На правах рукописи

ПЕСКОВА ОЛЬГА ВАДИМОВНА

РАЗРАБОТКА МЕТОДА АВТОМАТИЧЕСКОГО ФОРМИРОВАНИЯ РУБРИКАТОРА ПОЛНОТЕКСТОВЫХ ДОКУМЕНТОВ

Специальность 05.13.17 – Теоретические основы информатики

АВТОРЕФЕРАТ

диссертации на соискание ученой степени

кандидата технических наук

Москва – 2008

Работа выполнена в Московском государственном техническом университете им. Н.Э. Баумана.

Научный руководитель:	доктор технических наук, профессор Трусов Борис Георгиевич
Официальные оппоненты:	доктор технических наук, профессор Эрлих Александр Игоревич кандидат технических наук, доцент Цицин Алексей Георгиевич
Ведущая организация:	Всероссийский Институт Научной и Технической Информации

Защита диссертации состоится «___» _____ 200__ года на заседании диссертационного совета Д 212.141.10 в Московском государственном техническом университете им. Н.Э. Баумана по адресу 105005, Москва, 2-я Бауманская ул., д. 5.

С диссертацией можно ознакомиться в библиотеке Московского государственного технического университета им. Н.Э. Баумана.

Автореферат разослан « » 200__ г.

Ученый секретарь
диссертационного совета Д 212.141.10
к.т.н., доцент С.Р. Иванов

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы.

В связи с наблюдаемым на протяжении последних десятилетий стремительным ростом накапливаемых объёмов электронных документов особое значение приобретает разработка программных средств поиска информации. В настоящей работе рассматриваются коллекции полнотекстовых документов, т.е. множества документов, содержащие тексты на естественном языке, и доступные через средства телекоммуникации для поиска и доставки пользователю. Примерами могут служить фонды электронных библиотек, электронные архивы журнальных статей, различные собрания научно-технических материалов в локальных или глобальных сетях и другие.

Традиционными подходами к решению проблемы поиска информации в коллекциях полнотекстовых документов, являются: поиск по ключевым словам (а) и классификационный поиск (б). В настоящей работе наиболее перспективным считается второй подход. Эффективность поиска по ключевым словам существенно зависит от удачного описания информационных потребностей в форме запроса на естественном языке. В противоположность этому классификационный поиск благодаря интуитивно понятному навигационному интерфейсу позволяет легко формулировать и уточнять информационные потребности, что повышает эффективность и удобство поиска документов. Однако, традиционные механизмы классификационного поиска – универсальные библиотечные классификаторы (УДК, ГРНТИ, ББК) и специализированные предметные рубрикаторы, имеющие фиксированную структуру, не успевают изменяться вслед за темпом развития науки и техники или требуют высоких затрат как на адаптацию классификаторов, так и на классификацию по ним документов.

Современные методы классификационного поиска основаны на механизме автоматической классификации текстов. Данный подход, как правило, подразумевает применение методов категоризации, которые распределяют документы по предопределенному набору рубрик на основе знания, полученного из обучающего множества. Разработке и тестированию алгоритмов данного вида, а также связанным с ними алгоритмам представления текстов посвящены труды таких авторов как М. С. Агеев, И. Е. Кураленок, И. С. Некрестьянов , В. И. Шабанов, T. Joachims, D. D. Lewis, R. E. Schapire, H. Schutze, F. Sebastiani, Y.Yang, I. Dagan, S. T. Dumais и ряда других. Однако данный подход решает не все проблемы традиционного классификационного поиска: вопрос регулярной актуализации предметных областей классификаторов и связанных с ней затрат решается путём

повторного обучения систем категоризации, что в свою очередь приводит к высоким экспертным затратам на анализ ситуации и подготовку новых обучающих данных.

Таким образом, в настоящее время существует потребность в разработке методов классификации, которые способны на основе анализа текстов и внутренних связей между ними автоматически строить рубрикаторы коллекций полнотекстовых документов. Среди известных методов автоматического анализа текстовых данных потенциально способных решить представленную проблему следует выделить методы кластеризации, которые автоматически разбивают документы на группы (кластеры) на основе анализа тематической близости между ними. Разработке алгоритмов данного вида и способов оценки качества получаемого разбиения документов, а также связанным с ними алгоритмам представления текстов посвящены труды таких авторов как Д. В. Ландэ, М. В. Киселев, К. М. Кириченко, C. J. van Rijsbergen, G. Salton, D. Manning, H. Schutze, T. Kohonen, O. Eli Zamir, J. C. Bezdek, M. Halkidi и ряда других. Однако в большинстве работ разбиение документов на тематические группы рассматривается как промежуточный этап при формировании некоторого представления о составе анализируемых текстовых данных, не ставится задача формирования рубрикатора политематических коллекций в виде, близком к традиционному, способном служить механизмом классификационного поиска для конечного пользователя. Более того, требование применимости метода классификации к различным политематическим коллекциям вызывает необходимость разработки подхода к формированию представления документов при условии отсутствия специализированной априорной информации.

