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

ОЦЕНКА ЭФФЕКТИВНОСТИ МЕТОДА ПАРАЛЛЕЛЬНОЙ РЕАЛИЗАЦИИ ПРОЦЕССА КЛАСТЕРИЗАЦИИ ТЕКСТОВЫХ ДОКУМЕНТОВ НА ОСНОВЕ АЛГОРИТМА FRIS-CLUSTER [*]

Представлен вариант параллельного выполнения некоторых этапов кластеризации документов с использованием алгоритма FRiS-Cluster. Приведены количественные оценки времени выполнения процесса, наглядно демонстрирующие преимущества внедрения параллельной реализации на различных этапах обработки: при предварительном анализе документов, включающем вычисление мер сходства, а также частично при выполнении непосредственно процесса кластеризации.

Ключевые слова: кластеризация текстовых документов, параллельные алгоритмы.

Введение

Интенсивный рост объема генерируемой информации, представленной в электронном виде, в том числе и научных документов, делает актуальным решение задачи кластеризации документов, т. е. разбиения множества документов на заранее неизвестное количество подмножеств различных тематик. При вовлечении в процесс нового документа необходимо выполнить его отнесение к одной из существующих групп либо создать новую, в которой добавляемый документ будет являться центроидом. Добавление нового документа не является задачей с большой вычислительной сложностью при известных метриках, на основе которых выполняется расчет меры сходства документа с документами из групп известных тематик. Но данный способ вовлечения документов имеет свои недостатки: нельзя бесконечно долго добавлять новые элементы без учета параметров всех документов, имеющихся в разделяемой выборке. При полной рекластеризации имеющихся документов удается избежать ложного отнесения нового материала к уже сформированной группе, схожей с документом характеристикой, которая, вполне возможно, имеет довольно низкий приоритет. С другой стороны, периодическое выполнение процесса разбиения на кластеры является задачей, сложность которой возрастает с ростом количества обрабатываемых документов.

В настоящее время большой интерес представляет использование в процессе решения задач кластеризации вычислительных систем, состоящих из множества вычислительных узлов. Но, очевидно, при использовании таких систем необходимо иметь версии алгоритмов, которые могут эффективно работать на системах с параллельной архитектурой.

[*] Работа выполнена при частичной поддержке РФФИ (проекты № 10-07-00302, 11-07-00561, 12-07-00472), президентской программы «Ведущие научные школы РФ» (грант НШ 6293.2012.9) и интеграционных проектов СО РАН.

Принцип многопроцессорной обработки как способ повышения общей эффективности можно описать следующим образом: распределить вычислительный процесс на N вычислительных узлов, осуществляющих отдельные фрагменты вычислительного алгоритма. Хорошо распараллеливаемыми являются вычислительные задачи с многократно повторяемыми вычислениями при вариациях некоторых начальных условий для каждого цикла вычислений. При этом желательно, чтобы в таких задачах параметры последующих циклов вычислений имели бы минимально выраженную зависимость от результатов предыдущих циклов.

Алгоритмы, которые используются в настоящее время для задач кластеризации и категоризации документов, являются последовательными и используют имеющиеся вычислительные ресурсы далеко не в полную мощность. В них имеются ограничения, не позволяющие выполнять процесс обработки данных в несколько вычислительных потоков. Во-первых, это строгая последовательность выполняемых операций и обязательное завершение предыдущего логического этапа обработки до начала последующего. Во-вторых, все вычисления производятся на основе данных, полученных при анализе всего массива обрабатываемых документов, что, в свою очередь, довольно сильно усложняет реализацию, а также повышает накладные расходы, связанные с передачей информации между узлами системы.

Цель работы заключается в создании математического обеспечения и на его основе вычислительного комплекса, реализующего решение задачи кластеризации входящего множества документов при помощи алгоритма FRiS-Cluster, на основе функции конкурентного сходства. Процесс разбиения должен быть разбит на этапы, которые будут эффективно выполняться параллельно на N вычислительных узлах системы.

