УДК 004.912

ПРЕДСТАВЛЕНИЕ ДОКУМЕНТОВ В ЗАДАЧЕ КЛАСТЕРИЗАЦИИ АННОТАЦИЙ НАУЧНЫХ ТЕКСТОВ

С.В. Попова[a, b], В.В. Данилова[c]

[a] Санкт-Петербургский государственный университет, Санкт-Петербург, Россия, svp@list.ru
[b] Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики, Санкт-Петербург, Россия, svp@list.ru
[c] Автономный университет Барселоны, Барселона, Испания, maolve@gmail.com

Рассматривается проблема кластеризации узкотематических текстов короткой длины, таких как аннотации к научным публикациям. Цель решения данной задачи – группировка результатов запросов в поисковых системах по научным публикациям. Использованы наблюдения, полученные при решении задачи извлечения ключевых фраз из документов. Был применен расширенный список стоп-слов, построенный автоматически для решения задачи извлечения ключевых фраз и позволивший значительно улучшить качество информации, получаемой из научных публикаций. Приводится описание процедуры построения данного списка стоп-слов. Основной задачей является исследование возможности повысить качество и (или) скорость кластеризации аннотаций с помощью вышеупомянутого списка стоп-слов, а также информации о частях речи лексем. В последнем случае для представления документов применяется словарь, содержащий не все слова коллекции, а только существительные и прилагательные, или словарь, состоящий из последовательностей существительных и прилагательных.

Использованы два базовых алгоритма кластеризации: *k*-means и иерархическая кластеризация (метод межгруппового среднего). Показано, что использование расширенного списка стоп-слов и представление документов на основе существительных и прилагательных из словаря коллекции позволяют улучшить качество и скорость работы алгоритма *k*-means. Для метода межгруппового среднего в аналогичном случае может наблюдаться ухудшение качества кластеризации. Показано, что использование для представления документов последовательностей из существительных и прилагательных снижает качество кластеризации для обоих алгоритмов и оправдано только в тех случаях, когда требуется значительное снижение размерности пространства признаков.

Ключевые слова: кластеризация документов; представление документов; использование ключевых фраз, существительных и прилагательных; построение расширенного списка стоп-слов, представления результатов поиска.

DOCUMENT REPRESENTATION FOR CLUSTERING OF SCIENTIFIC ABSTRACTS

S. Popova[d, e], V. Danilova[f]

[d] Saint Petersburg State University, Saint Petersburg, Russia, svp@list.ru
[e] Saint Petersburg National Research University of Information Technologies, Mechanics and Optics, Saint Petersburg, Russia, svp@list.ru
[f] Autonomous University of Barcelona, Barcelona, Spain, maolve@gmail.com

The key issue of the present paper is clustering of narrow-domain short texts, such as scientific abstracts. The work is based on the observations made when improving the performance of key phrase extraction algorithm. An extended stop-words list was used that was built automatically for the purposes of key phrase extraction and gave the possibility for a considerable quality enhancement of the phrases extracted from scientific publications. A description of the stop- words list creation procedure is given. The main objective is to investigate the possibilities to increase the performance and/or speed of clustering by the above-mentioned list of stop-words as well as information about lexeme parts of speech. In the latter case a vocabulary is applied for the document representation, which contains not all the words that occurred in the collection, but only nouns and adjectives or their sequences encountered in the documents. Two base clustering algorithms are applied: k-means and hierarchical clustering (average agglomerative method). The results show that the use of an extended stop-words list and adjective-noun document representation makes it possible to improve the performance and speed of k-means clustering. In a similar case for average agglomerative method a decline in performance quality may be observed. It is shown that the use of adjective-noun sequences for document representation lowers the clustering quality for both algorithms and can be justified only when a considerable reduction of feature space dimensionality is necessary.

Keywords: document clustering, document representation, key phrases application, use of nouns and adjectives, extended list of stop-words creation, results retrieval representation.

Введение и обзор состояния дел в области

В работе рассматривается проблема кластеризации текстовых данных в режиме реального времени. Исследование проводится в рамках решения задачи представления результатов поиска пользователю в виде кластеров [1–3], в частности, в системах, осуществляющих поиск по базам научных публикаций [4, 5]. Целью решения является представление в структурированном виде коллекции документов с автоматическим разделением ее на группы тематически близких текстов.

Исходные данные – аннотации к научным публикациям, доступные в научных электронных библиотеках. Большинство таких библиотек предоставляет свободный доступ к аннотациям, получение полных текстов статей, как правило, требует специальной подписки или оплаты доступа. Тем не менее, аннотации представляют собой лаконичную выдержку основного содержания работы и могут быть достаточными для кластеризации научных текстов и получения адекватного результата [6].

Работа с аннотациями связана с рядом сложностей: во-первых, это короткие тексты, обработка которых обычно выделяется в отдельную проблему short-text clustering/classification (кластеризация/классификация текстов короткой длины) [7–10], во-вторых, обрабатываемые тексты могут иметь одну общую тему, в связи с чем требуется решение проблемы кластеризации узкотематических текстов короткой длины [11–14] (narrow-domain short text clustering). В более ранних исследованиях, посвященных решению указанных проблем, ученые сосредоточены на разработке алгоритмов безотносительно их вычислительной сложности, основной акцент делается на повышение качества кластеризации.