Таким образом, актуальность разработки метода автоматического формирования рубрикатора коллекции полнотекстовых документов, основанного на анализе тематической близости текстов документов, следует из недостаточной эффективности традиционных поисково-навигационных средств электронных библиотек и трудоёмкости обновления рубрикаторов вследствие динамичного развития областей научно-технического знания. Задача автоматического построения рубрикаторов актуальна как для полных коллекций документов, так и для их подмножеств, например, полученных в результате поиска по ключевым словам, что позволит пользователю оставаться в пределах интересующей его предметной области.

Целью диссертационной работы является создание метода автоматического формирования рубрикатора коллекции полнотекстовых документов, основанного на результатах кластеризации.

Для достижения этой цели в диссертации решены следующие **задачи**:

- выполнено обобщение известных методов и алгоритмов автоматической классификации полнотекстовых документов и создан модифицированный алгоритм послойной кластеризации, основанный на выделении компонент связности подграфов графа близости документов;
- разработан алгоритм формирования информационно-поисковых образов документов, включающий механизм редукции признаков, основанный на предложенном подходе к оценке тематической значимости признаков документов;
- создан программный комплекс для автоматического формирования рубрикатора коллекции полнотекстовых документов и его отображения в доступном для читателя виде с целью навигации по данной коллекции документов;
- с помощью программного комплекса выполнена оценка значений параметров разработанных алгоритмов и проверена работоспособность предложенного метода формирования рубрикатора.

Методы исследования.

При решении поставленных задач в данной работе использован математический аппарат теории множеств, теории графов, методы математической статистики, кластерного анализа и методы построения интеллектуальных систем и программных интерфейсов.

При разработке программного обеспечения применялись методы объектно-ориентированного программирования с использованием сред разработки Microsoft Visual Studio .NET 2003 и СУБД Microsoft SQL Server 2000.

Научная новизна.

В результате выполнения работы получены новые научные результаты:

- предложен новый метод автоматического формирования рубрикатора коллекции полнотекстовых документов, применимый для произвольных массивов научно-технических документов без ограничений на их объём и тематику, в условиях отсутствия специализированной априорной информации для формализации их содержания;
- разработана модификация алгоритма кластеризации документов, позволяющая автоматически разбивать тексты на естественном языке на тематические группы с возможностью простого управления глубиной и уровнем детализации иерархии этих групп;
- предложен подход к оценке тематической близости документов с использованием метода редукции пространства признаков, составляющих

информационно-поисковые образы, что позволило повысить качество и скорость выполнения кластеризации множества текстов.

На защиту выносятся:

- метод автоматического формирования многоуровневого рубрикатора политематической коллекции полнотекстовых документов, представленного в пригодном для пользователя виде;
- модифицированный алгоритм автоматической послойной кластеризации полнотекстовых документов, являющийся основой для формирования рубрикатора коллекции;
- алгоритм редукции пространства признаков документов для формирования информационно-поисковых образов документов;
- результаты экспериментальных исследований с помощью разработанного программного обеспечения, подтверждающие работоспособность предлагаемого метода автоматического формирования рубрикатора документов.

Практическая значимость.

Разработанный в диссертации метод и программная система предназначены для использования в электронных библиотеках в качестве элемента их поисковых систем. Предложенный подход к автоматической классификации документов позволяет решать проблему навигации как по всей коллекции документов, так и по её подмножествам, динамически формируя для каждого случая наиболее подходящий предметный рубрикатор, отражающий иерархические и родственные связи между областями знаний и обладающий автоматически получаемыми вербальными описаниями этих областей знаний. Такой элемент поисковой системы способен выполнять функции как самостоятельного поискового аппарата, так и служить средством повышения качества работы других поисковых механизмов.

Апробация и внедрение результатов работы.

Разработанный программный комплекс внедрен и используется в рамках единой Автоматизированной Библиотечной Информационной Системы МГТУ им. Н.Э. Баумана. Предложенные методы и алгоритмы применяются в подсистеме поддержки фонда электронных документов.

