БОРИСЮК ФЕДОР ВЛАДИМИРОВИЧ

СИСТЕМА ПОИСКА ТЕКСТОВЫХ ДОКУМЕНТОВ НА ОСНОВЕ АВТОМАТИЧЕСКИ ФОРМИРУЕМОГО ЭЛЕКТРОННОГО КАТАЛОГА

Специальность 05.13.18

Математическое моделирование, численные
методы и комплексы программ

АВТОРЕФЕРАТ

диссертации на соискание ученой степени
кандидата технических наук

Нижний Новгород

2010

Работа выполнена на кафедре математического обеспечения ЭВМ факультета вычислительной математики и кибернетики ГОУ ВПО «Нижегородский государственный университет им. Н.И.Лобачевского»

Научный руководитель доктор технических наук,
профессор
Швецов В.И.

Официальные оппоненты доктор технических наук, профессор,
Карпычев В.Ю. (г. Н.Новгород)

доктор технических наук, профессор,
Подольский В.Е. (г. Тамбов)

Ведущая организация ГОУ ВПО “Саратовский государственный технический университет”.

Защита состоится «28» декабря 2010 г. в ____ часов на заседании Совета по защите докторских и кандидатских диссертаций Д 212.166.13 при Нижегородском государственном университете им. Н.И. Лобачевского по адресу: 603950, Н. Новгород, пр. Гагарина, д. 23.

С диссертацией можно ознакомиться в фундаментальной библиотеке Нижегородского государственного университета им. Н.И. Лобачевского.
Автореферат размещен на сайте http://www.unn.ru университета ННГУ
им. Н.И. Лобачевского.

Автореферат разослан «___»_____2010 г.

Ученый секретарь диссертационного совета
кандидат физико-математических наук,
доцент _____ / Савельев В.П.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы исследований. В настоящее время в различных хранилищах знаний (электронных и традиционных) накоплены огромные массивы информации. При этом по причине больших объемов информации, ее слабой структурированности, представления в неэлектронном виде, получение актуальной и полной информации по конкретной теме является достаточно сложным, а также бесполезной становится большая часть накопленных информационных ресурсов из-за их необозримости. Можно отметить, что решение конкретной научной задачи требует высоких трудозатрат по поиску и анализу информации по теме. Поэтому, в связи с выше сказанным, возникает задача эффективного структурирования, хранения, обработки и поиска в информационных массивах.

Традиционными подходами к решению данной задачи являются: классификационный поиск и поиск по ключевым словам. К классификационному поиску относится поиск с использованием различных тематических классификаторов, рубрикаторов, электронных каталогов, которые позволяют искать (автоматически или вручную) документы в небольшом подмножестве исходной коллекции документов по интересующей пользователя тематике. Рубрикатор (электронный каталог) обычно представляет собой множество рубрик, объединенных в иерархию. К каждой рубрике приписываются соответствующие ее тематике документы. Традиционные каталоги (классификаторы) имеют фиксированную структуру и зачастую не поддерживают высоких темпов развития различных областей знаний в науке и технике, а также требуют высоких временных затрат на адаптацию классификаторов, и классификацию по ним документов.

При втором способе поиска пользователь вводит ключевые слова, отражающие его информационную потребность. При этом результатом поиска как правило является достаточно большое количество документов, среди которых пользователь должен выбрать нужные. Отметим, что одно и то же ключевое слово может соответствовать разным понятиям, поэтому результат поиска заведомо избыточен. Кроме этого, пользователь может ввести ключевые слова не соответствующие интересующему его документу. Для улучшения качества выдаваемых поисковых результатов в данной работе предлагается производить поиск по ключевым словам с использованием категориальной (тематической) информации электронных каталогов. Предлагается помещать (ранжировать) наиболее важные по тематике документы выше в списке результатов. Как было замечено ранее, подготовка нового или адаптация существующего классификатора является достаточно затратной, поэтому требуется

применение новых, более эффективных методов подготовки электронных тематических каталогов. В работе предлагается применить методы текстовой кластеризации для автоматического построения электронного каталога.

Таким образом, актуальной является задача создания новых моделей информационного поиска по ключевым словам с тематическим ранжированием результатов поиска, на основе автоматически построенного с использованием методов автоматической каталогизации (способных без участия человека строить электронные каталоги заданных коллекций текстовых документов) электронного каталога.

Цель работы. Цель работы заключается в создании моделей и методов поиска по ключевым словам с тематическим ранжированием, на основе электронного каталога заданных коллекций текстовых документов (автоматически построенного с использованием разработанных алгоритмов текстовой кластеризации).

Для достижения данной цели были поставлены и решены следующие **задачи**:

- проведены исследования подходов к извлечению текстовых признаков документов, и обзор существующих алгоритмов текстовой кластеризации;
- разработана модель поиска с тематическим ранжированием, на основе автоматически построенного электронного каталога;
- разработана модель автоматического построения электронного каталога, на основе предложенного в работе алгоритма иерархической кластеризации по областям;
- разработаны последовательные и параллельные варианты методов извлечения текстовых признаков, и алгоритмов иерархической текстовой кластеризации, (учитывающие недостатки существующих подходов).
- разработана программная система поиска (поисковая система) на основе модели поиска по ключевым словам с тематическим ранжированием, на основе автоматически построенного электронного каталога текстовых документов;
- проведено исследование эффективности и качества работы предложенных моделей, и алгоритмов с использованием разработанной программной системы.