Основные задачи работы можно сформулировать следующим образом.

1. Реализация варианта алгоритма кластеризации FRiS-Cluster с возможностью его выполнения параллельно на некотором количестве вычислительных узлов.

2. Оценка эффективности реализации, показывающая, в какой мере данный метод дает прирост производительности, т. е. уменьшение времени выполнения процесса, в зависимости от добавляемого количества вычислительных узлов.

3. Тестирование методики на множестве научных статей, находящихся в рамках узкой специализированной тематики.

4. Создание интерфейса для работы с предложенным комплексом в интерактивном режиме, что обеспечит решение практических задач, а также позволит выполнять дальнейшие исследования на реальной выборке большого объема при оценивании как качества процесса разделения на кластеры, так и скорости выполнения данного процесса.

Алгоритм кластеризации

В качестве алгоритма, на основе которого выполняется разработка параллельной реализации методики кластеризации документов, выбран алгоритм FRiS-Cluster, автором которого является Н. Г. Загоруйко В основе данного алгоритма лежит функция конкурентного сходства или FRiS-функция – мера сходства двух объектов, вычисляемая относительно некоторого иного объекта [1; 2].

FRiS-функция в отличие от других существующих мер сходства позволяет не просто оценивать понятия «далеко» или «близко», «похож» или «не похож», но и давать количественную оценку схожести. Такой подход дает возможность учитывать большее число факторов при классификации. Исследования показали, что FRiS-функция хорошо имитирует человеческий механизм восприятия сходства и различия. Это позволяет использовать ее как базовый элемент для различных типов задач, включая задачи кластеризации документов.

Методику кластеризации можно описать следующим образом. Если расстояния от объекта Z до двух ближайших объектов a и b равны R_a и R_b, то сходство Z с объектом a равно

$$F(a) = (R_b - R_a) / (R_a + R_b).$$

Значения F меняются в пределах от +1 до −1. Если контрольный объект Z совпадает с объектом a, то $R_a = 0$ и $F(a) = 1$, а $F(b) = -1$. При расстояниях $R_a = R_b$ значения $F(a) = F(b) = 0$, что указывает на границу между образами. При использовании алгоритма FRiS-Cluster на его первых шагах все N объектов заданного множества A принадлежат одному кластеру.

В связи с этим вводится конкурирующее множество из виртуальных объектов, отстоящих от каждого объекта множества на расстояние R^*. Произвольный объект M_i множества A назначается центральным объектом (называемым также столпом) первого кластера, оцениваются расстояния R_j до него от всех остальных объектов M_j и определяются значения функции

$$F(i) = (R^* - R_j) / (R^* + R_j).$$

Далее вычисляется сумма значений функций сходства Fs всех объектов первого кластера со своим центром. Затем в качестве центра второго кластера выбирается объект, набравший наибольшее значение Fs в конкуренции с центром первого кластера. Процесс увеличения числа кластеров k останавливается, когда достигается первый локальный максимум функции $Fs(k)$ либо максимальное количество кластеров, которое задается исходным параметром.

Вычисление меры сходства

В качестве шкал для определения меры сходства между двумя документами можно использовать атрибуты библиографического описания данных документов (метаданные): авторы, издание, аннотация, предопределенные ключевые слова. Но научные статьи обычно являются слабоструктурированными документами, т. е. они далеко не всегда имеют определенный набор метаданных [3]. Наиболее точно отражают тематику материала такие характеристики, как ключевые слова и ключевые выражения, полученные из основного текста документа. Это влечет за собой проблемы, связанные с морфологическим анализом содержания документа для определения списка ключевых одиночных слов и составных многословных ключевых выражений, наиболее точно описывающих действительную тематику документа. Причем желательно, чтобы данный анализ производился без априорного знания тематики материала и соответственно без использования предметных тезаурусов и словарей словоформ для них.