Основные результаты работы докладывались и обсуждались на Всероссийских конференциях студентов, аспирантов и молодых ученых «Технологии Microsoft в теории и практике программирования» (Москва, 2005 и 2006 гг.), 14-ой Международной конференции «Крым 2007: библиотеки и информационные ресурсы в современном мире науки,

культуры, образования и бизнеса» (Судак, 2007 г.), 7-ой Международной конференции «НТИ-2007: информационное общество, интеллектуальная обработка информации, информационные технологии» (Москва, 2007 г.).

Публикации по теме диссертации. По теме диссертации опубликовано 9 печатных работ (в том числе одна статья в журнале, входящем в перечень ведущих рецензируемых научных журналов и изданий) и получено 2 свидетельства об официальной регистрации программы для ЭВМ.

Структура и объем диссертации. Диссертация состоит из введения, четырех глав, выводов и списка литературы из 132 наименований. Диссертация изложена на 150 страницах, содержит 46 рисунков и 12 таблиц.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во **введении** обоснована актуальность проблемы создания методов и средств классификации полнотекстовых документов в электронных хранилищах, сформулирована цель исследования и разработки метода автоматического формирования рубрикатора коллекции полнотекстовых документов в условиях отсутствия априорных сведений о предметных областях документов.

В **первой главе** выполнена постановка задачи автоматического построения рубрикатора полнотекстовых документов, в основу решения которой положен алгоритм автоматической классификации текстов на естественном языке. Анализ известных подходов к автоматической классификации позволил выбрать подход, основанный на кластерном анализе. Алгоритмы кластерного анализа применяются при отсутствии предопределённого рубрикатора и документов-образцов и формируют группы (кластеры) документов на основе автоматического анализа тематической близости между ними. Количественная оценка качества кластеризации строится на основе анализа внешних и внутренних мер качества. Внешние меры основаны на сравнении автоматического разбиения данных с полученным от экспертов «эталонным» разбиением этих же данных. Внутренние меры основаны на оценке свойств отделимости и компактности полученного разбиения данных.

Входными данными в задаче кластеризации являются информационно-поисковые образы документов, которые представляют собой многомерные векторы в пространстве признаков документов и характеризуют смысловое содержание исходных документов. В процессе кластеризации сходство документов вычисляется как геометрическая близость векторов этих

документов в пространстве признаков. Формирование образов документов тесно связано с решением общей проблемы автоматической обработки текстов, обусловленной неоднозначностями естественного языка. В работе проведёно обобщение алгоритмов формирования образов документов, показавшее, что наиболее приемлемым является подход, использующий в качестве смысловых признаков одиночные слова из текстов, прошедшие морфологический анализ и оценку их значимости, основанную на частоте встречаемости слов в текстах.

Выходными данными задачи кластеризации является набор кластеров документов. Выбор алгоритма кластеризации обусловлен его видом, поскольку производимый набор кластеров служит основой для формирования рубрикатора. Рубрикатор предложено отображать в виде графа тематических групп (кластеров) документов, рёбра которого отражают как иерархические, так и родственные связи между ними. Глубина иерархии кластеров не более 2-3 уровней. Кластерам присваивается автоматически полученное вербальное описание, состоящее из краткого названия и списка ключевых слов. Такой способ организации рубрикатора документов выбран на основе опыта реализации и эксплуатации подсистемы систематизации АБИС МГТУ им. Н. Э. Баумана.

Анализ таких известных алгоритмов кластеризации, как: иерархические (основанные на правилах ближайшего соседа, наиболее удалённых соседей и попарного среднего, суффиксные деревья), квадратичной ошибки (алгоритм *k*-средних, нечёткий алгоритм *c*-средних), алгоритмы теории графов (алгоритм минимального остовного дерева, алгоритм послойной кластеризации), вероятностные (алгоритм максимального ожидания), основанные на концепции плотности (алгоритм *DBSCAN*), нейросетевые (самоорганизующиеся карты Кохонена и алгоритмы теории адаптивного резонанса) и эволюционные (генетические) – позволил выбрать кластеризационный подход, основанный на алгоритме послойной кластеризации, подвергнутый модификации.

Во **второй главе** описан предлагаемый подход к автоматическому формированию рубрикатора документов. Основные этапы *автоматического формирования рубрикатора коллекции полнотекстовых документов* включают (рис. 1):

- формирование информационно-поисковых образов полных текстов документов;
- формирование множества кластеров информационно-поисковых образов документов, содержащего иерархические связи между его элементами;

• преобразование множества кластеров в рубрикатор коллекции документов, соответствующий предлагаемому способу представления:

– выявление родственных связей между кластерами одного и того же уровня путём вычисления мер близости между их представителями; добавление при необходимости «родственного» ребра в граф рубрикатора.

