УДК 340.11(3)

В. Б. Барахнин, Д. А. Ткачев

Институт вычислительных технологий СО РАН
пр. Акад. Лаврентьева, 6, Новосибирск, 630090, Россия

Новосибирский государственный университет
ул. Пирогова, 2, Новосибирск, 630090, Россия
E-mail: bar@ict.nsc.ru; relk-tda@yandex.ru

КЛАСТЕРИЗАЦИЯ ТЕКСТОВЫХ ДОКУМЕНТОВ НА ОСНОВЕ СОСТАВНЫХ КЛЮЧЕВЫХ ТЕРМОВ *

Классический подход к координатному индексированию текстов с целью их последующей кластеризации заключается в использовании средств анализа на основе тезауруса обрабатываемой предметной области. Но если вести речь об обработке корпусов текстов достаточно узких тематик, то в таких случаях требуются очень подробные тезаурусы, которые имеются (по крайней мере в широком доступе) далеко не для всех предметных областей. Подход же, основанный на извлечении ключевых выражений без априорных ограничений, носит гораздо более универсальный характер. Однако при таком подходе возникает проблема отбора ключевых термов. Цель данной работы заключается в демонстрации практических преимуществ кластеризации документов на основе ключевых словосочетаний по сравнению с популярной кластеризацией на основе анализа только однословных ключевых термов, при этом для выделения ключевых словосочетаний используются общедоступные программные средства, не требующие особых вычислительных затрат.

Ключевые слова: кластеризация текстовых документов, координатное индексирование, составные ключевые термы.

Введение

Постоянный рост объема научной информации, представленной в электронной форме, делает актуальным решение задачи разработки методики автоматизированного вовлечения электронных документов в научно-информационный процесс. Одним из важнейших этапов этого процесса является классификация документов, поскольку при отсутствии классификационных признаков поиск документа человеком или его обработка интеллектуальной информационной системой может опираться только на простую проверку вхождения тех или иных терминов в текст документа. К сожалению, даже наиболее структурированные документы – журнальные статьи – далеко не всегда содержат классификационные признаки, к тому же классификатор источника может не совпадать с классификатором, используемым создателями информационной системы.

Другой распространенный подход к удовлетворению информационных потребностей научных работников – поиск «по аналогии» – заключается в нахождении документов, которые в том или ином смысле аналогичны документу (или множеству документов), уже известному данному лицу (подробнее см., например, [1]).

В качестве шкал для определения меры сходства между двумя документами в обоих названных случаях можно использовать атрибуты библиографического описания данных документов (метаданные), а также элементы контента электронного документа, в частности клю-

* Работа выполнена при частичной поддержке РФФИ: проекты № 08-07-00229, 09-07-00277, 10-07-00302 президентской программы «Ведущие научные школы РФ» (грант № НШ-6068.2010.9) и интеграционных проектов СО РАН.

чевые слова и ключевые словосочетания. В процессе автоматической категоризации документов ключевые слова являются основной характеристикой, отражающей содержание документа.

Проблема, возникающая в процессе индексирования документов, состоит в выборе структуры списка ключевых слов: должен ли он состоять исключительно из одиночных слов или может включать в себя и составные выражения? Конечно, составные ключевые слова более адекватно описывают предметную область, но при их использовании значительно усложняется морфологический анализ текста. Более того, в некоторых работах, например, в статье [2], содержащей подробный обзор современных методов классификации документов с использованием ключевых слов, утверждается, что использование одиночных ключевых слов является «наиболее приемлемым». Такой подход при наличии качественных средств морфологического анализа представляется недостаточно обоснованным, по крайней мере, для коллекций документов, относящихся к какой-либо определенной узкой тематике (данная оговорка сделана и в [2]), поскольку использование одиночных ключевых слов имеет серьезные теоретические недостатки [3]: возможность ложной координации, ложных синтагматических связей и др.

Цель данной работы заключается в демонстрации практических преимуществ кластеризации документов на основе ключевых словосочетаний по сравнению с кластеризацией на основе анализа только однословных ключевых термов, при этом для выделения ключевых словосочетаний используются общедоступные программные средства, не требующие особых вычислительных затрат.

Алгоритм извлечения ключевых термов

Важной задачей обработки текстовых документов, результат решения которой используется не только для их классификации (категоризации), но и для извлечения из них информации и знаний, является координатное индексирование, т. е. извлечение из текстов документов ключевых слов и словосочетаний.