Одной из основных проблем при кластеризации текстов короткой длины является сильная разреженность данных [11]. Так как мы имеем дело с короткими документами, а суммарный объем текста в коллекции может быть небольшим, затруднен сбор статистических характеристик, необходимых для обработки. Если документы взяты из одного источника (например, выбраны как релевантные одному запросу), велика вероятность того, что все они будут тематически близкими друг к другу. В этом случае, помимо небольшого размера, тексты могут иметь значительные перекрытия по общим словам, что еще больше усложняет задачу кластеризации [11]. Для тестирования и анализа качества работы алгоритмов кластеризации узкотематических текстов короткой длины был сформирован ряд коллекций, которые активно используются в области (CICling, EasyAbstracts, SEPLN-CICling) [6–8, 11–14]. Эксперименты в данной работе также базируются на данных коллекциях.

Выбор алгоритмов и мотивация

Поставлена задача исследовать возможность повысить качество и (или) скорость кластеризации коротких текстов узкой тематики, используя дополнительную информацию о частях речи слов и расширенный список стоп-слов. При постановке задачи для экспериментов мы опирались на исследования в области кластеризации/классификации текстов и на исследования в области извлечения ключевых понятий (ключевых фраз). Наблюдения показывают, что в обеих областях используются очень похожее подходы и инструменты [1–3, 15–17], что позволяет сделать вывод о возможном использовании достижений в одной области в рамках работ другого направления.

В области аннотирования документов ключевыми словами/фразами было показано, что существительные и прилагательные являются основными частями речи, необходимыми для извлечения ключевых понятий [18–21]. Данное наблюдение позволило сформировать гипотезу о том, что использование только существительных и прилагательных может положительно сказаться на качестве кластеризации и позволит понизить размерность пространства признаков. Поставленный эксперимент также связан с нашими исследованиями в области извлечения ключевых фраз [5, 22–24]. Показано, что последовательности, состоящие из существительных и прилагательных, часто содержат в себе основные понятия из документов и позволяют добиться неплохого качества аннотирования.

Второе наблюдение связано с исследованием, в рамках которого нам удалось показать эффективный способ автоматического построения списка стоп-слов для извлечения ключевых понятий из аннотаций к научным публикациям [24]. В настоящей работе мы исследуем возможность повысить качество кластеризации при использовании списка стоп-слов, полученного в работе [24].

Для кластеризации были выбраны алгоритмы *k*-means (*k*-средних) и hierarchical agglomerative clustering (иерархическая кластеризация), метод межгрупповых связей. Была использована известная в области библиотека Weka, предлагающая реализацию указанных выше алгоритмов. Алгоритмы являются классическими в задаче информационного поиска и хорошо описаны в литературе [25]. Авторы предполагают, что читатель знаком с данными методами или может самостоятельно ознакомиться с ними, например, в [25]. Отметим, что для иерархической кластеризации применяется алгоритм, оценивающий расстояние между кластерами как среднее расстояние между каждой парой объектов в разных кластерах (в [25] этот метод обозначен как «centroid», в Weka как «AVERAGE»). Данный метод не учитывает расстояния между объектами внутри объединяемых кластеров (в отличие от Group-average agglomerative clustering в [25]). В дальнейшем для выбранного алгоритма иерархический кластеризации мы будем использовать обозначение HAC.

При выборе алгоритмов кластеризации мы исходили из двух положений: вычислительная простота (*k*-средних) и потенциально возможный высокий результат кластеризации (HAC). В области автоматической группировки коротких текстов узкой тематики представлены результаты для таких алгоритмов иерархической кластеризации, как методы ближнего или дальнего соседа [11]. Нашей задачей было расширить данные по результатам тестирования алгоритмов иерархической кластеризации для указанной задачи.

Мотивацией для проведения исследования является задача кластеризации результатов запроса в поисковых системах [1–3], акцент был сделан на изучение данной проблемы для систем академического поиска (т.е. поиска по научным публикациям) [4, 5]. Введено предположение о неважности затрат на обработку данных, если данная обработка может быть проведена оффлайн, например, извлечение информации о части речи слов или извлечение из документов последовательностей из существительных и прила-

гательных. Основными требованиями к алгоритму кластеризации были высокая скорость работы (в пределах времени, допустимого для ожидания пользователя) и как можно более высокое качество работы.

Описание тестовых коллекций

Выбраны три коллекции (sites.google.com/site/merrecalde/resources), часто используемые для тестирования алгоритмов кластеризации документов короткой длины или узкотематических коротких текстов [7, 8, 11–14]. Все три коллекции представляют собой наборы аннотаций к научным публикациям, распределенные по четырем кластерам экспертом. Для каждой коллекции известен «золотой стандарт» – наилучший вариант группировки. Размер каждой из указанных коллекций – 48 документов. Коллекция CICling 2002 признана исследователями как одна из наиболее сложных коллекций для кластеризации [7, 8, 11–14].