– формирование вербального описания вершин графа рубрикатора, состоящего из краткого названия, или непосредственной подписи вершины кластера на графе рубрикатора, и списка ключевых слов, детально характеризующих тематическую направленность документов данного кластера.

Рис. 1. Функциональная схема формирования рубрикатора коллекции полнотекстовых документов

Формирование информационно-поисковых образов документов начинается с построения пространства признаков документов путём выявления признаков, т. е. псевдооснов слов всей коллекции документов. Слова, имеющие одинаковую псевдооснову, в дальнейшем считаются эквивалентными по смысловому значению. Метод отображения документов в пространство их признаков основан на взвешивании каждого признака для каждого документа по схеме, учитывающей как частоту признака в документе, так и обратную документную частоту каждого признака. В результате образы документов имеют вид N_P-мерных векторов признаков, где N_P – число элементов множества признаков всей коллекции документов P.

Существенными характеристиками пространства признаков текстов

являются их высокая размерность (N_P может достигать десятков тысяч даже для не больших коллекций текстов) и связанная с ней недостаточно выразительная ориентация векторов-образов в пространстве признаков, что приводит не только к высоким вычислительным затратам, но и к низкому качеству разбиения на кластеры. Анализ такой ситуации стал обоснованием необходимости разработки *алгоритма редукции исходного пространства признаков*.

В настоящей работе, во-первых, применена техника принудительной редукции пространства признаков, т. е. принудительно удалены из всех документов все те признаки, частоты и веса которых не соответствовали заданным порогам, без индивидуального подхода к оценке значимости признаков в различных документах. Данная техника редукции направлена на сокращение высокой размерности пространства признаков. Во-вторых, в работе предложен алгоритм избирательной редукции признаков, целью которой является повышение качества представления тематики отдельных документов их признаками. В основу данного алгоритма положен тот факт, что один и тот же признак может являться значимым для одной предметной области и не являться таковым для другой, но при этом иметь достаточно высокую частоту встречаемости в документах обеих областей. Следовательно, поиск и сокращение невыразительных признаков должно выполняться не для всей коллекции сразу (как в случае с принудительной редукцией), а для каждой группы тематически родственных документов в отдельности. Таким образом, алгоритм избирательной редукции заключается в группировке признаков документов в подпространствах документов, которые предположительно считаются тематически родственными, и принятии решения об удалении признаков в рамках каждой отдельной группы.

Для *формирования множества кластеров информационно-поисковых образов документов* разработан модифицированный алгоритм послойной кластеризации. Суть алгоритма заключается в представлении исходной информации о документах в виде графа близости $G = (V, E)$, вершины которого соответствуют документам. Рёбра, соединяющие вершины v_i и v_j, имеют длину равную значению меры близости между образами *i*-ого и *j*-ого документов $Sim(\vec{d}_i, \vec{d}_j)$ ($0 \le Sim(\vec{d}_i, \vec{d}_j) \le 1$). Тогда при экспериментально подобранной входной последовательности пороговых значений мер близости между документами $1 = \tau_0^{sim} > \tau_1^{sim} > \ldots > \tau_{m+1}^{sim} = 0$ алгоритм кластеризации определяет последовательность подграфов графа близости $G^0 \subseteq G^1 \subseteq \ldots \subseteq G^m \subseteq G^{m+1}$, где $G^t = (V, E^t)$ и $E^t = \left\{ e_{ij} \in E : Sim(\vec{d}_i, \vec{d}_j) \le \tau_t^{sim} \right\}$,

$G^0 = (V, \varnothing)$ и $G^{m+1} = G$. Алгоритм послойной кластеризации выделяет компоненты связности подграфа G^t, получая таким образом разбиение коллекции документов $\left(C_1^t, ..., C_{k_t}^t\right)$, называемое *кластеризацией на уровне* τ_t^{sim}. В результате на выходе алгоритма получается m вложенных разбиений $C^1 \subset ... \subset C^m$, или слоёв, которые отражают иерархические связи между кластерами документов. Модификация алгоритма послойной кластеризации выполнена с целью уменьшения влияния на результат кластеризации «узких перемычек» между кластерами и заключается в замене кластеров, полученных на предыдущих уровнях, их центроидами (средними элементами кластеров) при выявлении компонент связности на последующих уровнях.