Методы исследований, достоверность и обоснованность результатов. Для решения поставленных задач были использованы методы математического моделирования, системного анализа, методы математической статистики, кластерного анализа. Эффективность разработанных алгоритмов оценивалась с помощью математических методов анализа алгоритмов. В разработке программного обеспечения применялись методы объектно-ориентированного программирования с использованием

инструментов интегрированной среды разработки Eclipse. Для разработки параллельных версий алгоритма использовались программные средства платформы для распределенных вычислений Apache Hadoop. Достоверность и обоснованность результатов подтверждается корректностью разработанных математических моделей, согласованностью данных экспериментов и научных выводов, сделанных в работе, результатами апробации алгоритмов и разработанной программной системы.

Научная новизна. В работе предложена новая модель поиска по ключевым словам с тематическим ранжированием результатов поиска, на основе автоматически построенного электронного каталога заданных коллекций текстовых документов (без ограничения на тематику и размер исходной текстовой коллекции).

В рамках реализации этой модели разработаны:

- Новый метод тематического ранжирования, основанный на автоматически построенном электронном каталоге.
- Новая модель автоматического построения электронного каталога, основанная на предложенном в работе методе текстовой кластеризации - иерархическая кластеризации по областям текстовых документов, учитывающем недостатки существующих методов иерархической текстовой кластеризации (разработаны последовательные и параллельные варианты предложенного метода кластеризации).
- Методы извлечения текстовых признаков (разработаны последовательные и параллельные варианты предложенных методов), используемые для построения индекса текстовой коллекции, необходимого во время кластеризации и поиска.

Практическая значимость работы. Предложенные в работе новая модель информационного поиска по ключевым словам с тематическим ранжированием, модель автоматического построения электронного каталога, и алгоритмы тематического ранжирования могут быть использованы в качестве поисковой системы по специализированным коллекциям научной литературы, и электронным хранилищам библиотек.

Внедрение. Произведена апробация и внедрение предложенных в данной работе моделей и методов поиска по ключевым словам с тематическим ранжированием, на основе автоматически построенного электронного каталога, в качестве поисковой системы по статьям журнала “Вестник Нижегородского государственного университета им. Н.И. Лобачевского” (http://www.unn.ru/e-library/vestnik.html) на интернет-портале Нижегородского государственного университета им. Н.И. Лобачевского.

Апробация результатов. Результаты диссертации докладывались и обсуждались на всероссийской конференции «Технологии Microsoft в теории и практике программирования 2009» (Н.Новгород, 2009г.), международной научно-практической конференции по графическим информационным технологиям и системам «КОГРАФ-2009» (Н.Новгород,2009), 9-й международной конференции "Высокопроизводительные параллельные вычисления на кластерных системах" (Владимир, ВлГУ, 2009), всероссийской научной школе для молодежи "Управление информационными ресурсами образовательных, научных и производственных организаций" (Магнитогорск, Магнитогорский государственный университет, 2009), всероссийской конференции «Технологии Microsoft в теории и практике программирования 2010» (Н.Новгород, 2010), международном коллоквиуме SYRCoSE (Н.Новгород, 2010), на семинаре кафедры математического обеспечения ЭВМ факультета ВМК ННГУ.

Основные положения, выносимые на защиту

- Новая модель информационного поиска по ключевым словам с тематическим ранжированием, основанная на использовании автоматически построенного электронного каталога.
- Новая модель автоматического построения электронного каталога текстовых документов без ограничения на тематику и размер исходной текстовой коллекции.
- Новый метод текстовой кластеризации - иерархическая кластеризация по областям текстовых документов, учитывающий недостатки существующих алгоритмов иерархической текстовой кластеризации. Последовательные и параллельные версии предложенного метода.
- Архитектура и функциональные возможности разработанной программной системы.
- Результаты проведенных вычислительных экспериментов, подтверждающих работоспособность предлагаемого подхода к автоматическому построению электронного каталога.

Публикации и личный вклад автора. Основное содержание диссертации нашло отражение в 6 опубликованных научных работах, в том числе 1 статья в рецензируемом издании, рекомендованном ВАК РФ. Также, принята в печать научная статья "Распределенная реализация построения индекса поискового каталога" в №1

(2011 г.) журнала Вестник ННГУ им. Н.И. Лобачевского, входящего в список ВАК. Результаты совместных научных работ [1,2,4,6], использованные в диссертационной работе, принадлежат лично автору диссертации.

Структура и объем работы. Работа состоит из введения, трех глав, заключения, списка литературы. Общий объем работы составляет 115 страниц. Список литературы составляет 68 наименований. Основные результаты излагаются в главах 2 и 3.

Краткое содержание работы

1. Введение

Во **введении** обосновывается актуальность задачи создания новых моделей информационного поиска по ключевым словам с тематическим ранжированием результатов поиска, на основе автоматически построенного электронного каталога, сформулированы цели и задачи исследования. Приводится краткий обзор содержания диссертации.

2. Общая характеристика проблемы тематического ранжирования, на основе автоматически построенного электронного каталога текстовых документов

В **первой главе** производится описание предлагаемой модели поиска по ключевым словам с тематическим ранжированием (раздел 1.1), которое основано на использовании автоматически построенного электронного каталога коллекции текстовых документов. В разделе 1.2 первой главы производится описание предлагаемой модели автоматического построения электронного каталога. Построение электронного каталога производится автоматически с использованием предложенного в работе алгоритма иерархической кластеризации по областям текстовых документов.

Задача информационного поиска по ключевым словам заключается в автоматическом поиске документов, содержащих заданные ключевые слова, в текстовой коллекции. Ключевое слово – это слово в тексте, способное в совокупности с другими ключевыми словами представлять текст. Рассмотрим коллекцию текстовых документов $D = \{D_1, \dots, D_N\}$ на некотором естественном языке. В целях задачи поиска каждый из документов представляется в виде его информационно-поискового образа - вектора ключевых слов $\tilde{D}_i = \{t_1,\dots,t_V\}$. Выделение ключевых слов представляет собой процедуру выборки наиболее значимых слов документа. Множество векторов ключевых слов документов представляет пространство R^V. На рис. 1 изображена матрица отображения

множества документов в пространство признаков - ключевых слов. Элементами матрицы являются веса ключевых слов по отношению к рассматриваемому документу.

