УДК 004.045

ИЗВЛЕЧЕНИЕ КЛЮЧЕВЫХ СЛОВОСОЧЕТАНИЙ

С.В. Попова, И.А. Ходырев

Исследованы задачи аннотирования ключевыми словами и словосочетаниями семантически близких групп текстов в маленьких коллекциях узкотематических документов короткой длины. Показана возможность извлечения ключевых слов с одновременной фильтрацией общеупотребительных слов. Предложена формула для оценки взаимной зависимости слов и алгоритм извлечения на ее основе ключевых словосочетаний. Представлены результаты тестирования используемых алгоритмов.

Ключевые слова: ключевые слова, ключевые словосочетания, аннотирование, кластеризация, анализ коротких текстов, информационный поиск.

Введение

Задача извлечения ключевых слов из коллекций документов и составление коротких аннотаций к коллекциям, их частям, отдельным текстам или частям текстов (topic detection, topic interpretatition, summarization, TextTiling) [1, 2] получила широкое распространение вследствие высокой применимости на практике. Извлечение ключевых слов и словосочетаний из коллекций и текстов позволяет пользователю понять, насколько полезен для него некоторый документ, не требуя просмотра всего документа. Используя ключевые словосочетания, пользователь может найти документы, релевантные заинтересовавшему его документу. Появление новых словосочетаний в подборках научных статей, ранжированных по времени, может свидетельствовать о появлении нового направления или новой тенденции в некоторой научной области. В научной области важной является задача извлечения новой терминологии развивающейся области. Данная задача напрямую связана с выделением устойчивых словосочетаний [3]. Определение основных тем и ключевых слов, представленных в коллекциях новостей, является подзадачей направления, получившего широкое распространение под названием «Topic detection and tracking» (TDT) [4–7] и связанного с отслеживанием во времени изменений, модификаций, группировкой и возникновением «нового» в новостях. Решение задачи извлечения ключевых словосочетаний полезно также в задачах определения различных контекстов слова, например, при построении словарей или разработке систем автоматического машинного перевода. Сложность поставленной задачи связана с тремя основными проблемами, присущими узкотематическим коллекциям текстов короткой длины: низкая частота встречи термов (слов) в текстах, большое перекрытие по общим словам, недостаток данных для накопления статистической информации. Из-за низкой частоты встречи термов, маленького размера коллекций и небольшой частоты совместной встречаемости слов существующие методы извлечения словосочетаний, например, основанные на вычислении MI (меры взаимной информации), могут оказываться неэффективными, так как, в первую очередь, будут выделяться словосочетания, в состав которых входят слова, редко встречающиеся в коллекции. Целью данной работы является разработка алгоритмов для извлечения семантически близких групп текстов коллекции и аннотирования полученных групп ключевыми словами и словосочетаниями.

Постановка задачи и этапы решения

В работе рассматривается задача автоматического аннотирования коллекции документов ключевыми словами и словосочетаниями, характеризующими группы семантически близких документов данной коллекции. Извлечение из коллекции, помимо самих ключевых слов, контекстных словосочетаний с этими словами позволяет лучше отследить семантику использования выделенных слов. Задача извлечения ключевых словосочетаний и слов была разделена на два этапа: этап кластеризации и этап аннотирования ключевыми словосочетаниями полученных кластеров. В области кластеризации коротких текстов интересны работы [8–10]. Разработка собственного алгоритма потребовалась по следующим причинам. Если требуется ориентироваться на пользователя, результат кластеризации должен быть всегда достаточно высоким и желательно стабильным (не меняться при различных прогонах алгоритма). Стабильная работа алгоритма кластеризации необходима при использовании результатов кластеризации для аннотирования. В противном случае будут получаться различные аннотации для разных результатов кластеризации. Современные алгоритмы [8–10], дающие достаточно высокие результаты кластеризации для рассматриваемого типа коллекций, не дают этот результат стабильно (в случае фиксированного числа кластеров): при одном прогоне алгоритма результат может быть высоким, при другом относительно низким. В экспериментах авторов использованы те же коллекции, что и в работах [8–10]: CICling_2002, SEPLN_CICling, EasyAbstracts и Micro4News. Первые три коллекции содержат в себе аннотации научных статей и являются узкотематическими; последняя коллекция содержит короткие новости и относится к широкотематическим коллекциям. Названия внутренних тем коллекций (названия кластеров, которые должны быть получены на типе кластеризации) представлены в табл. 1. Данные коллекции, включая «Золотые стандарты» (Golden Standard, результат классификации коллекции экспертом, когда человек определяет, какие семантически близкие группы документов содержатся в коллекции), находятся в открытом доступе [11], где также можно найти информацию с описанием коллекций.

