 разносторонних категорий: международные отношения (GDIP); катастрофы и бедствия (GDIS); искусство, культура и развлечения (GENT); мода (GFAS); здоровье (GHEA); религия (GREL); наука и технологии (GSCI); спорт (GSPO); путешествия и туризм (GTOUR) и погода (GWEA).

Из всей коллекции были отобраны 5923 документа, определенных в одну или несколько вышеперечисленных категорий, и разделены на два множества. Обучающий набор содержит 3532 документа, тестовый набор — 1761.

Табл. 2. Эффективность классификаторов на коллекции «RCV1».

Классификатор	Микро *p*	Микро *r*	Макро *p*	Макро *r*
№1	.8499	**.8569**	.8610	.8230
№2	**.8564**	**.8569**	**.8759**	**.8236**

В табл. 2 приведены результаты повторного эксперимента, из которых видно, что точность классификации увеличилась на 1–2% при небольшом росте полноты. В этот раз механизм разрешения лексической многозначности оказал положительное влияние на результаты классификации.

Заключение

Таким образом, можно сделать вывод, что использование механизма разрешения лексической многозначности слов на базе контекстных векторов второго порядка в текстовом классификаторе приводит к улучшению эффективности классификации на корпусе разносторонних текстов, и к снижению на корпусе текстов узкой направленности. Разработанный метод работает без дополнительного механизма распознавания части речи слов в документах, что,

возможно, снижает его эффективность, но позволяет ускорить обработку документов.

Список литературы

1. Sebastiani F. Text Categorization // Text Mining and Its Applications. — 2005. P.109–129.
2. Barak L., Dagan I. Shnarch E. Text categorization from category name via lexical reference // Proceedings of Human Language Technologies. — 2009. P.33–36.
3. Gomez J.M., Buenaga M., Urena L.A., Martin M.T., Garcia M. Integrating Lexical Knowledge in Learning-Based Text Categorization // Proceedings of the 6th International Conference on the Statistical Analysis of Textual Data. — 2002. 10 p.
4. Gomez J.M., Buenaga M. Integrating a Lexical Database and a Training Collection for Text Categorization // Proceedings of ACL-EACL. — 1997. 12 p.
5. Patwardhan S., Pedersen T. Using WordNet-based context vectors to estimate the semantic relatedness of concepts // EACL 2006 Workshop Making Sense of Sense. — 2006. P.1–8.
6. Schutze H. Automatic word sense discrimination // Computational Linguistics. — 1998. — V. 24. P.97–123.
7. Yang Y. Expert network: effective and efficient learning from human decisions in text categorization and retrieval // Proceedings of SGIR-94. — 1994. — P.13–22.
8. Lewis D. The Reuters-21578 text categorization test collection. — 1999. E-print: http://www.daviddlewis.com/resources/testcollections/reuters21578/
9. T.G. Rose, M. Stevenson and M. Whitehead. The Reuters Corpus Volume 1 — from Yesterday's News to Tomorrow's Language Resources // Third International Conference on Language Resources and Evaluation. — 2002. 7 p.