	$\mathbf{t_1}$	$\mathbf{t_2}$	…	$\mathbf{t_V}$
$\mathbf{D_1}$	w_{11}	w_{12}	…	w_{1V}
$\mathbf{D_2}$	w_{21}	w_{22}	…	w_{2V}
…	…	…	w_{ij}	…
$\mathbf{D_N}$	w_{N1}	w_{N2}	…	w_{NV}

Рис.1. Отображение множества документов в пространство признаков.

Пусть имеется запрос $q=\{t_{j1},\ldots, t_{jL}\} \in R^V$, характеризующий некоторую информационную потребность. Тогда целью задачи поиска по ключевым словам является построение ранжированного по заданной функции ранжирования F списка документов $D_{i1}, \ldots, D_{iK}$ соответствующих данному запросу q. Функция ранжирования должна удовлетворять свойству упорядоченности, то есть для пары документов D_i, D_j выполняется $F(D_i) >= F(D_j)$, если документ D_i в большей степени соответствует запросу q, чем документ D_j. Наиболее известным примером функции ранжирования является Okapi BM25[1] (функция базового ранжирования), которая ранжирует документы в зависимости от их веса, вычисленного на основе встречаемости слов запроса в каждом из рассматриваемых документов (без учёта взаимоотношений между словами).

Модель поиска с тематическим ранжированием

Тематическое ранжирование результатов поиска по запросу из ключевых слов заключается в построении функции ранжирования зависящей от имеющихся тематических групп в заданной коллекции текстовых документов. В рассматриваемой текстовой коллекции предполагается наличие множества тематических групп документов $\exists\ K = \{K_1, \ldots, K_L\}$, которые возможно выделить.

В данной работе предлагается ранжировать документ выше, если он соответствует тематике поискового запроса. Под тематикой поискового запроса понимается превалирующая тематика K_b (которая может быть явно задана, или определяться по формуле ниже) из множества тематических групп K_q, документы которых присутствуют в результатах Q_r поискового запроса: $K_q=\{K_{i1},\ldots, K_{iM}\} \subset K$, где $\forall c \in \{i_1,\ldots,i_M\}, \exists D_i \subset Q_r, i \in [1,|Q_r|] : D_i \in K_c$,

$$K_b(q) = \max_c(\max_i(\text{Вес\_базовой\_функции\_ранжирования}(D_i)) : D_i \in K_c, K_c \subset K_q).$$

[1] Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. Okapi at TREC-3. In Proceedings of the Third Text Retrieval Conference (TREC 1994). Gaithersburg, USA, 1994.

В качестве тематической функции ранжирования в рамках данной работы были разработаны и предлагаются формулы (2) для подсчета веса документа D из результатов поискового запроса, где baseWeight(D) – вес документа, выдаваемый базовым алгоритмом ранжирования (например, алгоритмом BM25); Top – количество выдаваемых результатов на поисковой странице, Total – количество всех поисковых результатов, удовлетворяющих запросу q.

$$F(D) = Вес\_документа(D, K_b) = baseWeight(D) + categoryInfluence(D, K_b);$$
$$\deg\_cat(D) = baseWeight(D) * (1 - (\max Weight - baseWeight(D)));$$
$$\deg\_ncat(D) = (\max Weight - baseWeight(D)) * (1 - baseWeight(D));$$
$$categoryInfluence(D, K_b) = \begin{cases} \frac{2^{(\deg\_cat(D))} - 1}{norm}, если\_D \in K_b \\ -\frac{2^{(\deg\_ncat(D))} - 1}{norm}, если\_D \notin K_b \end{cases}; \quad (2)$$
$$norm = 2^{\max Weight} \sum_{i=1}^{Top} Log_2(1+i); \sum_{i=1}^{Total} baseWeight(D_i) = 1; \max Weight = \max_D(baseWeight(D));$$

Для тематического ранжирования полученного списка поисковых результатов требуется найти множество тематических групп заданной коллекции текстовых документов, поэтому рассматривается задача автоматического построения электронного каталога. Задача автоматического построения электронного каталога заключается в автоматическом построении иерархии электронного каталога, отражающей различные тематические области в исходной коллекции текстовых документов. В задачу также входит автоматическое формирование описания и названий рубрик электронного каталога, на основании содержащихся в них документах.

Для построения иерархии электронного каталога используются методы текстовой кластеризации. В частях 1.3-1.4 первой главы приводится постановка задачи текстовой кластеризации, и производится обзор и анализ существующих алгоритмов текстовой кластеризации. Входными данными задачи текстовой кластеризации являются образы текстовых документов $\tilde{D} = \{\tilde{D}_1, \ldots, \tilde{D}_N\}$, каждый из которых представлен многомерным вектором в пространстве признаков. В рамках задачи текстовой кластеризации в исследуемой коллекции текстовых документов предполагается наличие множества тематических групп (другими словами - кластеров) $K = \{K_1, \ldots, K_L\}$, которые возможно выделить. Выделение тематических групп (текстовая кластеризация) заключается в поиске неизвестного множества групп K схожих документов, в соответствии с некоторой мерой близости между объектами кластеризации. Наиболее часто применяемой

мерой близости является косинусная мера близости между векторами документов (3):

$$sim(\tilde{D}_i, \tilde{D}_j) = \cos(<(\tilde{D}_i, \tilde{D}_j)) = \frac{\sum_{l=0}^{V} \tilde{t}_{li} * \tilde{t}_{lj}}{\sqrt{\sum_{l=0}^{V} (\tilde{t}_{li})^2} * \sqrt{\sum_{l=0}^{V} (\tilde{t}_{lj})^2}} \qquad (3)$$

В частях 1.5-1.6 первой главы производится обзор общепринятых методов оценки результатов текстовой кластеризации и методов оценки полученного ранжирования поисковых результатов.