Предлагаемый *способ оценки кластеризации коллекции документов* основан на традиционном подходе – вычислении внешних и внутренних мер качества разбиения данных, а также на сравнении временных затрат алгоритма кластеризации. Для обобщения результата оценки предложен следующий обобщающий показатель эффективности алгоритма кластеризации F:

$$F = F_{exter} + F_{inter} + \frac{N_D}{t}, \quad (1)$$

где F_{exter} – обобщённый внешний критерий качества кластеризации; F_{inter} – обобщённый внутренний критерий качества кластеризации; t – время выполнения алгоритма кластеризации (без вычисления матрицы близости документов); $\frac{N_D}{t}$ – условная величина, показывающая количество документов, классифицируемых за секунду.

$$F_{exter} = F_1\text{-мера} - E, \quad (2)$$

где F_1-мера – это объединяющий показатель полноты и точности системы информационного поиска; E – погрешность классификации.

$$F_{inter} = |CPCC| + DI - DB + \frac{CH}{N_D} + I, \quad (3)$$

где $CPCC$ – кофенетический коэффициент корреляции; DI – индекс Дана; DB – мера Дейвиса-Булдина; CH – индекс Калинского и Гарабача; I – I-индекс.

Обобщённый показатель F предложен для сравнительной оценки эффективности различных алгоритмов кластеризации.

В **третьей главе** описана структура программного комплекса (рис. 2), реализующего алгоритм формирования образов документов, алгоритм

послойной кластеризации и метод формирования рубрикатора коллекции документов, а также структура базы данных, диаграмма состояний и компонентная модель программной системы.

Рис. 2. Структура программного комплекса

С помощью разработанной системы экспериментально исследован предлагаемый подход к кластеризации полнотекстовых документов. Тестовыми данными являлись русскоязычные документы он-лайн библиотеки по информационным технологиям CITFORUM (*http://www.citforum.ru*) от 23.02.2006, содержащей 1572 разнородных с точки зрения размера и содержательного уровня документов. Основными направлениями проведённого исследования являлись:

- оценка эмпирических значений параметров алгоритма формирования информационно-поисковых образов;
- испытание способа формирования образов документов, применяющего предложенный алгоритм редукции пространства признаков;
- испытание модифицированного алгоритма послойной кластеризации с оценкой эмпирических значений его входных параметров;
- исследование процесса формирования вербальных описаний

кластеров коллекции документов.

В результате экспериментальной *оценки входных параметров алгоритма формирования образов документов*, во-первых, подобраны значения параметров алгоритмов принудительной и избирательной редукции исходного пространства признаков, во-вторых, выработаны рекомендации по подбору их значений для других коллекций текстовых документов.

В процессе *испытания способа формирования образов документов, применяющего предложенный алгоритм редукции пространства признаков,* проведён анализ его влияния на качество кластеризации этих образов. Данный анализ показал, что применение разработанного алгоритма редукции, во-первых, заметно сократило количество признаков (в 3,5 раза) и связей типа «документ-признак» (в 5,7 раз), что привело к существенному увеличению скорости кластеризации. Во-вторых, позволило повысить качество кластеризации текстов, что подтверждается оценкой значений внутренних и внешних мер качества разбиения документов. Для рассмотренной коллекции документов (случайной выборки 200 документов из коллекции библиотеки CITFORUM) количественная оценка качества кластеризации повысилась почти в 11 раз с применением предложенного алгоритма редукции пространства признаков (рис. 3). Тот факт, что при заметном сокращении признаков получен рост значений мер качества разбиения документов позволил сделать вывод, что из документов были удалены именно невыразительные для их тематик признаки. Что подтвердило верность предположения, положенного в основу разработки алгоритма избирательной редукции.

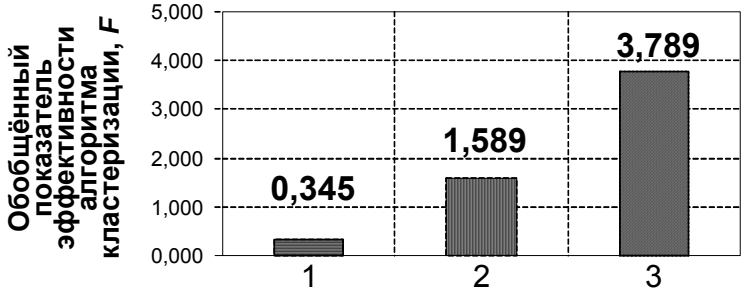

Рис. 3. Зависимость качества кластеризации от способа формирования образов документов

В результате *испытания модифицированного алгоритма послойной кластеризации* подобрана последовательность пороговых значений мер близости документов $\{\tau_1^{sim};\ \tau_2^{sim}\}$ и проведена оценка его эффективности в сравнении с результатами кластеризации с использованием: иерархического агломеративного алгоритма и исходного алгоритма послойной кластеризации

(С. А. Айвазян). Анализ значений обобщающего показателя эффективности алгоритма кластеризации F показал, что предлагаемый подход почти в 2,5 раза эффективнее, чем традиционный иерархический подход к кластеризации, и в 1,6 раза эффективнее исходного подхода послойной кластеризации применительно к выбранной тестовой коллекции полнотекстовых документов (рис. 4).