Название Коллекции	Названия внутренних тем коллекции
CICling_2002	Linguistic, Ambiguity, Lexicon and Text, Processing
SEPLN_CICling	Morphological – syntactic, analysis, Categorization of Documents, Corpus linguistics, Machine translation
Easy Abstracts	Machine Learning, Heuristics in Optimization, Automated reasoning and Autonomous intelligent agents
Micro4News	Sci.med, soc.religion.christian, rec.autos, comp.os.ms-windows

Таблица 1. Названия внутренних тем коллекций

Решение задачи кластеризации

Для решения задачи кластеризации были рассмотрены алгоритмы иерархической кластеризации: Single Linkage и Complete Linkage [12], Between Groups Linkage (UPGMA [13]. Исследовалось влияние на качество кластеризации данными алгоритмами сужения пространства кластеризации (пространства признаков, задаваемого словарем коллекции). Рассматривалось сужение пространства кластеризации, основанное на удалении термов с низкими и высокими значениями document frequency (где значением document frequency для терма является число документов, в которых данный терм встретился). Для оценки расстояния/подобия между двумя текстами были рассмотрены: расстояние Эвклида, Jaccard index, косинус угла между векторами, корреляция Пирсона. Описание данных оценок расстояния/подобия и результаты использования их для некоторых коллекций можно найти в работе [14].

Требовалось выбрать метод, дающий самые высокие результаты в среднем, при условии, что изначально точное число кластеров неизвестно и может быть задано в интервале от 3 до 8. В результате проведенных экспериментов с использованием описанных выше тестовых коллекций были выбраны метод Between Groups Linkage и коэффициент корреляции Пирсона. Данные алгоритм и оценка подобия легли в основу алгоритма кластеризации, предложенного в работе [15] и используемого в настоящей работе. Этот алгоритм использует векторную модель представления текстов.

Удалось добиться достаточно хорошей работы алгоритма за счет специального подбора техники отбора терминов (terms selection), которая позволяет сузить пространство кластеризации. Идея принципа отбора термов в том, что нужно из всего словаря коллекции оставить только такие термины, которые вносят «положительный» вклад при вычислении корреляции между двумя текстами. К таким терминам не относятся термины, встречающиеся в большом числе документов, так как эти термины усиливают корреляцию внутри группы текстов, заведомо превышающей размер наибольшего из кластеров. Для рассматриваемого в работе типа коллекций таких слов немного (обычно 2–4 слова). С другой стороны, слова, встречающиеся в коллекции редко, не несут в себе информации о корреляции. Именно таких слов оказывается большинство. Например, удаление слов, встречающихся менее чем в 3–4 текстах, может приводить к тому, что от всего словаря коллекции остается порядка 10% слов. Из-за большого числа таких слов в векторах, представляющих тексты, оказывается большое число не взаимосвязанных ненулевых значений, что оказывает на результат вычисления корреляции эффект шума. Исходя из этих наблюдений, можно считать, что целесообразно сужать пространство кластеризации с помощью удаления из словаря коллекции 90% слов с самыми низкими значениями document frequency и 3–4 слова с самыми высокими значениями document frequency.