Для оценки методов кластеризации в данной работе решено использовать общепринятую меру оценки качества - стандартную F1-меру[2]. F1-мера (4) оценивает, насколько разбиение построенное алгоритмом кластеризации совпадает с “эталонным” разбиением, подготовленным экспертом:

$$F1 = \frac{2 * Точность * Полнота}{Точность + Полнота} \qquad (4)$$

Целью является максимизация F1 меры.

Для измерения качества работы алгоритма тематического ранжирования в работе используется общепринятая метрика NDCG@L_q [3], описываемая формулой (5):

$$NDCG@L_q = \frac{100}{Z} \sum_{r=1}^{L_q} \frac{2^{l(r)} - 1}{Log(1+r)}; \quad \text{NDCG@L} = \frac{1}{Q_T} \sum_{q=1}^{Q_T} NDCG@L_q \qquad (5)$$

Так же как и для F1-меры, метрика NDCG дает оценку ранжирования, подготовленного алгоритмом ранжирования для заданного запроса q, по сравнению с “эталонным” ранжированием, подготовленным экспертом. В качестве итоговой оценки алгоритма ранжирования подсчитывается среднее значения метрики NDCG@L (5), для всей коллекции подготовленных экспертом запросов (размером Q_T).

Качество формирования рубрик электронного каталога напрямую зависит от алгоритмов формирования образов текстовых документов. В частях 1.7-1.10 первой главы диссертации производится постановка задачи формирования образов текстовых документов. Производится обзор существующих методов формирования информационных образов текстовых документов, их анализ и выбор оптимального подхода.

[2] B. Stein, S. Meyer zu Eissen, F. Wißbrock. On Cluster Validity and the Information Need of Users. In Proc. 3rd IASTED International Conference on Artificial Intelligence and Applications (AIA'03), 2003. p. 404-413.

[3] K. Jarvelin and J. Kekalainen. IR evaluation methods for retrieving highly relevant documents. In SIGIR, 2000, p. 41-48.

3. Разработка тематического ранжирования на основании автоматического построения электронного каталога текстовых документов

Во **второй главе** приводится описание разработанных автором алгоритмов построения образов текстовых документов, алгоритмов автоматического построения электронного каталога, а также, алгоритмов тематического ранжирования, основанных на автоматически построенном электронном каталоге.

В целях поиска и кластеризации каждый документ представляется в виде информационно поискового образа - вектора ключевых слов в пространстве R^V. Зачастую количество слов документа может достигать несколько тысяч. Для повышения производительности и качества работы алгоритма кластеризации, из документа выбираются наиболее значимые ключевые слова. В этих целях применяются алгоритмы понижения размерности пространства признаков. Для понижения размерности пространства ключевых слов используются следующие приемы: выделение основ слов (различные морфологические формы слова отображаются в одну координату пространства признаков), удаление стоп слов (исключение из исследуемого текста некоторых слов, которые не несут самостоятельной смысловой нагрузки, являются «шумом» (например, предлоги, союзы)), принудительная редукция пространства признаков с применением порогов к частоте встречаемости слова в коллекции документов (понятно, что слова, встречающиеся в только одном документе, или в большей части коллекции не смогут иметь качественное влияние на процесс разделения документов на группы).

Для оценки важности рассматриваемой основы слова по отношению к документу в данной работе применялись формулы TF-IDF (6). Выделяют понятия TF_i (term frequency) – частоты основы t_i в документе D, а также IDF (inversed document frequency) – обратная частота встречаемости основы t_i в документах коллекции, N – общее количество документов коллекции, DN_i – количество документов, содержащее основу t_i.

$$IDF(t_i) = \log\left(1+\frac{N}{DN(t_i)}\right); \qquad Bec_D(t_i) = \frac{TF(t_i)*IDF(t_i)}{\sqrt{\sum_j \left(TF(t_j)*IDF(t_j)\right)^2}} \qquad (6)$$

Для автоматического построения иерархии электронного каталога применяется предложенный в работе метод иерархической кластеризации по областям текстовых документов. Алгоритм иерархической кластеризации по областям строит

иерархическое дерево областей, которое состоит из документов текстовой коллекции. Узлы строящегося дерева назовем областями. Итоговыми кластерами являются узлы дерева областей. Узлы дерева областей содержат в себе документы наиболее близкие друг к другу. При этом иерархия дерева отражает отношения соподчиненности между областями. Структура двухуровневого дерева областей отображена на рис.2.

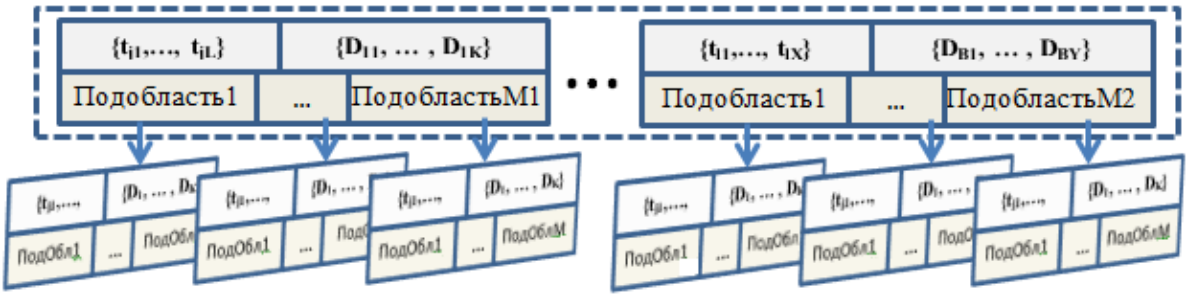

Рис. 2. Структура дерева областей.