Оценка качества кластеризации

В работе была использована классическая оценка качества кластеризации, основанная на комбинировании информации о точности (Precision) и полноте (Recall) полученных кластеров [25–27]:

$$FM = \sum_i \frac{G_i}{|D|} \max_j F_{ij}, \text{ где } F_{ij} = \frac{2 \cdot P_{ij} \cdot R_{ij}}{P_{ij} + R_{ij}}, \quad (1)$$

$$P_{ij} = \frac{|G_i \cap C_j|}{G_i}, \; R_{ij} = \frac{|G_i \cap C_j|}{C_j},$$

$G = \{G_i\}_{i=\overline{1,m}}$ – кластеры, полученные в результате кластеризации, $C = \{C_j\}_{j=\overline{1,n}}$ – классы, построенные экспертом вручную, D – множество документов в коллекции.

Описание эксперимента

Исследование состояло из двух основных блоков: работа с алгоритмом *k*-средних и с алгоритмом HAC. Для каждой тестовой коллекции требовалась кластеризация по четырем кластерам (аналогично работам [7, 8, 11–14], для алгоритмов, требующих данные о числе кластеров, указывалось число классов в золотом стандарте). Для представления документов использовалась векторная модель: каждый документ был представлен вектором в пространстве признаков (словаря коллекции). Вес каждого признака (слова) w для документа d определялся с помощью *tf – idf* [25] (кроме случая представления документов с помощью последовательностей слов):

$$tf - idf(w,d) = tf^d(w) \cdot \log \frac{|D|}{df(w)},$$

где D – множество документов в коллекции; $tf^d(w)$ – частота слова w в документе d; $df(w)$ – число документов, в которых встретилось слово w.

В обоих алгоритмах для определения расстояния между объектами использовался косинус угла между векторами, представляющими документ.

Для каждого из алгоритмов проверялись две гипотезы:

- **гипотеза 1**: качество кластеризации можно повысить, используя для представления документов только существительные и прилагательные, а также расширенный список стоп-слов;
- **гипотеза 2**: повысить качество кластеризации можно, используя для представления документов последовательности из существительных и прилагательных.

Во всех экспериментах для представления документов и пространства признаков были использованы стемы слов, полученные при помощи стеммера Портера (The Porter Stemming Algorithm[1]). Во всех экспериментах применялся стандартный список стоп-слов, кроме отдельно отмеченных экспериментов с расширенным списком.

Проверка гипотезы 1. В эксперименте для построения пространства признаков были использованы все существительные и прилагательные из словаря коллекции, встретившиеся хотя бы раз, по крайней мере, в одном тексте. Каждый документ был представлен вектором в полученном пространстве признаков. Также был рассмотрен случай, когда из пространства признаков были удалены все слова, попавшие в расширенный список стоп-слов. Способ построения данного списка приводится далее.

Построение расширенного списка стоп-слов. Нами проводились исследования в области излечения ключевых фраз для аннотаций к научным публикациям [5, 22–24]. Для построения фраз были использованы только существительные и прилагательные, представленные в тексте. В работе [24] был предложен алгоритм автоматического построения словаря стоп-слов, включающего общеупотребитель-

[1] http://tartarus.org/martin/PorterStemmer

ные слова, которые не отражают тематической направленности публикации и, как следствие, не входят в состав ключевых фраз (например: предыдущий, новый, предшествующий, статья и т.д.). Была использована одна из основных в области извлечения ключевых фраз коллекция INSPEC [18, 19, 21–24, 28, 29]. Подробное описание коллекции приводится в [18]. INSPEC содержит в себе несколько подколлекций, в частности, «trial», которая используется для обучения алгоритмов, и «test» – для оценки качества работы алгоритмов аннотирования и сравнения алгоритмов. Каждая из коллекций имеет «золотой стандарт» – результат аннотирования ключевыми фразами каждого документа. Для построения списка стоп-слов использовалась подколлекция «trial»: последовательно отбиралось каждое существительное или прилагательное из словаря коллекции и добавлялось в стандартный список стоп-слов. Запускался алгоритм извлечения ключевых фраз (см. полное описание в [24]), и, если добавление этого слова приводило к повышению качества полученной информации более чем на заданную величину (0,0002, см. в [24]), то добавленное в список слово помечалось как «плохое». Примеры таких «плохих» слов приводятся в табл. 1.

№	Слово (оригинал извлеченного слова)	Русскоязычный перевод слова
1	Actual	Существующий, текущий
2	Possible	Вероятный, возможный
3	Excellent	Отличный, превосходный
4	Results	Результаты
5	Number	Число, количество, номер
6	Pure	Строгий, чистый, простой
...	...	...

Таблица 1. Примеры стоп-слов, попавших в расширенный список стоп-слов

После того как описанная выше процедура была выполнена для всех существительных и прилагательных из словаря коллекции, все слова, получившие отметку «плохое», были собраны вместе и добавлены к стандартному списку стоп-слов. Полученный таким образом расширенный список был использован для извлечения ключевых фраз из документов коллекции «test» и позволил значительно повысить качество полученной информации по сравнению со случаем использования стандартного списка. В настоящей работе мы применяем вышеописанный расширенный список стоп-слов.