Выбор иерархической кластеризации оправдан тем, что с ее помощью можно получить стабильный результат кластеризации, зависящий только от определения числа кластеров. Так как точное определение числа кластеров часто затруднительно, было введено предположение, что для используемых в настоящей работе коллекций число кластеров может быть определено в интервале от 3 до 8. Использование описанных выше средств (алгоритма кластеризации, меры подобия между текстами и техники отбора терминов) позволило получить результаты, сравнимые с результатами кластеризации алгоритмами, опубликованными в работах [8–10]. Оценка результатов кластеризации проводилась с помощью той же меры

$$measure(F) = \sum_i \frac{G_i}{|D|} \max F_{ij} \ , \ F_{ij} = \frac{2 \cdot P_{ij} \cdot R_{ij}}{P_{ij} + R_{ij}} \ , \ P_{ij} = \frac{|G_i \cap C_j|}{G_i} \ , \ R_{ij=} \frac{|G_i \cap C_j|}{C_j} \ .$$

Здесь G_i задает кластеры, полученные в результате автоматической обработки; C_j задает кластеры, выделенные экспертами; $|D|$ – размер коллекции. В табл. 2 приведена оценка результатов кластеризации описанным выше алгоритмом (*) в лучшем (Max), в худшем случае (Min), и в среднем (Avg) в зависимости от того, какое число кластеров порождалось (от 3 до 8 кластеров). Для сравнения в табл. 2 приводятся данные, опубликованные в работе [8] для алгоритма AntSA-CLU и в работе [9] для алгоритма CLUDISPO.

Название Коллекции		CICling_2002	SEPLN_CICling	Easy Abstracts	Micro4News
*	Max	0,73	0,84	0,82	0,96
	Avg	0,65	0,72	0,79	0,87
	Min	0,59	0,65	0,72	0,79
CLUDISPO	Max	0,73	0,85	0,98	1
	Avg	0,6	0,72	0,92	0,93
	Min	0,47	0,58	0,85	0,85
AntSA-CLU	Max	0,75	0,85	0,98	1
	Avg	0,61	0,75	0,96	0,96
	Min	0,47	0,63	0,92	0,88

Таблица 2. Оценка результата автоматической кластеризации для тестовых коллекций

Результаты работы алгоритмов K-Means [12], MajorClust [16], DBSCAN [17] в таблице не приводятся, так как качество кластеризации с помощью этих алгоритмов хуже, чем с помощью AntSA-CLU и CLUDISPO. Для последних алгоритмов результаты могут быть найдены в работе [8].

Решение задачи выделения ключевых слов и словосочетаний

Второй частью работы является задача выделения ключевых слов для кластеров и определение контекста использования выделенных слов (определение ключевых словосочетаний для найденных слов, биграмм, колокаций). Отбор происходит только среди слов, полученных в результате сужения пространства кластеризации. В основе алгоритма лежит простая идея: словами, характеризующими тематическую направленность кластера, являются слова, встречающиеся в большом числе документов данного кластера и в малом числе документов за его пределами. Слово отбирается как ключевое для кластера, если число текстов, в которых частота слова меньше α (мы выбирали $\alpha = 4$), не превышает размер кластера, и если данное слово не встречается только в β текстах данного кластера. В работе [15] показано, что эти два условия позволяют выделить ключевые для кластера слова, которые встречаются во многих документах кластера и появляются в ряде документов за пределами кластера (первая группа слов). Эти же два условия позволяют отсеять слова, частые в нескольких кластерах и типичные для данной коллекции (вторая группа слов). Связано это с тем, что практически нет текстов, в которых слова из второй группы встречались бы часто, в отличие от слов первой группы.

Параметр β является подвижным, его увеличение приводит к выделению большого числа ключевых слов, однако качество выделяемых слов при этом падает. В работе данный параметр автоматически изменялся в зависимости от того, какое число ключевых слов требуется выделить. В настоящей работе на этапе выделения ключевых слов для каждого кластера отбирались ключевые слова до тех пор, пока не было отобрано как минимум 10 слов. Из каждых отобранных слов на следующем этапе оставлялось всего 3 слова, встречающихся в наибольшем числе документов кластера.