Функционирование алгоритма иерархической кластеризации по областям можно разбить на два этапа:

1) Инициализация алгоритма иерархической кластеризации по областям.

В качестве исходных данных на данном этапе алгоритма выступают документы $D=\{D_1,\ldots,D_N\}$, представленные множеством ключевых слов $D_i=\{t_1,\ldots, t_V\}$. Первоначально все поступающие документы попадают в специальную область дерева, называемую корзиной, до тех пор, пока количество документов не превысит заранее заданный предел. При превышении предела область-корзина разбивается на подобласти (используется алгоритм K-средних), которые присоединяются к верхнему уровню дерева.

2) Этап обработки входящего потока документов.

В качестве исходных данных на данном этапе алгоритма выступают документ, представленный множеством ключевых слов, и инициализированное дерево областей. На первом шаге второго этапа алгоритма происходит проверка возможности правильной вставки поступившего документа в дерево областей. Возможность вставки определяется измерением близости между документом и областями первого уровня. Если близость не превышает динамически установленный предел, который определяется как минимум близости между уже обработанными документами, то документ сохраняется в области-корзине и временно не может быть встроен в дерево. Если же документ имеет близость,

превышающую установленный предел, то он направляется по дереву к самой близкой подобласти. На следующих шагах алгоритма документ спускается по дереву до тех пор, пока не встретится наиболее близкая к документу область. Документ помещается в найденную область. При превышении размера области определенного ограничения, происходит разбиение ее на подобласти (при этом используется алгоритм К-средних). Если количество подобластей превзошло определенное ограничение, то выполняется операция интеграции подобластей. Операция интеграции подобластей состоит из двух основных операций: поиск и объединение под единым началом наиболее близких друг к другу подобластей.

Для применения предложенных выше последовательных алгоритмов для формирования электронного каталога больших коллекций текстовых документов были разработаны их **параллельные версии**, которые описаны в **частях 2.6-2.8 данной работы**. В качестве технологии параллельного (распределенного) программирования было предложено использовать программную модель MapReduce[4]. Концепция MapReduce состоит в том, что производимые вычисления разбиваются на функции Map и Reduce. Функция Map трансформирует входные данные в промежуточный список пар ключ/значение. Функция Reduce берет список пар ключ/значение, который генерирует Map и свертывает его по ключу (на выходе одна пара ключ/значение для каждого ключа). В части 2.7 второй главы данной работы представлено описание распределенной версии алгоритма подготовки образов текстовых документов. Распределенная версия алгоритма построения образов текстовых документов состоит из двух шагов.

Входные данные Документы в текстовом формате.

Map Построить вектор частот встречаемости основ слов документа.

Reduce Из построенного на шаге Map вектора основ слов - рассчитать вес каждой основы слова, отбросить стоп слова и слова, частота которых в коллекции не входит в заданные пределы. В качестве результата (образ документа) построить вектор основ слов, имеющих наибольший вес.

Рис.3. Параллельная реализация подготовки информационных образов текстовых документов.

На первом шаге происходит построение вектора DF(встречаемости слов в документах коллекции), который состоит из записей <основа слова, количество документов, содержащее данную основу>: в функции Map производится подсчет уникальных вхождений основ слов в рассматриваемый документ, в функции

[4] Jeffrey Dean, Sanjay Ghemawat. MapReduce: simplified data processing on large clusters // Communications of the ACM. 2008. V. 51 P. 107-113.

Reduce аккумулируются результаты подсчетов для каждой основы слова по всей коллекции документов. На втором шаге, происходит формирование образов текстовых документов (рис.3).

В части 2.8 второй главы диссертации представлена реализация параллельной версии алгоритма иерархической кластеризации по областям текстовых документов, которая позволяет масштабировать решение задачи на необходимое количество вычислительных узлов. Работа распределенной версии алгоритма представлена в виде двух шагов. На первом шаге алгоритма производится подготовка дерева областей верхнего уровня (рис.4).

Входные данные Образы документов в виде векторов ключевых слов.

Map Построить дерево областей из входящего потока документов последовательным алгоритмом кластеризации, в результат выдать вектор областей верхнего уровня (с одинаковым ключом для всех процессов Map; каждая область характеризуется вектором ключевых слов).

Reduce Произвести кластеризацию областей верхнего уровня полученных на шаге Map последовательным алгоритмом кластеризации. В результат выдать дерево областей верхнего уровня.

Рис. 4. Формирование дерева областей верхнего уровня.

На втором шаге алгоритма, запускается распределенная кластеризация документов, которая работает следующим образом: при поступлении на обработку в функцию Map для каждого документа определяется область дерева верхнего уровня, далее он отправляется на вычислительный узел, ответственный за обработку документов данной области, и на этом вычислительном узле последовательным алгоритмом иерархической кластеризации по областям строится поддерево областей (рис. 5).

Входные данные Образы документов в виде векторов ключевых слов, дерево верхнего уровня.

Map Для каждого входного образа документа определить область дерева верхнего уровня. Выходные пары (ключ, значение) = (идентификатор области верхнего уровня, образ документа). Внутренними средствами Hadoop выходные пары Map группируются по ключу и далее передаются на вход Reduce.