Классический подход к решению данной проблемы заключается в использовании средства анализа на основе тезауруса обрабатываемой предметной области. Но метод выделения ключевых слов и словосочетаний, основанный на анализе тезауруса предметной области, имеет существенный недостаток: таким способом нельзя производить индексирование корпусов текстов произвольных тематик. Более того, если вести речь об обработке корпусов текстов достаточно узких тематик, то в таких случаях требуются очень подробные тезаурусы, которые имеются (по крайней мере, в широком доступе) далеко не для всех предметных областей. Подход же, основанный на извлечении ключевых выражений без априорных ограничений, носит гораздо более универсальный характер, хотя, естественно, несколько проигрывает в адекватности индексирования.

Ввиду того, что в русском языке имена существительные и прилагательные при склонении изменяют свою форму, разработка эффективного алгоритма автоматизации извлечения ключевых слов является нетривиальной задачей, так как необходимо учитывать и те случаи, когда слова, образующие термин (т. е. ключевое слово), находятся не только в именительном, но и в косвенных падежах.

Для решения этой задачи мы опирались на морфологический анализ текстов и выделения ключевых словосочетаний по морфологическим шаблонам с использованием программного продукта компании Яндекс (http://company.yandex.ru/technology/mystem/), который является бесплатным для некоммерческих целей. При фильтрации и разборе производился отсев стоп-слов. Ключевые словосочетания отбирались по морфологическим шаблонам с учетом словоформ языка.

Для определения ключевых словосочетаний использовались классические морфологические шаблоны, которые достаточно качественно определяют искомые ключевые выражения:

(Причастие) (Существительное)

(Прилагательное) (Существительное)

(Существительное) (Существительное в творительном падеже)

(Существительное) (Существительное в родительном падеже)

После завершения подсчета вхождений ключевых слов и словосочетаний в документе необходимо произвести отделение наиболее значимых слов, отражающих контекстное содержание корпуса. Количество вхождений слов в текст в большинстве случаев поддается закону распределения частот Ципфа: если все слова упорядочить по убыванию частоты их использования, то частота *n*-го слова в этом списке окажется примерно обратно пропорциональной его порядковому номеру (рангу). Для отделения одиночных ключевых слов использовался именно закон Ципфа.

Однако данный закон не работает для частоты распределения ключевых словосочетаний. Для ограничения числа составных ключевых фраз, наиболее точно описывающих содержание электронного документа, использовалась следующая закономерность, замеченная эмпирическим путем, которая проверялась на достаточно большом количестве корпусов текстов средней и большой величины:

$$KeyPhrase\,(i):\frac{\max(Frequency)}{Frequency\,(i)}<\frac{word\_num}{3},$$

где max(*Frequency*) – максимальная частота встречаемости 1-го (т. е. наиболее часто встречаемого) терма и всех его словоформ терма в корпусе текстов); *Frequency*(*i*) – частота встречаемости *i*-го, проверяемого, терма; *word_num* – желаемое (ориентировочно) количество отобранных термов.

Разумеется, данное условие (как и закон Ципфа) плохо работает на документах небольшого размера (типа аннотаций), поскольку в них частоты всех однословных и многословных ключевых терминов приблизительно равны и стремятся к единичному вхождению в рамках контекста документа.

Для демонстрации качества отбора ключевых слов и ключевых словосочетаний на основе морфологических шаблонов приведем результаты их отбора из текста романа Л. Н. Толстого «Война и мир», а также текста научной статьи философской направленности (З. О. Османов «К вопросу о различении эпистемологических категорий»), размер которой является совершенно обычным (средним) в пределах рамок, используемых для публикации научных материалов.

В табл. 1 и 2 приводятся результаты выделения ключевых терминов из документов различной тематики и величины. Рядом с каждым термом приведено количество его вхождений (в различных формах) в текст документа. Для составных ключевых термов: фраза в скобках – форма, в которой данное выражение встречалось последний раз («крайнее вхождение»), которое приводится для облегчения (в необходимых случаях) понимания семантики термов, фраза до скобок – первообразная форма фразы, т. е. форма, по которой производился подсчет вхождений.

Нетрудно видеть, что приведенные в таблицах термы вполне адекватно отражают тематику документов, при этом, если говорить о философской статье, процент стандартных «общенаучных» лексических конструкций (к каковым из числа полученных двухсловных термов можно отнести выражения «обладающее свойством», «следующим образом» и «изучаемых объектов») достаточно мал. Это позволяет сделать вывод о том, что данная методика выбора однословных и двухсловных ключевых термов показала достаточно хорошее качество работы.