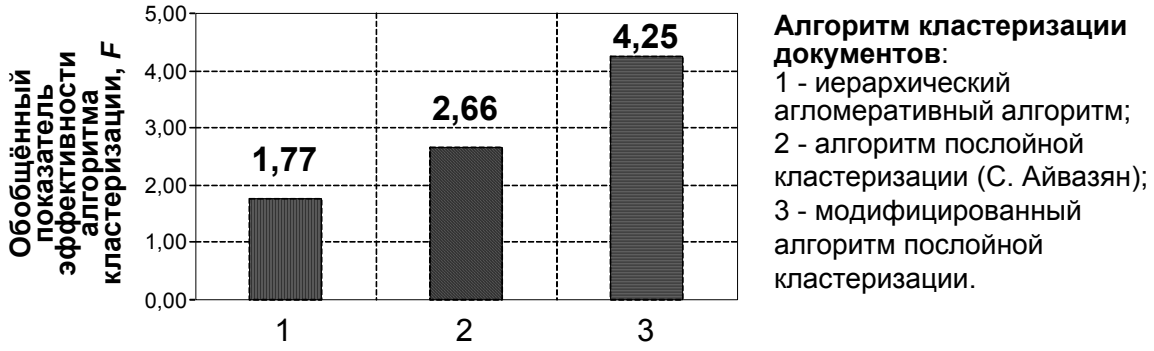

Рис. 4. Сравнительная оценка эффективности алгоритмов кластеризации

Анализ каждой составляющей, учтённой в обобщающем показателе эффективности алгоритма кластеризации F, показал, что:

- затраты времени на выполнение модифицированного алгоритма послойной кластеризации t оказалась в 2 раза ниже, чем у исходного алгоритма, и в 20 раз ниже, чем у иерархического алгоритма (рис. 5).

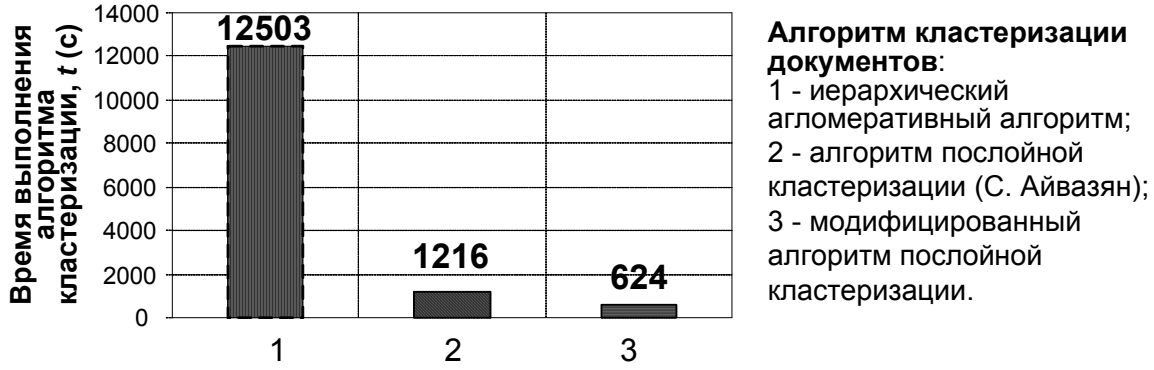

Рис. 5. Сравнительная оценка времени выполнения алгоритма кластеризации *t* (с) для коллекции из 1572 разнородных документов

- по критерию F_{inter} заметное преимущество показал предложенный алгоритм – модифицированный алгоритм послойной кластеризации.
- по критерию F_{exter} получены незначительные различия качества кластеризации всеми тремя алгоритмами. Заметим, что критерии данного типа носят субъективный характер особенно для документов с нечётко очерченной тематикой.

Проведённые эксперименты подтвердили работоспособность предлагаемого в настоящей работе алгоритма кластеризации коллекции документов.

В процессе экспериментального *исследования способа формирования вербальных описаний кластеров коллекции документов* принято решение о:

- формировании краткого названия кластера как первого слова из множества слов, полученных путём пересечения ранжированного по весу списка слов центроида кластера и ранжированного по частоте списка слов заглавий документов, входящих в кластер (включая дочерние кластеры);
- формировании списка ключевых слов кластера как самых значимых слов из ранжированного по весу списка слов центроида кластера.

На рис. 6 представлен пример интерфейса навигации по выборке из коллекции документов библиотеки CITFORUM с помощью автоматически построенного рубрикатора коллекции.