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 производили выделение ключевых словосочетаний по морфологическим шаблонам с использованием программного продукта компании «Яндекс» – так называемого стеммера (http://company.yandex.ru/technology/mystem/), который является бесплатным для некоммерческих целей. При фильтрации и разборе производился отсев стоп-слов. Ключевые словосочетания отбираются по морфологическим шаблонам с учетом словоформ языка [4] [1]. Подробно методика выделения ключевых слов и ключевых словосочетаний описана в статье [5]. Для демонстрации эффективности работы методики выделения значимой информации из текстовых документов в табл. 1 приведены результаты работы по одной из статей экспериментальных данных: Астафьев Д. А. «Общие и индивидуальные особенности эволюции, регионального и глубинного строения осадочных и нефтегазоносных бассейнов Земли» (Материалы трудов международной конференции «Современное состояние наук о Земле», посвященной памяти Виктора Ефимовича Хаина, Москва, 1–4 февраля 2011 г.).

Как видно из приведенных данных, качество анализа содержания документа находится на хорошем уровне. Список значащих слов и выражений достаточно точно описывает тематику анализируемого документа. Кроме того, с помощью фильтрации в список не попадают незначащие стоп-слова, которые могут существенно ухудшить качество оценки меры сходства между документами.

Также на основе сделанных ранее экспериментов можно утверждать, что смешанный критерий сходства документов, вычисленный на основе данных о ключевых словах и словосочетаниях, является оптимальным для решения задач кластеризации документов заранее неизвестной тематики. На основе анализа содержания вычислялась мера сходства m документов,

[1] См. также: *Шаров С. А.* Частотный словарь русского языка. URL: http://www.artint.ru/projects/frqlist.asp

Таблица 1

Пример полученных ключевых слов и словосочетаний *

Ключевое		Количество вхождений
слово	словосочетание	
	Осадочный чехол	16
	Осадочный бассейн	8
	Коромантийный оболочка	7
	Надрифтовый депрессия	6
	Земной кора	6
	Нефтегазоносный бассейн	5
	Глубинный строение	5
	Переходный комплекс	4
	Бассейн земля	4
	Столбчатый тело	3
	Размещение зона	3
Бассейн		19
Чехол		16
Строение		14
Зона		14
Формирование		11
Земля		10
Модель		9

* Слова приведены в начальной форме.

которая может принимать значение в интервале от 0 до 1,причем, значение меры равняется 0 при полном различии документов и равняется 1 при полном их сходстве. В алгоритме FRiS-Cluster с помощью данной меры производилось вычисление расстояния между документами, которое вычислялось как $1 - m$. Другими словами, чем меньше мера сходства между сравниваемыми документами, тем на большем расстоянии они находятся.

Распараллеливание алгоритма

При большом количестве обрабатываемых документов время выполнения процесса кластеризации очень сильно растет, и вполне оправдана цель ускорения обработки. В нашем случае можно выделить следующие этапы работы.

1. *Сбор и загрузка исходных данных*. Это действие зависит от источника данных. Это могут быть данные в текстовых форматах, хорошо поддающиеся обработке и загрузке, импорт из внешних веб-ресурсов, который зависит от производительности внешнего веб-сервера, а также от пропускной способности сети, либо импорт из форматов, обработка которых имеет некоторую специфику, например pdf. Оценивать трудоемкость данного действия нужно в каждом конкретном случае. В нашей работе параллельная реализация на этом этапе не использовалась.

2. *Подготовительный этап*. Включает первичную обработку документов. В нашем случае сюда входят процессы выделения ключевых термов и словосочетаний из текста документов, а также вычисление меры сходства между документами. На этом этапе внедрение параллельной обработки может дать существенный выигрыш в производительности. Причем действия по анализу содержания являются независимыми друг от друга и не требуют наличия полной информации на всех вычислительных узлах.

3. *Процесс кластеризации*. Исходя из методики работы алгоритма FRiS-Cluster, можно сделать вывод, что самым сложным вычислительным процессом является обход всех объек-

тов выборки и проверка каждого на роль столпа. Понятно, что данный процесс может хорошо выполняться параллельно, хотя и требует наличия информации о расстояниях между объектами на всех вычислительных узлах.