Проверка гипотезы 2. В эксперименте каждый документ был описан с помощью последовательностей существительных и прилагательных, следующих в документе друг за другом (не более пяти слов подряд), разделенных словами других частей речи, пунктуацией, стоп-словами. Также в описание документов были добавлены пересечения полученных последовательностей. Например, если в тексте есть последовательности «русский язык» и «английский язык», то последовательность «язык» также будет добавлена в описание документа.

Для кластеризации было использовано пространство признаков, состоящее из всех последовательностей, встретившихся хотя бы раз, по крайней мере, в одном описании документа. Каждый документ был представлен бинарным вектором в полученном пространстве признаков, где вес признака оценивался как 1, если последовательность, соответствующая признаку, присутствует в описании документа, и 0 в противном случае.

Результаты экспериментов и обсуждение

Результаты экспериментов представлены в табл. 2, 3. В табл. 2 приводятся результаты для алгоритма *k*-средних, в табл. 3 – итоги вычислений для алгоритма HAC. В таблицах приняты следующие обозначения. В строке «Базовый эксперимент» представлены результаты оценки качества алгоритма для каждой из трех коллекций (CICling, SEPLN-CICling, EasyAbstracts). Для построения пространства признаков были использованы все слова из словаря коллекции. Так как в алгоритме *k*-means центроиды выбираются произвольным образом, то качество работы алгоритма меняется от эксперимента к эксперименту. Для *k*-means приводятся результаты, полученные на основе 1000 запусков алгоритма: средний результат (avg), худший результат (min), лучший результат (max). В строке «Эксперимент: только существительные и прилагательные» приводятся результаты экспериментов для случая, когда построение пространства признаков основано только на существительных и прилагательных из словаря коллекции. В строке «Эксперимент: только существительные и прилагательные, использован расширенный список стоп-слов» представлены результаты эксперимента, аналогичного последнему, но пространство признаков сужено за счет расширенного списка стоп-слов. В строке «Эксперимент с использованием словосочетаний» приведены результаты экспериментов, в рамках которых для построения пространства признаков и представления документов использовались только последовательности из существительных и прилагательных; в строке «Эксперимент с использованием словосочетаний и расширенного списка стоп-слов» представле-

ны итоги экспериментов с использованием расширенного списка стоп-слов. Оценка качества кластеров, полученных в результате работы каждого из алгоритмов, проводилась с помощью (1).

Название эксперимента	Название коллекции								
	CICling			SEPLN-CICling			EasyAbstracts		
	avg	min	max	avg	min	max	avg	min	max
Базовый эксперимент	0,48	**0,34**	0,68	0,56	0,34	0,80	0,57	0,35	0,83
Эксперимент: только существительные и прилагательные	0,48	0,33	0,68	0,57	**0,36**	**0,83**	0,57	**0,37**	0,83
Эксперимент: только существительные и прилагательные, использован расширенный список стоп-слов	**0,49**	**0,34**	**0,73**	**0,58**	0,35	**0,83**	**0,58**	**0,37**	**0,86**
Эксперимент с использованием словосочетаний	0,45	0,33	0,66	0,50	0,34	0,71	0,49	0,34	0,77
Эксперимент с использованием словосочетаний и расширенного списка стоп-слов	0,46	0,32	0,67	0,53	0,36	0,77	0,50	0,33	0,66

Таблица 2. Результаты оценки качества работы алгоритма *k*-means

Название эксперимента	Название коллекции		
	CICling	SEPLN-CICling	EasyAbstracts
Базовый эксперимент	0,56	**0,84**	**0,92**
Эксперимент: только существительные и прилагательные	**0,58**	**0,84**	**0,92**
Эксперимент: только существительные и прилагательные, использован расширенный список стоп-слов	0,54	0,82	**0,92**
Эксперимент с использованием словосочетаний	0,44	0,43	0,51
Эксперимент с использованием словосочетаний и расширенного списка стоп-слов	0,52	0,63	0,54

Таблица 3. Результаты оценки качества работы алгоритма HAC

Результаты, указанные в таблицах, показывают, что использование для представления документов последовательностей из существительных и прилагательных приводит к снижению качества кластеризации по сравнению со случаем, когда для представления документов и построения пространства признаков использованы все слова. Основная причина – краткость обрабатываемых документов и, как следствие, небольшое число последовательностей, представляющих текст, а также очень низкое пересечение по общим последовательностям между документами. Использование расширенного списка стоп-слов при представлении документов с помощью последовательностей из существительных и прилагательных в большинстве экспериментов позволяет повысить качество кластеризации по сравнению со случаем использования указанных последовательностей и стандартного списка стоп-слов. Наблюдение позволяет сделать предположение о целесообразности использования расширенного списка стоп-слов для описания документов с помощью последовательностей из существительных и прилагательных как в задаче извлечения ключевых фраз, так и в задаче кластеризации.