Таблица 1

Выделенные термы из романа Л. Н. Толстого «Война и мир»

Однословные термы		**Двухсловные термы**	
князь –	2011	княжна марья –	93 (княжною Марьей)
человек –	1755	старый князь –	92 (старого князя)
княжна –	885	полковой командир –	76 (полкового командира)
граф –	734	старый граф –	53 (старого графа)
время –	714	русский армия –	50 (русская армия)
москва –	644	русский войска –	41 (русскими войсками)
француз –	595	молодой человек –	32 (молодого человека)
государь –	591	исторический лицо –	30 (исторические лица)
солдат –	581	выражение лицо –	30 (выражением лица)
наполеон –	575	французский армия –	28 (французской армией)
жизнь –	572	главный квартира –	27 (главная квартира)
слово –	566	французский войска –	26 (французские войска)
рост –	544	старый графиня –	23 (старой графини)
офицер –	543	князь андрей –	23 (князем Андреем)
кутузов –	533	военный министр –	23 (военного министра)
армия –	463	французский офицер –	21 (французских офицеров)
лошадь –	450	великий князь –	20 (великого князя)
графиня –	441	расположение дух –	19 (расположении духа)
войска –	435	лицо наташа –	19 (Лицо Наташи)

Таблица 2

Выделенные термы из статьи З. О. Османова
«К вопросу о различении эпистемологических категорий»

Однословные термы		**Двухсловные термы**	
знание –	131	эпистемологический категория –	5 (эпистемологических категорий)
суждение –	85	познавательный процесс –	5 (познавательный процесс)
истина –	71	обладающее свойством –	5 (обладающее свойством)
вера –	50	достоверный знание –	5 (достоверное знание)
мнение –	38	эпистемологический статус –	4 (эпистемологический статус)
сведение –	37	тематический словарь –	4 (тематический словарь)
заблуждение –	37	следующим образом –	4 (следующим образом)
истинность –	36	некий суждение –	4 (некоему суждению)
отношение –	32	мнение вера –	4 (мнения веры)
объект –	29	эмоциональный оценивание –	3 (эмоциональное оценивание)
субъект –	28	познавательный деятельность –	3 (познавательной деятельности)
состояние –	26	ложный вера –	3 (ложная вера)
слово –	26	логический круг –	3 (логических кругов)
		истинный суждение –	3 (истинных суждений)
		истинный вера –	3 (истинная вера)
		изучаемых объектов –	3 (изучаемых объектов)
		аффективный точка –	3 (аффективной точки)

Алгоритм кластеризации текстов

Кластеризация наборов электронных документов выполнялась с использованием так называемого жадного алгоритма [4], который признан методом, дающим достаточно хорошие результаты при кластеризации корпуса научных статей близкой тематики (см. например, [5]), хотя и обладающим сравнительно большой вычислительной сложностью.

Для лучшего понимания результатов кластеризации и объяснения большой вычислительной сложности работы коротко опишем метод его работы. Процесс можно описать шагами, циклически повторяемыми до тех пор, пока не будет «свободных» документов, которые не включены ни в один из результирующих кластеров.

0. Строится матрица схожести парных сочетаний каждого документа с каждым, т. е. матрица $N \times N$, где N равняется количеству документов в кластеризуемой выборке. На пересечении задаются меры сходства документов в шкале [0; 1], причем 0 соответствует полному различию документов, а 1 – полному их сходству. Разумеется, матрица заполняется только до главной диагонали.

1. Ищется строка матрицы, сумма компонент которой будет максимальной. Эта строка содержит в себе все коэффициенты подобия i-го документа ко всем остальным документам. Этот документ объявляется центром 1-го кластера. Затем в кластер добавляются все документы, коэффициенты подобия к которым больше либо равно некоторому наперед заданному пороговому значению, являющемуся параметром данного метода и позволяющему управлять процессом кластеризации.

2. Исключаются все документы, попавшие в кластер, т. е. из матрицы вычеркиваются все строки и столбцы, соответствующие документам, добавленным в кластер. Далее пункты 1 и 2 повторяются до тех пор, пока не останется документов, не включенных в какой-либо кластер.