4. *Визуализация результата*. Является вспомогательным действием, облегчающим работу с исходными данными и полученными кластерами. Не требуется больших вычислительных мощностей для работы, независимо от количества документов в выборке. Оптимизация работы на данном этапе достигается оптимизацией на уровне хранения данных, т. е. на уровне сервера базы данных.

Исходя из вышеизложенного для использования параллельной реализации подходят этап подготовки данных и часть действий этапа непосредственной кластеризации.

Вычислительные эксперименты

В качестве исходных данных для выполнения анализа и кластеризации использовались материалы трудов международной конференции «Современное состояние наук о Земле», посвященной памяти Виктора Ефимовича Хаина, проходившей 1–4 февраля 2011 г. в Москве. Исходная выборка включает 488 документов. Выбор исходных данных обусловлен двумя причинами. Во-первых, важно было проверить работу методики выделения ключевых термов из текстов именно геологической тематики без использования тезауруса предметной области, поскольку эта область знаний достаточно сильно насыщена специфическими терминами и определениями, проблемы с анализом которых потенциально могли выявиться в процессе работы. Во-вторых, все документы находятся в рамках одной, уже достаточно узкой тематики, и дальнейшее их разбиение вызывает сложности при использовании мер сходимости, основанных только на библиографических описаниях либо заголовках документов.

В первую очередь, производилась разбивка исходного множества на установленное количество кластеров – на 10, 20, 30. Этот параметр (количество результирующих кластеров) является опциональным в алгоритме работы и устанавливается перед началом процесса кластеризации. На рис. 1–3 изображено распределение документов по результирующим кластерам.

Как видно из представленных графиков, несмотря на очевидную близость документов внутри узкой тематики, алгоритм успешно справился с разбиением выборки на части. Кластер с номером центроида, равным 0, включает документы, которые не удалось отнести ни к одной из формируемых групп. Количество таких документов, в зависимости от количества кластеров, варьируется в интервале 7,3–12,7 %, что является приемлемым для работы алгоритмов кластеризации.

Разумеется, более точно оценить качество разбиения выборки настолько узкой тематики возможно только с помощью эксперта в предметной области, так как обрабатываемые материалы не имеют кодов принадлежности к уровням того или иного классификатора. Неэксперту трудно произвести точную оценку корректности разбиения даже после ознакомления с заголовками и короткими аннотациями статей.

Измерение времени выполнения выполнялось следующим образом. Были произведены измерения времени процессов подготовки данных (анализ содержания и вычисление мер близости) и непосредственно кластеризации для 10, 20 и 30 формируемых кластеров на одном вычислительном узле и на нескольких вычислительных узлах для параллельной реализации. Причем структура вычислительной сети при выполнении измерений параллельной реализации алгоритма не была однородной и состояла из трех рабочих станций различной конфигурации (в том числе отличающихся производительностью), связанных в единую сеть. Понятно, что данная сеть не может быть бесконечно расширена для ускорения процесса, так как накладные расходы, связанные с передачей информации будут расти с ростом количества вычислительных узлов.

Результаты измерения времени выполнения последовательной однопоточной реализации приведены в табл. 2, а параллельной реализации, выполняемой на трех вычислительных узлах, в табл. 3.

На рис. 4 наглядно представлена зависимость времени выполнения процесса кластеризации в зависимости от количества вычислительных узлов, а также итогового количества кластеров.