Результаты работы алгоритма HAC улучшают результаты *k*-means. HAC обеспечивает высокие результаты по сравнению с другими алгоритмами, используемыми в области [7, 8, 11–14]. Для сравнения приведем результаты работы других алгоритмов, опубликованные в [14]. Нужно учитывать, что HAC дает один стабильный результат кластеризации, который не изменяется от запуска к запуску и является одновременно средним, лучшим и худшим значением. При сравнении рекомендуем читателю ознакомиться с входными параметрами, которые были использованы для алгоритмов в [14]. Так, существует некоторое различие в результатах, представленных для *k*-means в настоящем исследовании и в [14]. В первую очередь влияет на это различное число запусков алгоритма *k*-means: 1000 в данной работе и 50 в [14], а также рандомный выбор центроидов во время каждого из запусков. В табл. 4 представлены результаты из [14],

полученные по результатам 50 запусков каждого из алгоритмов: CLUDIPSO [7, 30], Ant-Tree [31], Major-Clust [32], DBSCAN, AntSA-CLU [14].

Algorithms	EasyAbstracts			SEPLN-CICling			CICling 2002		
	avg	min	max	avg	min	max	avg	min	max
Major-Clust	0,69	0,44	0,98	0,59	0,4	0,77	0,43	0,37	0,58
DBSCAN	0,66	0,62	0,72	0,63	0,4	0,77	0,47	0,42	0,56
Ant-Tree	0,6	0,5	0,67	0,49	0,41	0,64	0,41	0,38	0,48
CLUDIPSO	0,92	0,85	0,98	0,72	0,58	0,85	0,6	0,47	0,73
AntSA-CLU	0,96	0,92	0,98	0,75	0,63	0,85	0,61	0,47	0,75

Таблица 4. Результаты работы алгоритмов, представленные в [14]

Несмотря на хорошие результаты кластеризации HAC (табл. 3), нужно отметить бóльшую вычислительную сложность HAC по сравнению с *k*-means, что допускает использование первого только на небольших объемах данных в случае ввода ограничения на время работы алгоритма. Ту же особенность имеют алгоритмы CLUDISPO и AntSA-CLU, качество работы которых значительно выше, чем у других алгоритмов, рассмотренных в табл. 4.

Сложность алгоритма *k*-means оценивается как $\Theta(I*K*N*M)$ [25], где M – временные затраты на расчет расстояния между парой объектов (документом и центроидом), $K*N$ – количество производимых расчетов расстояния между документом и центроидом (K – число центроидов, N – число документов), I – число итераций алгоритма. Отмечено [25], что число итераций, как правило, не является большой величиной, более того, может быть фиксировано в алгоритме (например, ограничено 10-ю итерациями). Если число итераций фиксировать, то оценка сложности станет $\Theta(K*N*M)$. Если ввести предположение, что число кластеров ограничено некоторым небольшим значением, и фиксировать максимально возможное число кластеров (например, в пределах 1000 штук), то сложность алгоритма можно оценить как $\Theta(N*M)$, т.е. сложность будет линейно зависеть от числа документов и скорости вычисления расстояния между документом и центроидом. Сложность иерархической кластеризации (HAC) оценивается как:

- время, необходимое для построения таблицы расстояний между каждой парой объектов $\Theta(N^2*M)$, где N – число документов, M – время расчета расстояния между парой документов;
- время, необходимое для работы алгоритма, когда таблица расстояний между объектами уже построена $\Theta(N^2 * \log N)$.

Как отмечено в [25], можно считать сложность HAC как $\Theta(N^2*M)$ (в [25] авторы пишут о $\Theta(N^2)$, указывая, что это сложность относительно временных затрат на расчет расстояний между каждой парой объектов). Обоснование: временные затраты, требуемые на расчет расстояний между парами объектов и построение таблицы расстояний, как правило, являются более значительными, чем временные затраты на последующую работу алгоритма. Таким образом, в отличие от *k*-means, в HAC наблюдается не линейная, а квадратичная зависимость от размера входных данных, что приводит к значительно более быстрому росту временных затрат на работу алгоритма с увеличением объема данных и делает невозможным использование HAC в режиме реального времени, начиная с некоторого момента. По этой причине при увеличении объема обрабатываемых данных потребуется перейти к использованию *k*-means. Аналогичное произойдет и в случае использования CLUDIPSO или AntSA-CLU.

Алгоритм CLUDIPSO, также лежащий внутри AntSA-CLU, является реализаций подхода к решению задачи кластеризации как задачи оптимизации. Задача алгоритма – разбить документы на кластеры так, чтобы доставить оптимум выбранной внешней оценки качества. Для решения задачи оптимизации в CLUDIPSO используется метод роя частиц (PSO) [7]. В экспериментах [14] в качестве оптимизируемой функции был выбран Global Silhouette coefficient (GS), расчет которого требует вычисления для каждого документа:

- средней непохожести этого документа с другими документами его кластера;
- средней непохожести документа с другими документами ближайшего кластера [7].