Очевидно, что таких операций будет не более чем N (на самом деле значительно меньше). При подобном подходе можно пройти весь массив документов, сформировав некоторое количество кластеров, которое будет варьироваться в зависимости от информационной потребности (это реализуется посредством изменения порогового значения).

Вычислительные эксперименты

Были проведены две серии экспериментов: кластеризация достаточно большого множества документов правовой направленности (около 1 300 документов) и набор научных документов математической направленности, содержащих классификационные признаки классификатора MSC2000.

Целью первой серии экспериментов была проверка работы механизма кластеризации в целом (от этапа автоматического выделения ключевых термов до итогового получения разбиения на кластеры множества документов) на примере достаточно большого массива документов. Во второй серии экспериментов априорное знание классификационных признаков позволило произвести вычисление мер качества и сравнить работу методики при работе алгоритма с использованием однословных и двухсловных ключевых выражений, варьируя при этом параметрический коэффициент жадного алгоритма кластеризации.

Так как вычисление ошибки кластеризации в классическом виде в первом эксперименте не выполнялось (поскольку не производилось экспертного разбиения выборки документов), то для демонстрации корректности работы методики, основанной на анализе ключевых однословных и составных термов, были случайным образом выбраны 3 кластера и выписаны названия случайно выбранных документов, включенных в них. В данном эксперименте на меру сходства между двумя документами оказывали влияние как однословные, так и составные ключевые термы.

Как видно из приведенных названий статей, документы каждого из кластеров относятся к определенной, явно выраженной тематике.

Кластер 1, общая тематика – налогообложение и уклонение от уплаты налогов.
1. Бухгалтер в России.
2. Функции государства – налогообложение и взимание налогов.
3. Налоговые преступления.
4. Уклонение от уплаты налогов с организаций.
5. Уклонение физического лица от уплаты налога или страхового взноса.

Кластер 2, общая тематика – управление и государственная служба.
1. Понятие, принципы и порядок прохождения государственной службы.
2. Управление: основные понятия, система управления, ее признаки [...].
3. Основные принципы создания, [...] организации арбитражных управляющих.
4. Особенности государственной службы субъекта Российской Федерации.
5. Органы внутренних дел Российской Федерации, правовые основы [...].

Кластер 3, общая тематика – имущественные права.
1. Институциональные аспекты землепользования.
2. О возможности защиты права собственности на недвижимость путем виндикации.
3. Природа виндикационного притязания и элементы виндикационного иска.
4. Правовое положение лица, владеющего имуществом [...].
5. Критика понятия «объект правоотношения».

Итоговые кластеры не являются чем-то отдельно стоящим: через некоторые ключевые выражения они могут быть связаны с другими группами, а также с другими документами, как в своей, так и в чужой группе. Диаграмма на рис. 1 показывает взаимосвязь через некоторое ключевое выражение кластеров, которые, несмотря на свою «непохожесть», имеют что-то общее.