В работе для решения задачи выделения словосочетаний предлагается оценка взаимосвязанности слов в виде

$$c = \frac{t_1 + t_2}{2 \cdot f(t^1 t^2)},$$

где t_1 и t_2 отражает число различных пар с первым или вторым словом из рассматриваемого словосочетания, $f(t^1 t^2)$ отражает число появлений словосочетания типа «первое–второе слово вместе». Чем меньше значение c, тем лучшей считается пара слов (t^1, t^2). Словосочетанием, поясняющим ключевое слово, является устойчивое словосочетание, где одно из слов является ключевым. Пара слов (t_1, t_2) является устойчивым словосочетанием, если для слов t_1 и t_2 величина c меньше некоторого фиксированного порога, а сами значения t_1, t_2, $f(t^1 t^2)$ больше 3. Максимальное пороговое значение, используемое в работе, равно 20. Хорошим является также порог, равный 15, менее хорошим – 30, при пороге, равном 10, очень мало словосочетаний, определяемых как устойчивые. Идея, положенная в основание предложенной формулы, состоит в следующем: устойчивые словосочетания часто встречаются вместе и редко порознь.

В настоящей работе приводятся результаты, полученные для коллекции Easy Abstracts. В табл. 3 приведены результаты для случая восьми кластеров (при числе кластеров в диапазоне от 3 до 8 результаты похожи). Пустые графы в табл. 3 говорят о том, что кластеры содержат всего 1–2 документа, т.е. такого числа документов недостаточно для выделения ключевых слов кластера.

Ключевые слова	Ключевые словосочетания
search, objective, function	objective+single, objective+genetic objective+multi, search+local search+space, search+tabu function+approximation
proof, theorem, based	theorem+proving
agent, models, agents	agent+oriented, agent+communication, agent+patterns, agents+esrl
learning, machine, function	machine+learning, machine+boltzmann machine+tabu, function+approximation, learning+machine, learning+classifier earning+reinforcement
selection, large, probabilistic	large+data, selection+feature, probabilistic+svm
---	---
---	---
---	---

Таблица 3. Результат аннотирования ключевыми словосочетаниями коллекции Easy Abstracts в случае восьми кластеров

Просмотрев документы коллекции Easy Abstracts, можно убедиться, что найденные словосочетания являются типичными для конкретных кластеров этой коллекции. К сожалению, у нас сейчас нет иной меры оценки качества найденных словосочетаний, чем оценка человеком-экспертом.

Заключение

В работе предложен алгоритм для автоматического аннотирования узкотематических маленьких коллекций коротких текстов, заключающийся в извлечении из коллекций ключевых словосочетаний и ключевых слов на основе предварительной кластеризации. Предложены алгоритмы кластеризации, аннотирования полученных кластеров ключевыми словами и алгоритм выделения ключевых словосочетаний. Предложенный алгоритм кластеризации дает достаточно высокие результаты для указанных коллекций. С помощью алгоритма выделения ключевых словосочетаний удается выделить словосочетания, отражающие специфику каждого кластера коллекции. Авторам неизвестны попытки выделения ключевых слов для рассматриваемых в этой работе коллекций, в частности, для коллекции Easy Abstracts. Это затрудняет оценку качества представленного алгоритма. В дальнейшем планируется модификация предложенного в работе алгоритма выделения ключевых словосочетаний: для вычисления $f(t^1t^2)$ планируется использовать информацию об отдельных кластерах, а не информацию обо всей коллекции в целом, как это сделано в настоящей работе.