Reduce Для всех документов отображенных на шаге Map в подобласть дерева верхнего уровня – построить поддерево областей.

Рис.5. Параллельная реализация иерархической кластеризации по областям.

Тематическое ранжирование списка результатов поиска

В части 2.9 второй главы диссертации представлено описание алгоритмов тематического ранжирования, основанных на автоматически построенном электронном каталоге. Поиск по ключевым словам с тематическим ранжированием представлен в виде двух этапов:

1) Выборка из индекса документов, содержащих заданные ключевые слова, и определение веса каждого из результатов поиска на основе базовой ранжирующей функции (BM25), для каждого извлеченного документа определена тематическая группа из автоматически построенного электронного каталога (тематическая информация является частью индекса).

2) Тематическое ранжирование списка результатов. Наиболее распространенным представлением результатов поиска является последовательный список страниц с заданным числом результатов на каждой странице (обычно 10) – в этом случае в работе предлагается производить ранжирование списка результатов в соответствии с тематикой наилучшего, по мнению пользователя, документа. Пользователь поисковой системы обозначает наилучший результат, и поисковая система ранжирует результаты поиска по формулам (2). При отсутствии информации от пользователя о превалирующей тематической группе в данной работе было предложено как альтернативное экспериментальное представление - группировать результаты поиска в тематические группы на каждой из страниц результатов, при этом список тематических групп сортируется по весу документа, входящего в группу, имеющего наибольший вес (базового алгоритма ранжирования).

4. Программная реализация системы поиска с тематическим ранжированием, на основе автоматически построенного электронного каталога, результаты проведенных испытаний

В **третьей главе** приводится описание архитектуры разработанной программной системы, реализующей модели и методы поиска по ключевым словам с тематическим ранжированием, на основе электронного каталога заданных коллекций текстовых документов, автоматически построенного с использованием предлагаемого в работе алгоритма текстовой кластеризации. Приведено описание тестовых текстовых коллекций и результаты испытаний предлагаемого метода автоматического построения электронного каталога, исследование качества тематического ранжирования результатов поиска по ключевым словам.

Программная поисковая система по ключевым словам с тематическим ранжированием, на основе автоматически построенного электронного каталога, разработана с использованием технологий объектно-ориентированного анализа с использованием инструментов интегрированной среды разработки Eclipse и платформы для распределенных вычислений Apache Hadoop. Общая схема разработанной программной системы изображена на рис.6.

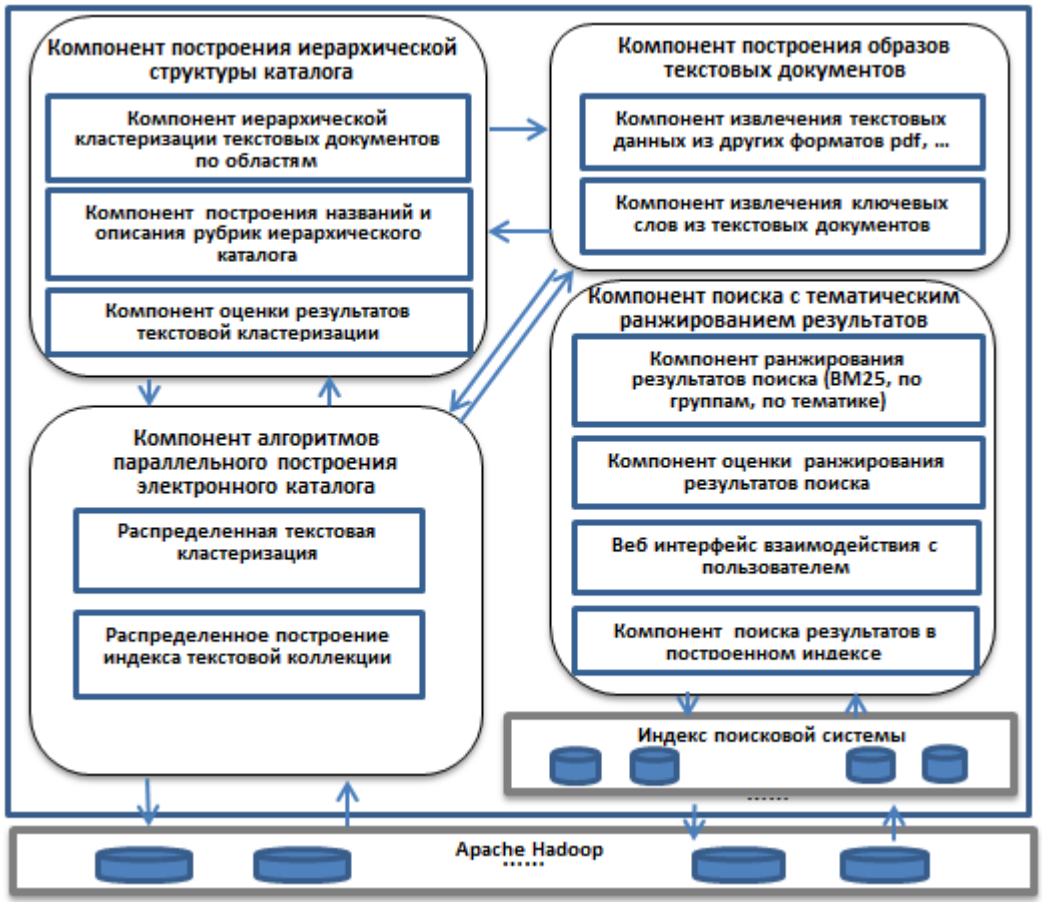

Рис. 6. Архитектура разработанной программной системы поиска по ключевым словам с тематическим ранжированием.