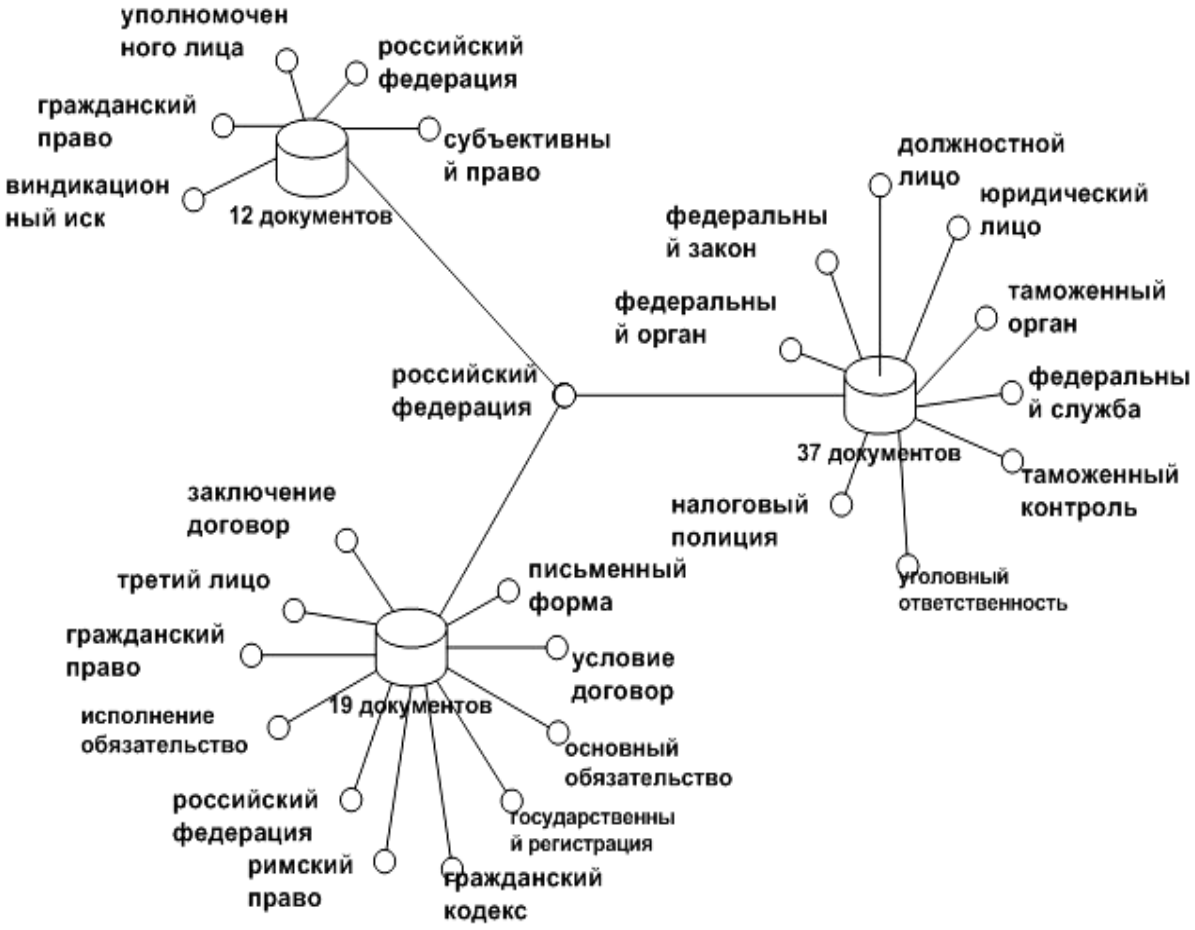

Рис. 1. Взаимосвязь кластеров через ключевые выражения

Рис. 2 показывает, что удельный объем кластеров, содержащих достаточно большое количество элементов, довольно велик, т. е. разделение документов по тематикам выполнялось на хорошем уровне, с учетом того, что все документы принадлежали одной обширной области знаний.

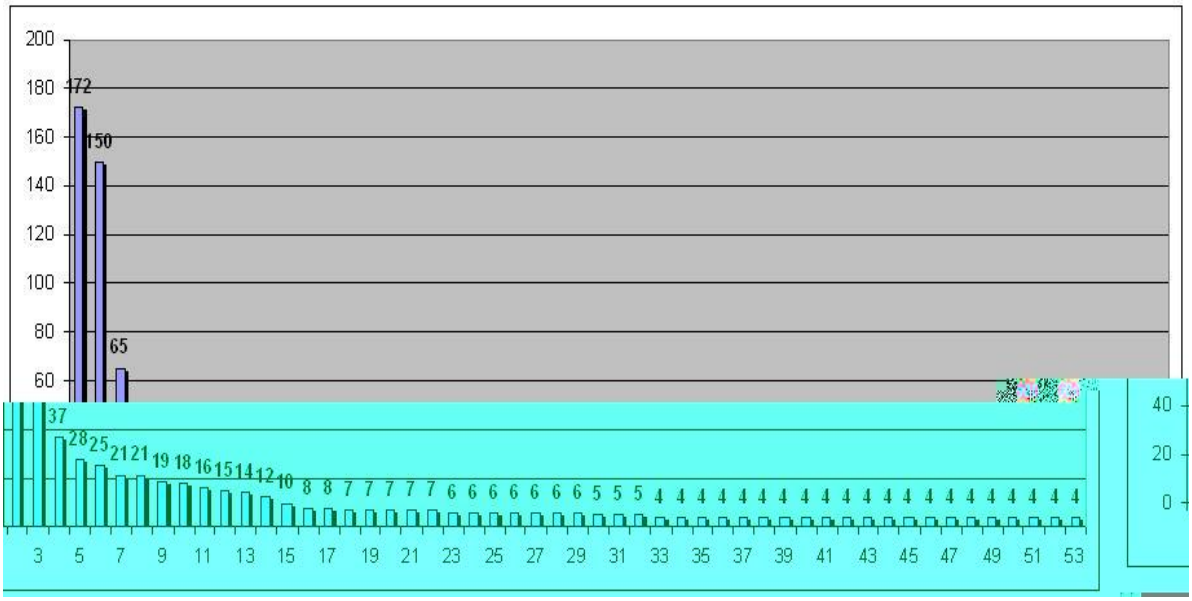

Рис. 2. Количество документов в кластерах

Документов, не включенных ни в какой кластер (иными словами, количество кластеров, состоящих из 1-го документа), оказалось 202, т. е. около 15 % от общего количества документов, которые участвовали в данном эксперименте.