Литература

1. Lloret E. Topic Detection and Segmentation in Automatic Text Summarization [Электронный ресурс]. – Режим доступа: http://www.dlsi.ua.es/~elloret/publications/SumTopics.pdf, св. Яз. англ. (дата обращения 01.10.2011).
2. Teufel S., Moens M. Summarizing Scientific Articles: Experiments with Relevance and Rhetorical Status [Электронный ресурс]. – Режим доступа: http://acl.ldc.upenn.edu/J/J02/J02-4002.pdf, св. Яз. англ. (дата обращения 01.10.2011).
3. Ягунова Е.В., Пивоварова Л.М. Извлечение и классификация коллокаций на материале научных текстов. Предварительные наблюдения // V Международная научно-практическая конференция «Прикладная лингвистика в науке и образовании» памяти Р.Г. Пиотровского (1922–2009): Материалы. – СПб: 2010. – С. 356–364.
4. Makkonen J. Semantic Classes in Topic Detection and Tracking [Электронный ресурс]. – Режим доступа: http://www.doria.fi/bitstream/handle/10024/48180/semantic.pdf, св. Яз. англ. (дата обращения 01.10.2011).
5. Smith S.C., Rodr´ıguez M.A. Clustering-based Searching and Navigation in an Online News Source [Электронный ресурс]. – Режим доступа: http://captura.uchile.cl/jspui/bitstream/2250/6257/1/Smith_Simon.pdf, св. Яз. англ. (дата обращения 01.10.2011).
6. Shih C., Peng T. Building Topic / Trend Detection System based on Slow Intelligence [Электронный ресурс]. – Режим доступа: http://www.cs.pitt.edu/~chang/265/proj10/57shih.pdf, св. Яз. англ. (дата обращения 01.10.2011).
7. He Q., Chang K., Lim E., Banerjee A. Keep It Simple with Time: A Re-examination of Probabilistic Topic Detection Models [Электронный ресурс]. – Режим доступа: http://www-users.cs.umn.edu/~banerjee/papers/09/pami-tdt.pdf, св. Яз. англ. (дата обращения 01.10.2011).
8. Errecalde M., Ingaramo D., Rosso P. A new AntTree-based Algorithm for Clustering Short-text Corpora // Journal of Computer Science and Technology. – 2010. – V. 10. – № 1. – P. 1–7.

9. Ingaramo D., Cagnina L., Errecalde M., Rosso P. A Particle Swarm Optimizer to cluster short-text corpora: a performance study // Proc. Workshop on Natural Language Processing and Web-based Technologies, 12th edition of the Ibero-American Conference on Artificial Intelligence. IBERAMIA. – 2010. – P. 71–79.
10. Pinto D. Analysis of narrow-domain short texts clustering. Research report for Diploma de Estudios Avanzados. DEA // Department of Information Systems and Computation. UPV. – 2007 – [Электронный ресурс]. – Режим доступа: http://users.dsic.upv.es/~prosso/resources/PintoDEA.pdf, св. Яз. англ. (дата обращения 01.10.2011).
11. PLN Resources // Data Sets for Short-texts Experimental Works [Электронный ресурс]. – Режим доступа: https://sites.google.com/site/merrecalde/resources, св. Яз. англ. (дата обращения 01.10.2011).
12. Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze Introduction to Information Retrieval. – Cambridge University Press. – 2008. – C. 377–402.
13. Local methods – UPGMA (Unweighted Pair Group Method) // Phylogentics workshop 09: Methods in Bioinformatics and Molecular Evolution [Электронный ресурс]. – Режим доступа: http://www.adelaide.edu.au/acad/events/workshop/LockhartUPGMA&NJ_calculation.pdf, св. Яз. англ. (дата обращения 01.10.2011).
14. Huang A. Similarity Measures for Text Document Clustering. Department of Computer Science The University of Waikato, Hamilton, New Zealand [Электронный ресурс]. – Режим доступа: http://nzcsrsc08.canterbury.ac.nz/site/proceedings/Individual_Papers/pg049_Similarity_Measures_for_Text_Document_Clustering.pdf, св. Яз. англ. (дата обращения 01.10.2011).
15. 15.Popova S.V., Khodyrev I.A. Local theme detection and annotation with key words for narrow and wide domain short text collections // The Fifth International Conference on Advances in Semantic Processing. SEMAPRO. 2011. – Lisbon: Portugal, 2011. – P. 49–55.
16. 16.Stein B., Niggemann O. On the Nature of Structure and its Identification // In Proc. of the 25th International Workshop on Graph Theoretic Concepts in Computer Science. LNCS. – Springer-Verlag, 1999. – V. 1665. – P. 122–134.
17. Ester M., Kriegel H., Sander J., Xu X. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise // Proc. of 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96). – 1996. – P. 226–231.

Попова Светлана Владимировна – Санкт-Петербургский государственный университет, spbu@bk.ru
Ходырев Иван Александрович – ОЛИМП, программист, kivan.mih@gmail.com