На разработанной системе была проведена экспериментальная проверка качества работы предложенной модели поиска с тематического ранжированием и разработанных алгоритмов автоматического построения электронного каталога. В качестве тестовых текстовых коллекций для проверки работоспособности алгоритма были выбраны две коллекции – русскоязычная и англоязычная: коллекция статей журнала “Вестник Нижегородского государственного университета им. Н.И. Лобачевского” (1972 русскоязычных текста, 1 ГБ данных), коллекция новостных статей“20NewsGroups” (20000 англоязычных текстов, 46 МБ данных).

Используя введенный критерий оценки качества кластеризации (F1-мера), были проведены численные эксперименты по измерению качества работы предлагаемого алгоритма и традиционных алгоритмов иерархической кластеризации, которые показали превосходство предлагаемого алгоритма иерархической кластеризации по областям, используемого для автоматического построения иерархии каталога, по сравнению с традиционными алгоритмами иерархической кластеризации. Результаты сравнения приведены на рис.7 и рис.8.

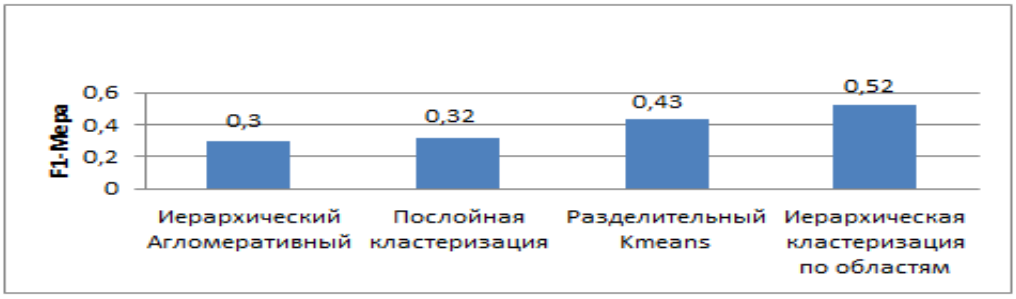

Рис. 7.Сравнение качества работы алгоритмов текстовой кластеризации по F1-мере качества кластеризации на тестовой коллекции “Вестник”.

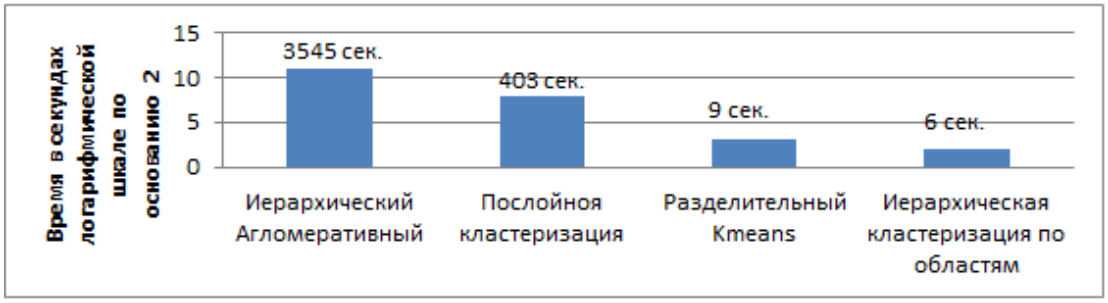

Рис.8. Сравнение времени работы последовательных алгоритмов текстовой кластеризации (в логарифмической шкале, по основанию 2).

Исследование параллельных версий предложенных алгоритмов

В результате применения парадигмы распределенного программирования MapReduce удалось существенно сократить время построения индекса коллекции текстовых документов и проведения кластеризации текстовых данных. На рис.10 приведены результаты вычислительных экспериментов для распределенной версии алгоритма кластеризации текстовой коллекции. Отметим, что время вычислений и эффект получаемый от количества используемых процессорных ядер зависит от объема данных, который необходимо обработать. При уменьшении объема данных (из-за накладных расходов связанных с распределенными вычислениями) наблюдается уменьшение эффекта ускорения получаемого от увеличения объема вычислительных ресурсов.

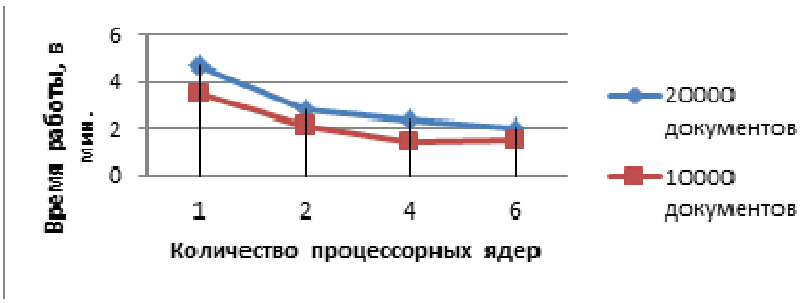

Рис.10. Зависимость времени работы иерархической кластеризации по областям от количества задействованных в вычислениях процессорных ядер.

Реализация поиска, как правило, осуществляется на информационных серверах. В настоящее время аппаратные средства в большинстве своем содержат от двух до шести ядер, поэтому проведенные вычислительные эксперименты позволяют утверждать о современности и эффективности применения предложенных распределенных версий алгоритмов.

Исследование качества тематического ранжирования

Используя введенный критерий оценки качества ранжирования (NDCG@10), были проведены численные эксперименты по измерению качества работы предлагаемых алгоритмов ранжирования на коллекции статей журнала “Вестник Нижегородского государственного университета им. Н.И. Лобачевского”, которые показали превосходство предлагаемого алгоритма ранжирования по сравнению с базовым алгоритмом Okapi BM25 (рис.11).