В худшем случае такой расчет даст $O(N^2*M)$, где M – время, затрачиваемое на расчет непохожести двух документов, а N – число документов, в отличие от *k*-means: $O(N*M)$, где M – время на расчет расстояния между парой объектов.

Фактором, влияющим на время работы алгоритма *k*-means, являются затраты на расчет расстояний между объектами и центроидами. Снижение пространства признаков уменьшает размерность векторов, представляющих документы, и, как следствие, затраты на расчет расстояний между объектами. Использование только существительных и прилагательных для построения пространства признаков позволяет снизить размерность для коллекции CICling на 27%, для коллекции SEPLN-CICling на 25%, для коллек-

ции EasyAbstracts на 23%. Можно ожидать снижения временных затрат на кластеризацию с помощью *k*-means при таком изменении размерности. Как показывает табл. 2, в этом случае не происходит потери в качестве кластеризации, напротив, наблюдается повышение качества для всех трех тестовых коллекций, особенно в случае применения расширенного списка стоп-слов. Таким образом, обосновывается использование только существительных и прилагательных, а также расширенного списка стоп-слов для представления документов в случае алгоритма *k*-means. Для получения экспериментального подтверждения мы провели оценку времени работы алгоритма:

– для случая, когда для задания пространства признаков были использованы все слова из словаря коллекции (далее – «базовый эксперимент»);
– для случая, когда были использованы только существительные и прилагательные, не попавшие в расширенный список стоп-слов (далее – «улучшенный эксперимент»).

Для каждого из двух случаев было поставлено по 20 экспериментов на каждой из трех тестовых коллекций. Результатом каждого из 20 экспериментов явилось среднее время работы алгоритма кластеризации по результатам 1000 запусков. В каждом блоке результатов из 20 экспериментов выделялось минимальное, максимальное и среднее значения времени. Эксперимент показал, что для каждой из трех тестовых коллекций во время «улучшенного эксперимента» наблюдается сокращение временных затрат не менее чем на 28% по сравнению с «базовым экспериментом». При этом сравнивались лучшие, худшие и средние значения, полученные по результатам 20 экспериментов, для обоих случаев. Теоретические выводы и экспериментальные результаты показывают, что использование для представления пространства признаков только существительных и прилагательных из словаря коллекции, не попавших в расширенный список стоп-слов, приводит к повышению качества работы алгоритма *k*-means и одновременно позволяет сократить время работы алгоритма.

Заключение

Для алгоритма *k*-means было выявлено преимущество описания пространства признаков при помощи существительных и прилагательных (за исключением слов, входящих в расширенный список стоп-слов) вместо всего словаря коллекции. Показано, что в этом случае одновременно достигается улучшение качества кластеризации и снижение временных затрат. Преимуществом *k*-means по сравнению с другими алгоритмами является вычислительная простота, благодаря которой возможно его применение на достаточно больших массивах данных в режиме реального времени, в то время как использование других алгоритмов оказывается затруднительным.

Показано, что алгоритм иерархической кластеризации, определяющий расстояние между кластерами как среднее расстояние между каждой парой объектов в разных кластерах, дает высокие результаты по сравнению с *k*-means, поэтому на небольших массивах данных использование метода межгруппового среднего является более выгодным. Применение для описания пространства признаков только существительных и прилагательных вместо всего словаря коллекции в случае иерархической кластеризации приводит к ухудшению качества работы, вследствие чего не является оправданным. Полученные результаты экспериментов дополняют имеющиеся данные о качестве работы алгоритмов иерархической кластеризации коротких текстов узкой тематики [11]. Показано, что метод «межгрупповое среднее» иерархической кластеризации дает высокие показатели по сравнению с рядом других алгоритмов, используемых в области [7, 8, 11–14].

В лучшем случае *k*-means показывает достаточно хорошие результаты для выбранных коллекций. Напомним, что качество работы данного алгоритма в каждом из экспериментов определялось выбором исходных центроидов. Можно ввести предположение о приемлемости использования алгоритма *k*-means для решения задачи кластеризации узкотематических текстов короткой длины при условии выработки стратегии поиска исходных центроидов. Такая стратегия должна позволить значительно повысить результат работы алгоритма в среднем и приблизить его к лучшим результатам. Тем не менее, выработка такой стратегии (особенно за линейное время) является отдельной нетривиальной задачей и требует специального исследования, на котором мы планируем сосредоточиться в следующей работе.

Литература

1. Bernardini A., Carpineto C., D'Amico M. Full-Subtopic Retrieval with Keyphrase-Based Search Results Clustering // Proc. of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology. IEEE Computer Society, 2009. V. 1. P. 206–213.
2. Zhang D., Dong Y. Semantic, Hierarchical, Online Clustering of Web Search Results // Proc. of the 6th Asia-Pacific Web Conference (APWeb 2004). Lecture Notes in Computer Science. 2004. V. 3007. P. 69–78.
3. Zeng H.-J., He Q.-C., Chen Z., Ma W.-Y., Ma J. Learning to cluster web search results // Proc. of the 27th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '04). NY: ACM Press, 2004. P. 210–217.