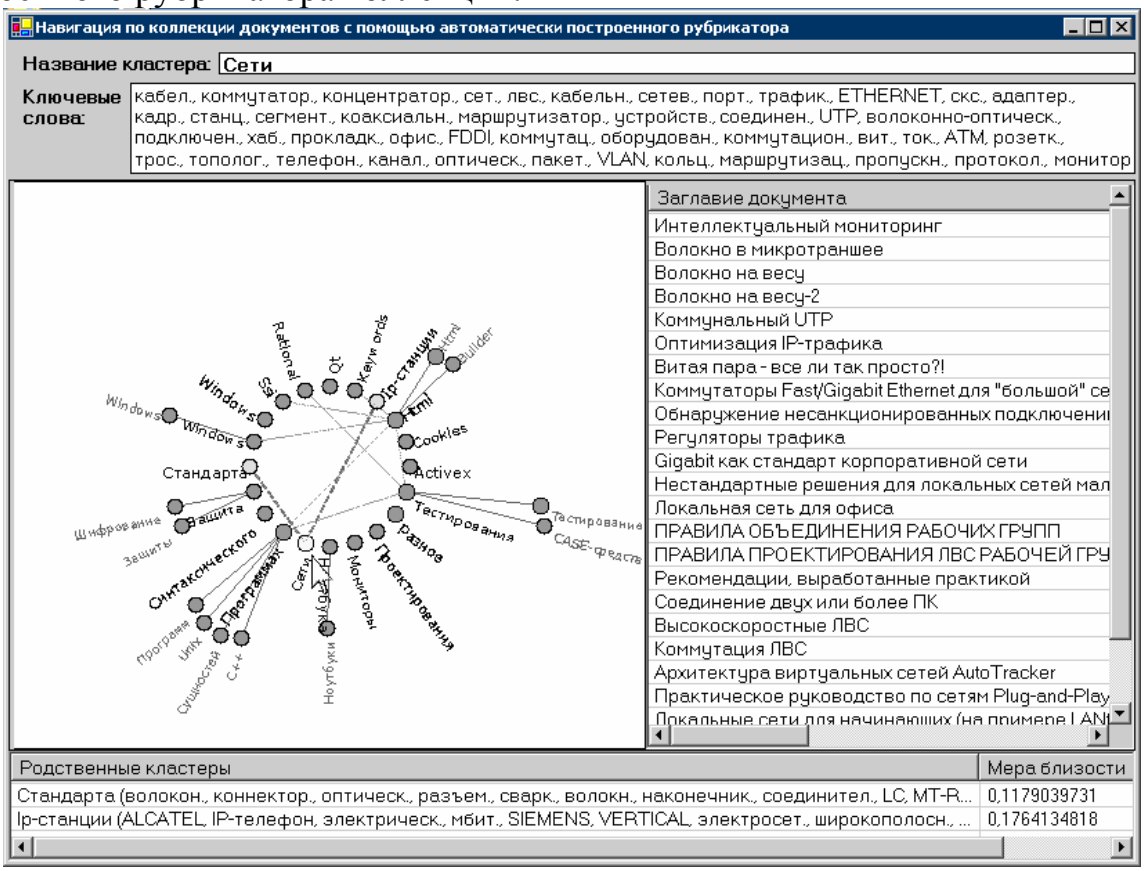

Рис. 6. Пример интерфейса навигации по выборке из коллекции документов с помощью автоматически построенного рубрикатора (выбрана рубрика «Сети»)

В **четвёртой главе** приведены результаты практической апробации разработанной системы автоматического формирования рубрикатора документов

на электронных ресурсах библиотеки МГТУ им. Н. Э. Баумана − коллекции полных текстов авторефератов диссертаций. Коллекция авторефератов насчитывала 234 документа научно-технической направленности, общим объёмом 18,14 МБ простого текста. Оценка качества её кластеризации выполнена путём вычисления погрешности классификации по сравнению с:

- индексом УДК, присвоенным авторами авторефератов диссертации. В этом случае погрешность автоматической классификации составила 3,2%;
- областью знания по номенклатуре ВАК, по которой планировалась защита диссертации. В этом случае погрешность составила 13,6%, что объясняется тематическим перекрытием укрупнённых направлений, по которым осуществляется подготовка и защита диссертаций.

Проведённые эксперименты показали работоспособность предложенного в настоящей работе метода автоматического формирования рубрикатора документов и положенного в его основу алгоритма кластеризации полных текстов документов.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ РАБОТЫ

1) Предложен метод автоматического формирования рубрикатора коллекции электронных полнотекстовых документов, применимый для совокупности научно-технических текстов произвольной тематики и объёма в условиях отсутствия специализированной априорной информации об их содержании.