Исходные данные второй серии экспериментов позволили произвести вычисление мер, с помощью которых можно оценить качество работы методики.

Тестирование алгоритмов проводилось на полных текстах статей журнала «Вычислительные технологии». Электронная версия данного журнала содержит библиографические описания, а также полные тексты статей, вышедших в период с 1996 по 2010 гг. [1]. Некоторым статьям на сайте журнала, помимо стандартных атрибутов (название, автор, год издания и т. п.), приписаны соответствующие коды классификатора из «Классификации математических сущностей» (MSC2000). Совпадение данных кодов для группы документов является объективным критерием совпадения тематики данных документов.

Работа по оцениванию мер качества кластеризации была разбита на следующие этапы.

1. Кластеризация подготовленных текстов статей на основе жадного алгоритма с различными значениями входного параметра (порогового значения).

2. Получение результатов кластеризации с использованием однословных ключевых терминов и результатов, основанных на смешанном критерии, т. е. с использованием как простых, так и составных ключевых выражений.

3. Вычисление внешних мер для полученных результатов. Нахождение оптимального метода кластеризации (и задание оптимального параметра порогового значения для жадного алгоритма), который даст результат, наиболее близкий к результату разбиения на основе кодов классификатора MSC2000.

Выделяют следующие два вида мер качества кластеризации документов: внешние и внутренние (см., например, [6; 7]).

[1] См.: http://www.ict.nsc.ru/jct.

Внешние меры основаны на сравнении автоматического разбиения с полученным от экспертов эталонным разбиением этих же данных. Идея, положенная в основу этих мер, заключается в том, чтобы для каждой пары документов автоматически сопоставить два решения о сходстве этих тематик.

Примерами внешних мер являются традиционные для оценки систем поиска такие характеристики, как полнота (*Recall*), точность (*Precision*), ошибка классификации (*Error*), F_1-мера и др. Эти характеристики подсчитываются по формулам:

$$Recall = \frac{a}{a+b};\ Precision = \frac{a}{a+c};\ Error = \frac{b+c}{a+b+c+d};\ F_1 = \frac{2 \cdot Precision \cdot Re.call}{Precision + Re.call},$$

где коэффициенты *a, b, c, d* определяются в соответствии с табл. 3.

Таблица 3

Коэффициенты для подсчета внешних мер сходства

Для каждой пары документов d_j и d_i	d_j и d_i принадлежат одному кластеру в «эталонном» разбиении	d_j и d_i принадлежат разным кластерам в «эталонном» разбиении
d_j и d_i принадлежат одному кластеру в автоматическом разбиении	*a*	*c*
d_j и d_i принадлежат разным кластерам в автоматическом разбиении	*b*	*d*

Оценка и анализ внутренних мер (основанных на оценке оценка свойств отделимости и компактности полученного разбиения документов) в данной работе не производились.

В табл. 4 приводятся вычисленные меры при выполнении анализа только на основе одиночных ключевых слов, в табл. 5 приводятся результаты процесса кластеризации на основе анализа как одиночных ключевых слов, так и двухсловных ключевых выражений (значение параметра работы жадного алгоритма, показывающего, при какой мере сходства можно говорить, что один документ достаточно похож на другой, выбиралось в диапазоне 0,4–0,6).