4. Gutwin C., Paynter G., Witten I., Nevill-Manning C., Frank E. Improving browsing in digital libraries with keyphrase indexes // J. Decision Support Systems. 1999. V. 27. N 1-2. P. 81–104.
5. Popova S., Khodyrev I., Egorov A., Logvin S., Gulyaev S., Karpova M., Muromtsev D. Sci-Search: Academic Search and Analysis System Based on Keyphrases // Proc. of the 4th Conference on Knowledge Engineering and Semantic Web (KESW 2013). Communications in Computer and Information Science series. 2013. V. 394. P. 281–288.
6. Alexandrov M., Gelbukh A., Rosso P. An Approach to Clustering Abstracts // Proc. of the 10th International Conference NLDB-05. Lecture Notes in Computer Science. 2005. V. 3513. P. 8–13.
7. Cagnina L., Errecalde M., Ingaramo D., Rosso P. A discrete particle swarm optimizer for clustering short text corpora // Proc. of the 3rd International Conference on Bioinspired Optimization Methods and their Applications (BIOMA08). Ljubljana, Slovenia, 2008. P. 93–103.
8. Errecalde M., Ingaramo D., Rosso P. ITSA: An Effective Iterative Method for Short-Text Clustering Tasks // Proc. 23rd International Conference on Industrial, Engineering & Other Applications of Applied Intelligent Systems (IEA/AIE 2010). Lecture Notes in Artificial Intelligence. 2010. V. 6096. P. 550–559.
9. Ramírez-de-la-Rosa G., Montes-y-Gómez M., Solorio T., Villaseñor-Pineda L. A document is known by the company it keeps: neighborhood consensus for short text categorization // Lang Resources and Evaluation. 2012. V. 47. P. 127–149.
10. Romero F.P., Julián-Iranzo P., Soto A., Ferreira-Satler M., Gallardo-Casero J. Classifying unlabeled short texts using a fuzzy declarative approach // Lang Resources and Evaluation. 2013. V. 47. P. 151–178.
11. 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, свободный. Яз. англ. (дата обращения 23.12.2013).
12. Pinto D., Rosso P., Jiménez H. A Self-Enriching Methodology for Clustering Narrow Domain Short Texts // Computer Journal. 2011. V. 54. N 7. P. 1148–1165.
13. Pinto D., Jimenez-Salazar H., Rosso P. Clustering abstracts of scientific texts using the transition point technique // Proc. of the 7th International Conference CICLing 2006. Lecture Notes in Computer Science. 2006. V. 3878. P. 536–546.
14. Errecalde M., Ingaramo D., Rosso P. A new AntTree-based algorithm for clustering short-text corpora // J. Computer Sci. Technol. V. 10. N 1. P. 1–7.
15. Stein B., Meyer zu Eissen S., Potthast M. Syntax versus Semantics: Analysis of Enriched Vector Space Models // Third International Workshop on Text-Based Information Retrieval (TIR 06)/ Eds B. Stein, O. Kao. Trento, Italy: University of Trento, 2006. P. 47–52.
16. Meyer zu Eissen S., Stein B., Potthast M. The Suffix Tree Document Model Revisited // Proc. of the 5th International Conference on Knowledge Management (I-KNOW 05). Graz, Austria, 2005. P. 596–603.
17. You W., Fontaine D., Barhes J.-P. An automatic keyphrase extraction system for scientific documents // Knowledge and Information Systems. 2013. V. 34. N 3. P. 691–724.
18. Hulth A. Improved automatic keyword extraction given more linguistic knowledge // Proc. of the 2003 Conference on Empirical Methods in Natural Language Processing (EMNLP'03). Stroudsburg, 2003. P. 216–223.
19. Mihalcea R., Tarau P. TextRank: Bringing order into texts // Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP '04). Stroudsburg, 2004. P. 404–411.
20. Wan X., Xiao J. Exploiting Neighborhood Knowledge for Single Document Summari-zation and Keyphrase Extraction // ACM Transactions on Information Systems. 2010. V. 28. N 2. Article 8.
21. Zesch T., Gurevych I. Approximate Matching for Evaluating Keyphrase Extraction // Proc. of the International Conference on Recent Advances in Natural Language Processing (RANLP 2009). 2009. P. 484–489.
22. Popova S., Khodyrev I. Ranking in keyphrase extraction problem: is it useful to use statistics of words occurrences? // RuSSUR 2013. Режим доступа: http://romip.ru/russiras/doc/2013_for_participant/russirysc2013_submission_18_1.pdf, свободный. Яз. англ. (дата обращения 27.12.2013).
23. Попова С.В., Ходырев И.А. Извлечение и ранжирование ключевых фраз в задаче аннотирования // Научно-технический вестник информационных технологий, механики и оптики. 2013. № 1 (83). С. 81–85.
24. Popova S., Kovriguina L., Muromtsev D., Khodyrev I. Stop-words in Keyphrase Extraction Problem // Proc. of 14th Conference of Open Innovations Association FRUCT. Helsinki, Finland, 2013. P. 113–121.
25. Manning C., Raghavan P., Schütze H. Introduction to Information Retrieval. Cambridge University Press, 2009.544 p.
26. Meyer zu Eissen S., Stein B. Analysis of Clustering Algorithms for Web-based Search // Proc. of the 4th International Conference on Practical Aspects of Knowledge Management (PAKM 2002). Lecture Notes in Artificial Intelligence. 2002. V. 2569. P. 168–178.