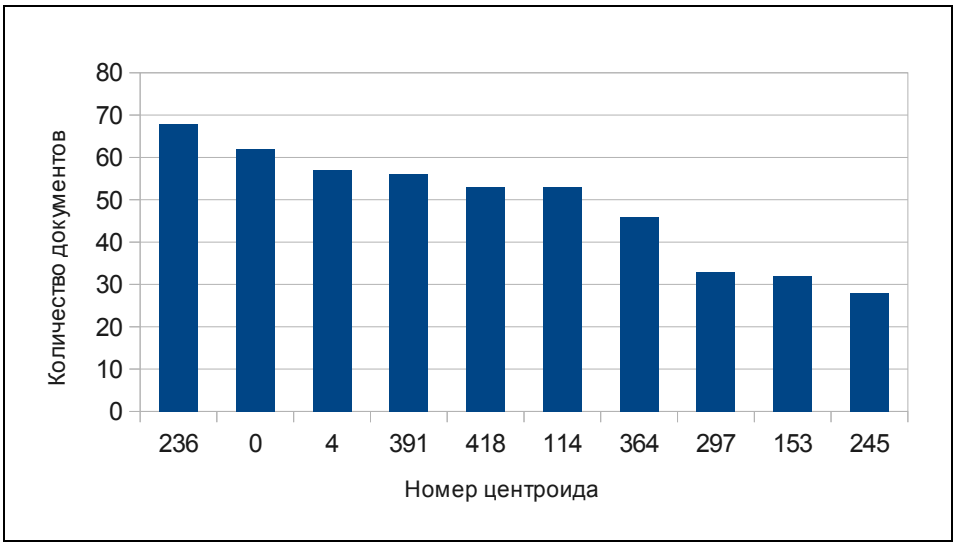

Рис. 1. Количество документов в результирующих кластерах, 10 кластеров

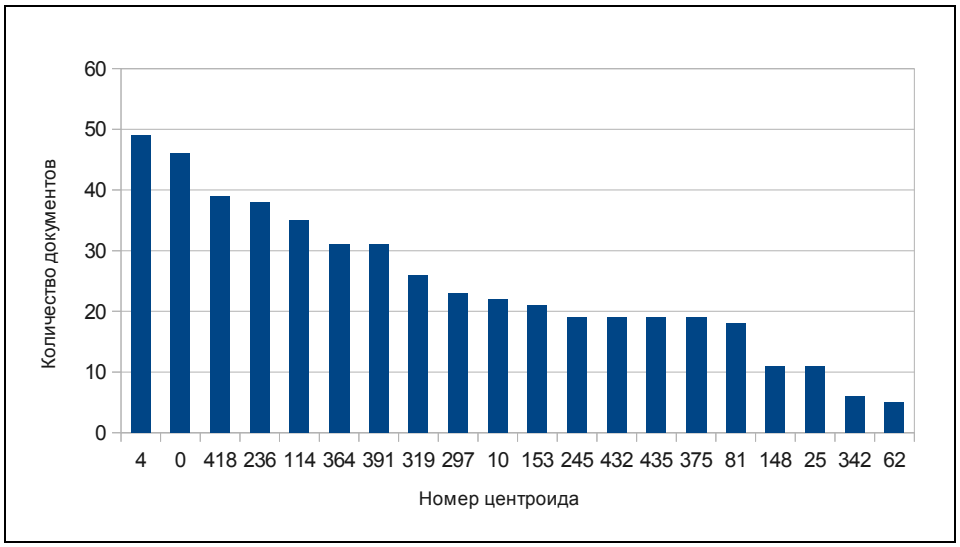

Рис. 2. Количество документов в результирующих кластерах, 20 кластеров

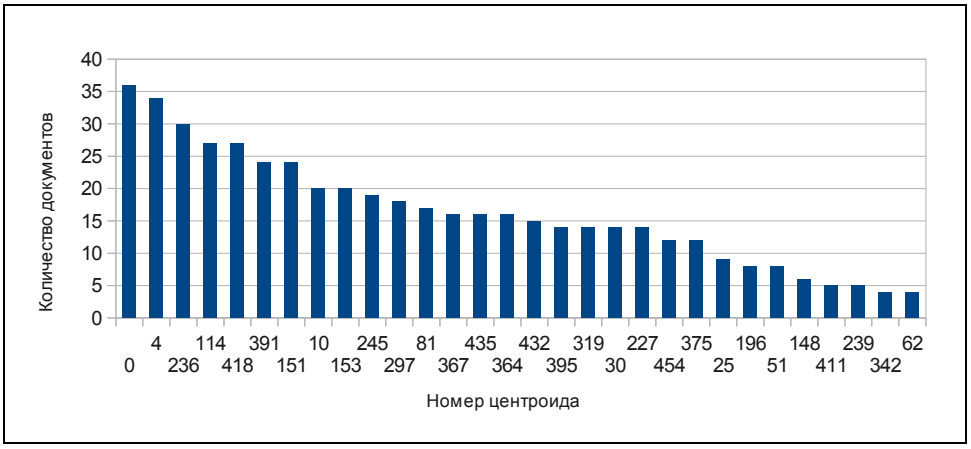

Рис. 3. Количество документов в результирующих кластерах, 30 кластеров

Таблица 2

Время последовательного выполнения процесса, с

Этап работы	Количество кластеров		
	10	20	30
Предварительный анализ данных	290	290	290
Подбор столпов	31	96	200
Итоговое уточнение столпов	8	13	19
Итоговое время	329	399	509

Таблица 3

Время параллельного выполнения процесса, с

Этап работы	Количество кластеров		