2) Разработан модифицированный алгоритм послойной кластеризации, позволяющий автоматически разбивать тексты на естественном языке на тематические группы с возможностью простого управления глубиной и уровнем детализации иерархии этих групп.

3) Предложен подход к оценке тематической близости документов с использованием метода редукции пространства признаков, составляющих информационно-поисковые образы, и на его основе разработан алгоритм формирования информационно-поисковых образов документов, позволяющий повысить качество и скорость выполнения автоматической кластеризации документов.

4) Разработан программный комплекс, реализующий предложенный метод автоматического формирования рубрикатора, а также средства визуального отображения полученных результатов для навигации по коллекции документов. Автоматически построенные рубрикаторы отражают иерархические и родственные связи между областями знаний, обладают

автоматически получаемыми вербальными описаниями этих областей знаний и способны служить как самостоятельным поисковым аппаратом, так и средством повышения качества работы других поисковых механизмов.

5) Экспериментально подтверждена эффективность предложенных алгоритмов формирования образов документов и их кластеризации. Формирование образов документов с применением предложенного алгоритма редукции привело на тестовой коллекции к увеличению в 11 раз значения критерия эффективности кластеризации по сравнению с формированием образов без использования механизма редукции. Кластеризация документов с применением модифицированного алгоритма послойной кластеризации привела к увеличению критерия эффективности кластеризации в 2,5 раза по сравнению с кластеризацией на основе традиционного иерархического алгоритма.

6) Итоговая проверка метода на политематической коллекции из 234 авторефератов диссертаций показала, что автоматическая классификация документов привела к погрешности в 3,2% по сравнению с классификацией по УДК каждого автореферата диссертации.

РАБОТЫ ПО ТЕМЕ ДИССЕРТАЦИИ

1) Автоматизированная библиотечно-информационная система технического университета / А.Е. Шиваров, Г.В. Абрамов, О.В. Пескова, Н.А. Белостоцкий // Вестник МГТУ им. Н.Э. Баумана. Приборостроение. – 2007. – №4. – С. 21-32.

2) Пескова О. В. Автоматизация работы с классификаторами документов библиотеки МГТУ им. Н. Э. Баумана // Культура народов Причерноморья. – 2004. – Т. 2, № 48. – С. 38-41.

3) Пескова О. В. Автоматическая классификация электронных текстовых документов с применением механизма обратной связи // Технологии Microsoft в теории и практике программирования: Труды всероссийской конференции студентов, аспирантов и молодых учёных. – Москва, 2005. – С. 54-55.

4) Пескова О. В. Автоматическое формирование рубрикатора полнотекстовых документов // НТИ-2007: Материалы 7-ой международной конференции. – Москва, 2007. – С. 241-242.

5) Пескова О. В. Автоматическое формирование тематической схемы коллекции документов // Технологии Microsoft в теории и практике программирования: Труды всероссийской конференции студентов, аспирантов и молодых учёных. – Москва, 2006. – С. 66-68.

6) Пескова О. В. Исследование и разработка метода автоматического анализа документов для формирования индексов УДК // Информатика и системы управления в XXI веке: Сборник трудов молодых учёных, аспирантов и студентов МГТУ им. Н.Э. Баумана. – 2005. – №3. – С. 90-92.

7) Пескова О. В. Классификация документов в электронных библиотеках [Электронный ресурс] / О. В. Пескова // Крым 2007. Библиотеки и информационные ресурсы в современном мире науки, культуры, образования и бизнеса: Материалы 14-ой Международной конференции. – Судак, 2007. – Режим доступа: http://www.gpntb.ru/win/inter-events/crimea2007/cd/proceeding.html, свободный.

8) Пескова О. В. Методы автоматической классификации текстовых электронных документов // Научно-техническая информация. Сер. 2. – 2006. – №3. – С. 13-20.

9) Пескова О. В. Методы автоматической классификации электронных текстовых документов без обучения // Научно-техническая информация. Сер. 2. – 2006. – № 12. – С. 21-32.

10) Свидетельство об официальной регистрации программы для ЭВМ №2007610196. Автоматизированная библиотечно-информационная система «Яуза» / А.Е. Шиваров, Г.В. Абрамов, Н.А. Белостоцкий, О.В. Пескова. – Москва, 2007. – 1с.

11) Свидетельство об официальной регистрации программы для ЭВМ №2007614766. Информационная система автоматического формирования рубрикатора коллекции полнотекстовых документов «Авторубрикатор» / О. В. Пескова – Москва, 2007. – 1с.