Таблица 4

Меры внешнего сходства,
полученные на основе только одиночных ключевых слов

Анализ на основе кодов классификатора 2-го уровня (вида 76Mxx)

Мера	Значение параметра (величина схожести)		
	0,4	0,5	0,6
Recall	0,2263	0,2394	0,3223
Precision	0,2194	0,1026	0,0404
Error	0,2251	0,1799	0,1536
F_1	0,2228	0,1437	0,0718

Таблица 5

Меры внешнего сходства, полученные на основе
как одиночных ключевых слов, так и двухсловных выражений

Анализ на основе кодов классификатора 2-го уровня (вида 76Mxx)			
Мера	Значение параметра (величина схожести)		
	0,4	0,5	0,6
Recall	0,2771	0,3796	0,4624
Precision	0,1105	0,0387	0,0081
Error	0,1732	0,1507	0,1473
F_1	0,1580	0,0702	0,0159

Как видно из представленных результатов, кластеризация документов на основе ключевых словосочетаний показывает меньшее значение ошибки по сравнению с кластеризацией на основе анализа только однословных ключевых термов. Кроме того, она лучше справляется с поставленной задачей, если исходная выборка документов принадлежит к одной, достаточно узкой, тематике, чего существенно сложнее добиться при обработке такой выборки алгоритмом, опирающимся только на анализ однословных ключевых термов.

Сравнительный анализ табл. 4 и 5 показывает стабильно более высокую точность алгоритма кластеризации, использующего смешанный подход для всех значений порогового параметра жадного алгоритма. Также данный подход дает более высокие значения коэффициента полноты. Наиболее оптимальный результат с коэффициентом ошибки *Error* = 0,1473 достигается с использованием порогового значения, равного 0,6.

Заключение

Проведенное сравнение результатов кластеризации документов, принадлежащих корпусам близких по тематике текстов, показало целесообразность применения общедоступных средств морфологического анализа текстов для извлечения составных ключевых термов, поскольку использование последних для подсчета меры сходства между документами дает заметно лучшие результаты по сравнению с получаемыми при использовании лишь одиночных ключевых слов, позволяя во многих случаях избежать ошибок ложной координации, при этом рост вычислительных затрат на обработку одного текста незначителен. Разумеется, полученные результаты несколько уступают тем, которые возможны при использовании для выделения ключевых слов и словосочетаний тезауруса предметной области, однако, поскольку речь идет об обработке корпусов текстов, близких по тематике, в таких случаях потребовались бы весьма подробные тезаурусы, которые имеются (по крайней мере, в широком доступе) далеко не для всех предметных областей. Рассматриваемый же в статье подход носит гораздо более универсальный характер.

Список литературы

1. *Федотов А. М., Барахнин В. Б.* К вопросу о поиске документов «по аналогии» // Вестн. Новосиб. гос. ун-та. Серия: Информационные технологии. 2009. Т. 7, вып. 4. С. 3–14.

2. *Пескова О. В.* Автоматическое формирование рубрикатора полнотекстовых документов // Тр. X Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» (RCDL’2008). Дубна, 7–11 октября 2008 г. С. 139–148.

3. *Михайлов А. И., Черный А. И., Гиляревский Р. С.* Основы информатики. М.: Наука, 1968.

4. *Кормен Т., Лейзерсон Ч., Ривест Р. М.* Алгоритмы: построение и анализ. М.: МЦНМО, 2001.

5. *Барахнин В. Б., Нехаева В. А., Федотов А. М.* О задании меры сходства для кластеризации текстовых документов // Вестн. Новосиб. гос. ун-та. Серия: Информационные технологии. 2008. Т. 6, вып. 1. С. 3–9.

6. *Bezdek J. C., Pal N. R.* Some New Indexes of Cluster Validity // IEEE Transactions On Systems, Man And Cybernetics. 1998. Vol. 28, No. 3. P. 301–315.

7. *Halkidi M., Batistakis V., Vazirgiannis M.* On Clustering Validation // Journal of Intelligent Information Systems. 2001. Vol. 17 (2/3). P. 107–145.

Материал поступил в редколлегию 11.05.2010

V. B. Barakhnin, D. A. Tkachev

CLUSTERING OF TEXT DOCUMENTS BASED ON COMPOSITE KEY TERMS

The classical approach to the coordinate indexing texts with a view to their subsequent clustering is to use analysis tools based on the thesaurus treated he subject area. But if we talk about the processing of texts rather narrow topics, in such cases requires a very detailed thesauri, which are (at least, widely available), not for all subject fields. The approach is based on the extraction of key phrases without a priori constraints is much more universal. However, this approach has the problem of selection of key terms. The purpose of this article is to demonstrate the practical advantages of clustering documents based on key phrases compared to the very popular clustering based on the analysis of only one-word key terms. At the same time to highlight the key phrases used publicly available software tools that do not require special computing costs.

Keywords: clustering text documents, coordinate indexing, composite key terms.