	10	20	30
Предварительный анализ данных	99	99	99
Подбор столпов	11	33	68
Итоговое уточнение столпов	8	13	19
Итоговое время	118	145	186

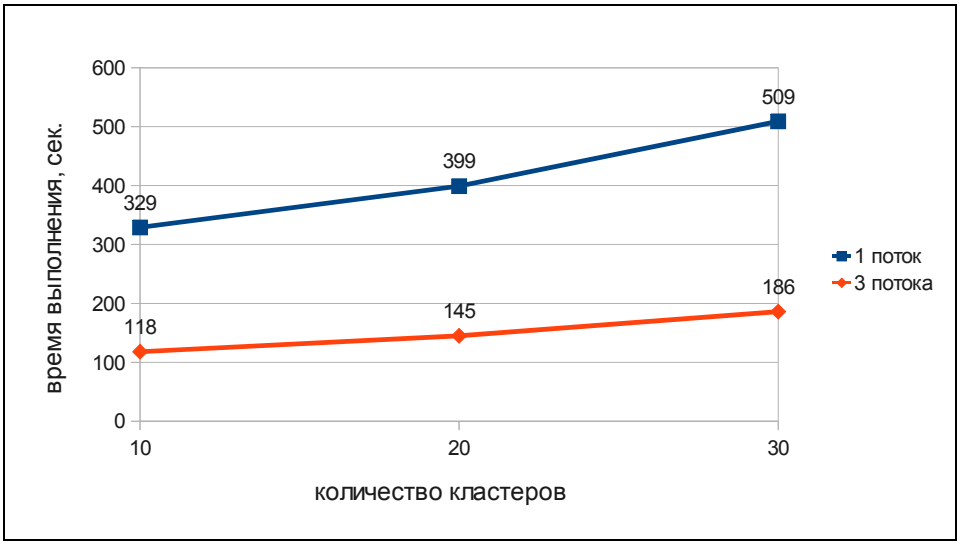

Рис. 4. График зависимости времени выполнения от количества вычислительных узлов

Так как разбиение выборки производилось без учета производительности каждого из вычислительных узлов, сложилась ситуация, что время выполнения не может быть быстрее, чем обработка $1/N$ части выборки на самом медленном узле. Тем не менее приведенные данные наглядно демонстрируют целесообразность использования параллельной реализации различных стадий обработки данных в процессе кластеризации.