27. Stein B., Meyer zu Eissen S., Wißbrock F. On Cluster Validity and the Information Need of Users // Proc. of the 3rd IASTED International Conference on Artificial Intelligence and Applications (AIA 03). Benalmádena, Spain, 2003. P. 216–221.
28. Tsatsaronis G., Varlamis I., Norvag K. SemanticRank: Ranking Keywords and Sentences Using Semantic Graphs // Proc. of the 23rd International Conference on Computational Linguistics (Coling'10). Beijing, China, 2010. P. 1074–1082.
29. Hasan K.S., Ng V. Conundrums in Unsupervised Keyphrase Extraction: Making Sense of the State-of-the-Art // Proc. of the 23rd International Conference on Computational Linguistics: Posters (Coling'10). Beijing, China, 2010. P. 365–373.
30. Ingaramo D., Errecalde M., Cagnina L., Rosso P. Particle Swarm Optimization for lustering short-text corpora // Computational Intelligence and Bioengineering/ Eds F. Masulli, A. Micheli, A.Sperduti. IOS Press, 2009. P. 3–19.
31. Azzag H., Monmarche N., Slimane M., Venturini G. AntTree: A new model for clustering with artificial ants // Proc. of the 2003 Congress on Evolutionary Computation (CEC '03). IEEE Press, 2003.V. 4. P. 2642–2647.
32. Stein B., Meyer zu Eißen S. Document Categorization with MAJORCLUST // Proc. of the 12th Workshop on Information Technology and Systems (WITS 02) / Eds A. Basu, S. Dutta. Barcelona, Spain: Technical University of Barcelona, 2002. P. 91–96.

Попова Светлана Владимировна – инженер, ст. преподаватель, Санкт-Петербургский государственный университет; Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики, Санкт-Петербург, Россия, svp@list.ru

Данилова Вера Владимировна – аспирант, Автономный университет Барселоны, Барселона, Испания, maolve@gmail.com

Svetlana Popova – engineer, senior lecturer, Saint Petersburg State University; Saint Petersburg National Research University of Information Technologies, Mechanics and Optics, Saint Petersburg, Russia, svp@list.ru

Vera Danilova – postgraduate, Autonomous University of Barcelona, Barcelona, Spain, maolve@gmail.com

УДК 004.925.4

ИСПОЛЬЗОВАНИЕ КОНТЕЙНЕРА BC7 ДЛЯ ХРАНЕНИЯ ТЕКСТУР С ГЛУБИНОЙ ЦВЕТА 10 БИТ

И.В. Перминов[a], Т.Т. Палташев[a, b, c]

[a] Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики, Санкт-Петербург, Россия, i.am.perminov@gmail.com
[b] Северо-Западный политехнический университет, Фримонт, Калифорния, США, timpal@mail.npu.edu
[c] Advanced Micro Devices (AMD), Калифорния, США, timpal@mail.npu.edu

Изображения с высокой глубиной цвета обладают гораздо лучшими возможностями воспроизведения цветов и плавных градиентов, особенно при использовании устройств с широким цветовым охватом. В работе рассматривается проблема сжатия текстур с глубиной цвета 10 бит, применяемых в трехмерной компьютерной графике. Предложен метод хранения подобных текстур с использованием стандартного формата сжатия текстур BC7, рассмотрен тип блока BC7, с помощью которого можно закодировать цвета с точностью, превышающей 8 бит, и показаны необходимые изменения в аппаратуре декодера. Предложенный подход обладает обратной совместимостью с существующими декодерами. В ходе исследования была разработана программная реализация кодека на основе компрессора bc7_gpu. Сравнение исходного и предлагаемого кодека показало уменьшение ошибок сжатия для изображений с высокой глубиной цвета.

Ключевые слова: сжатие текстур, BC7, block compression, глубина цвета, Direct3D.

USAGE OF BC7 CONTAINER FOR STORING TEXTURES WITH 10-BIT COLOR DEPTH

I. Perminov[d], T. Paltashev[d, e, f]

[d] Saint Petersburg National Research University of Information Technologies, Mechanics and Optics, Saint Petersburg, Russia, i.am.perminov@gmail.com
[e] Northwestern Polytechnic University, Fremont, California, USA, timpal@mail.npu.edu
[f] Advanced Micro Devices (AMD), California, USA, timpal@mail.npu.edu

High color depth images can more accurately reproduce colors and smooth color transitions without banding artifacts, especially for wide color gamut devices. The paper deals with texture compression with 10-bit color depth applied in 3D graphics. A method for storing of such textures in standard BC7 blocks is proposed. The example of BC7 block applicable for storing of smooth texture information with an accuracy exceeding 8-bit is also given. The proposed approach has a backward compatibility with current hardware. Additional hardware cost to support 10-bit decoding is expected to be low. An overview of