Рис.11. Качество работы алгоритмов ранжирования по метрике NDCG@10.

Отметим, что для оценки качества работы алгоритма ранжирования необходимо подготовить тестовое множество запросов и сортированные списки поисковых результатов, соответствующие различным алгоритмам ранжирования. Для каждого из результатов заданных списков эксперт выставляет оценку соответствия рассматриваемого документа по отношению к запросу. При выборе числа запросов тестового множества ориентируются в основном на критерий стоимости трудозатрат экспертов для его подготовки[5]. Также, предлагается оценить устойчивость оценки ранжирования (если при увеличении количества запросов тестового множества оценка ранжирования остается в заданных переделах, то далее предлагается не увеличивать мощность тестового множества запросов).

На рис.12 приведен график изменения метрики NDCG@10 в зависимости от количества запросов: при увеличении количества запросов до 30 наблюдается

[5] Труды РОМИП 2009. Российский семинар по оценке методов информационного поиска. Санкт-Петербург. 2009. 198 с.

стабилизация значения критерия качества NDCG@10.

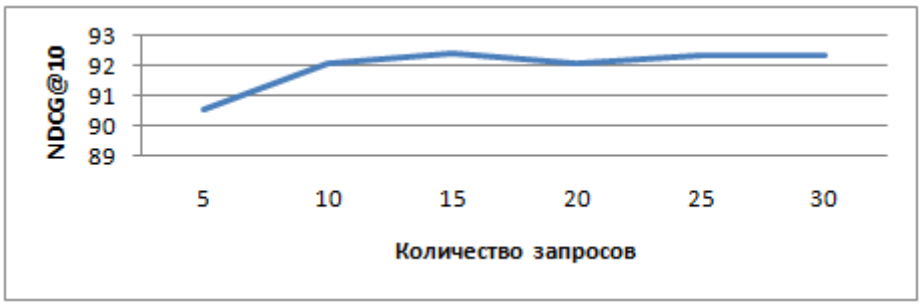

Рис.12. График изменения метрики NDCG@10 в зависимости от количества запросов.

В целом по итогам проведенных исследований предлагаемый способ тематического ранжирования, на основе автоматически построенного электронного каталога, показал результаты, превосходящие по качеству и скорости существующие подходы. Таким образом, предложенная модель поиска с тематическим ранжированием, на основе автоматически построенного электронного каталога удовлетворяет всем требованиям, вытекающим из цели исследования.

Основные результаты:

1)Разработана новая модель информационного поиска с тематическим ранжированием, основанным на автоматически построенном электронном каталоге. Предложенные методы тематического ранжирования показали улучшение качества ранжирования результатов поиска от 1% до 15% по метрике NDCG@10.

2) Разработана новая модель автоматического построения электронного каталога текстовых документов без ограничения на тематику и размер исходной текстовой коллекции.

3) Предложен новый метод текстовой кластеризации - иерархическая кластеризация по областям текстовых документов, учитывающий недостатки существующих алгоритмов иерархической текстовой кластеризации. Разработаны последовательные и параллельные версии предложенного алгоритма. Проведенные испытания показали преимущество в качестве предложенного алгоритма иерархической кластеризации по областям по сравнению с тремя традиционными алгоритмами кластеризации. Улучшение качества кластеризации составило от 9 % до 22%.

4) Разработаны параллельные версии алгоритмов извлечения текстовых

признаков и иерархической кластеризации по областям текстовых документов. Проведенные эксперименты показали линейное ускорение в зависимости от количества вычислительных узлов.

5) Разработан метод автоматического выбора названия и описания для сформированных кластеров автоматически построенного электронного каталога.

6) Разработан, апробирован и внедрен в качестве поисковой системы по публикациям программный комплекс, реализующий предложенную модель поиска по ключевым словам с тематическим ранжированием, основанным на использовании предложенного подхода к автоматическому построению электронного каталога.

Публикации в изданиях, рекомендованных ВАК РФ

1) Борисюк Ф.В. “Новый метод поиска на основе иерархической кластеризации по областям текстовых документов” / Борисюк Ф.В., Швецов В.И. // Вестник ННГУ им. Н.И. Лобачевского. 2009, № 4, с. 165–171.

Публикации в прочих изданиях

2) Борисюк Ф.В. “Иерархическая кластеризация по областям текстовых документов”. / Борисюк Ф.В., Швецов В.И. // Материалы всероссийской конференции Технологии Майкрософт в теории и практике программирования. 2009, с. 48–54.

3) Борисюк Ф.В. “Выделение ключевых слов в научной коллекции гипертекстовых документов”. / Борисюк Ф.В. // Сборник материалов всероссийской научной школы для молодежи “Управление информационными ресурсами образовательных, научных и производственных организаций”. Магнитогорск, Магнитогорский государственный университет. 2009, с. 31-32.

4) Борисюк Ф.В. “Параллельная реализация построения индекса поисковой системы с использованием платформы Hadoop”. / Борисюк Ф.В., Швецов В.И, Белоусова И.И. // 9-я международная конференция “Высокопроизводительные параллельные вычисления на кластерных системах”. Владимир, Владимирский государственный университет. 2009, с. 48-61.

5) Борисюк Ф.В. “Параллельная реализация иерархической кластеризации по областям текстовых документов”. / Борисюк Ф.В. // Материалы всероссийской конференции Технологии Майкрософт в теории и практике программирования. 2010, с. 38-40.

6) Borisyuk F.V. “Adaptation of Hierarchical clustering by areas for automatic construction of electronic catalogue”/ Borisyuk F.V., Shvetsov V.I. // Proceedings of SYRCoSE, 2010, p. 141.-145.