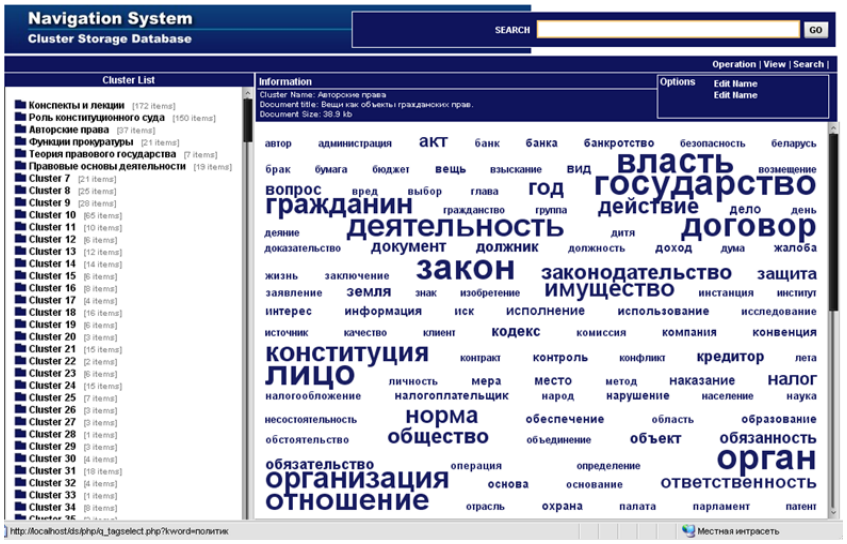

Рис. 5. Вид навигационной системы

Заключение

Реализация алгоритма выполнялась в рамках навигационной системы, которая упрощает работу по добавлению исходных документов и дальнейшей навигации по ним, по итоговым кластерам. Поиск документов через пользовательский интерфейс возможен по полному тексту статьи, по ключевым словам или словосочетаниям, по схожести просматриваемой статьи с другими документами, на основе меры сходства. Вид навигации по ключевым словам можно увидеть на рис. 5. Для удобства и понимания наиболее смежных ключевых слов размер шрифта становится больше с ростом количества вхождений терма в документах обрабатываемой выборки.

Предложенная методика автоматической кластеризации документов в электронном виде позволяет выполнять процесс обработки на системах, состоящих более чем из одного вычислительного узла. В работе приводятся количественные величины оценок времени выполнения при различных исходных данных.

Оценка эффективности процесса при использовании параллельной реализации алгоритма FRiS-Cluster на основе функции конкурентного сходства по сравнению с классическим линейным методом интеллектуальной обработки данных демонстрирует неоспоримый выигрыш в производительности, даже несмотря на то, что не все этапы обработки данных в процессе кластеризации могут выполняться параллельно на наборе вычислительных узлов.

Список литературы

1. *Борисова И. А., Загоруйко Н. Г.* Функции конкурентного сходства в задаче таксономии // Материалы Всерос. конф. с международным участием «Знания – Онтологии – Теории» (ЗОНТ-07). Новосибирск, 2007. Т. 2. С. 67–76.

2. *Борисова И. А., Загоруйко Н. Г.* Использование FRiS-функций для решения задачи SDX // International Conference «Classification, Forecasting, Data Mining» CFDM 2009. Varna, 2009. P. 110–116.

3. *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.

4. *Киселев М.* Метод кластеризации текстов, основанный на попарной близости термов // Сборник работ участников конкурса «Интернет-математика 2007». Екатеринбург: Изд-во Уральского университета, 2007. С. 74–83.

5. *Барахнин В. Б., Ткачев Д. А.* Кластеризация текстовых документов на основе составных ключевых термов // Вестн. Новосиб. гос. ун-та. Серия: Информационные технологии. 2010. Т. 8, вып. 2. С. 5–14.

Материал поступил в редколлегию 09.11.2012

V. B. Barakhnin, D. A. Tkachev

EVALUATING THE EFFECTIVENESS OF THE METHOD OF THE PARALLEL IMPLEMENTATION OF THE PROCESS OF CLUSTERING TEXT DOCUMENTS ON THE BASIS OF THE ALGORITHM FRIS-CLUSTER

This paper presents a variant of the parallel execution of certain phases of the clustering of documents using the algorithm FRiS-Cluster. We give quantitative values of time the process is to demonstrate the benefits of implementing the parallel implementation of the various stages of processing: a preliminary analysis of documents, which includes calculation of similarity measures, and partly in the performance of the clustering process itself.

Keywords: clustering text documents, parallel algorithms.