Автоматическое формирование рубрикатора полнотекстовых документов

© Пескова Ольга Вадимовна

МГТУ им. Н. Э. Баумана
opeskova@mail.ru

Аннотация

Рассмотрены вопросы построения информационно-поисковой системы на основе механизма автоматической классификации полнотекстовых документов. Изложены подходы и алгоритмы формирования образов документов, их кластеризации и оценки полученного разбиения данных. Предложен метод автоматического формирования рубрикатора коллекции документов в виде, унаследованном от традиционного библиотечного предметного рубрикатора. Проверена работоспособность метода на небольших коллекциях, содержащих русскоязычные документы различного размера и содержания.

1 Введение

На протяжении последних десятилетий во всём мире наблюдается стремительный рост количества и объёмов коллекций полнотекстовых документов, т. е. множеств документов, содержащих тексты на естественном языке, и доступных 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. В связи с этим одной из важнейших задач является создание программных средств, предоставляющих пользователю эффективные механизмы поиска документов.

Традиционными механизмами поиска в коллекциях полнотекстовых документов, являются: *поиск по ключевым словам* (а) и *классификационный поиск* (б). Поиск по ключевым словам из-за кажущейся простоты использования применяется в большинстве поисковых систем. Однако его эффективность существенно зависит от удачного описания пользователем своей информационной потребности в форме запроса на естественном языке. Следовательно, пользователь, малознакомый с искомой предметной областью или малоопытный в вопросах использования поисковых машин, не находит требуемых документов. К такому же результату часто приводит необходимость отбора пользователем документов среди огромных списков, найденных поисковой машиной документов.

Классификационный поиск самостоятельно или в его сочетании со поиском по ключевым словам благодаря интуитивно понятному интерфейсу навигации позволяет легко формулировать и уточнять информационные потребности, что повышает эффективность и удобство поиска документов. Во-первых, в результате классификации всей коллекции документов пользователю становится доступным средство тематической навигации по коллекции, что позволит малоопытному пользователю легко углубляться в искомую предметную область. Во-вторых, механизмы поиска по запросу могут использовать информацию о классификации документов для уменьшения ширины поисковой области, таким образом сокращая число нерелевантных документов в результатах поиска. В-третьих, трудоёмкость выбора среди результата поиска может быть снижена, если документы поисковой выборки предоставлять пользователю не в виде огромных списков, а в виде набора тематических групп, на которые автоматически разбиваются релевантные запросу документы. В таком случае пользователь легко отбросит документы неинтересующих его тематик.

Настоящая работа посвящена классификационному поиску. 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. Выделяют два вида методов автоматической классификации полных текстов: (а) *методы категоризации* и (б) *методы кластеризации*. В обоих случаях входными данными являются информационно-поисковые образы документов, представляемые в виде

множества признаков, характеризующих содержание текста документа.

Методы категоризации 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 T. Joachims, D. D. Lewis, R. E. Schapire, H. Schutze, F. Sebastiani, Y.Yang, I. Dagan, S. T. Dumais, М. С. Агеев, И. Е. Кураленок, И. С. Некрестьянов, В. И. Шабанов и ряда других. Однако данный подход решает не все проблемы традиционного классификационного поиска, поскольку регулярная актуализация рубрикатора связана с высокими экспертными затратами на анализ ситуации и подготовку новых обучающих данных. Широко известные методы категоризации подробно изложены в работе [4] и далее рассмотрены не будут.

Методы кластеризации классифицируют документы в условиях отсутствия предопределенного набора рубрик и множества документов-образцов, разбивая документы на группы (кластеры) на основе анализа тематической близости между ними. Разработке алгоритмов данного вида и способов оценки качества получаемого разбиения данных, а также связанным с ними алгоритмам представления текстов посвящены труды таких авторов как C. J. van Rijsbergen, G. Salton, D. Manning и H. Schutze, T. Kohonen, O. Eli Zamir, J. C. Bezdek, M. Halkidi, Д. В. Ландэ, М. В. Киселев, К. М. Кириченко и ряда других. Решить представленную проблему потенциально способны только алгоритмы кластеризации. Их применение позволит получить средство навигации как по всей коллекции документов, так и по её подмножествам, динамически формируя для каждого случая наиболее подходящий предметный рубрикатор, например, для результата поиска по ключевым словам.

Поскольку данная работа зародилась в процессе автоматизации университетской библиотеки, то особое значение имеет вид автоматического рубрикатора, который должен быть унаследован от традиционного библиотечного предметного рубрикатора, основанного на многолетнем опыте обслуживания читателей. Кроме того, метод построения рубрикатора должен быть применим к коллекциям русскоязычных документов, которые могут сильно различаться по размеру (статьи, учебники, книги, технические руководства и др.) и по тематике (физика, металлургия, транспорт, информатика, радиотехника, авиация, энергетика, приборостроение и др.). Таким образом, поставлена задача разработать метод автоматического построения рубрикатора в ориентированном на читателя виде для политематической коллекции документов без ограничения на их объём и оценить применимость для её решения метода кластерного анализа.

2 Обзор известных методов

2.1 Способы формирования образов документов

Формирование образов документов заключается в преобразовании исходного текста на естественном языке в некоторое формализованное представление его смысла. В задаче кластеризации образы документов представляют собой многомерные векторы в пространстве их признаков, например, слов документов. Тогда метод кластерного анализа определяет сходство документов путём вычисления геометрической близости их векторов. В работе выполнено обобщение известных способов формирования образов документов (рис. 1), направленных на решение проблемы автоматической обработки текстов, обусловленной неоднозначностями естественного языка [22, 2, 8, 25, 23и др.].

Рис. 1. Этапы формирования образа полнотекстового документа для задачи кластеризации и способы к их реализации (статистический подход)

Выполненное обобщение показало, что наиболее приемлемым является подход, который использует в качестве смысловых признаков одиночные слова, прошедшие морфологический анализ и оценку их значимости по схеме TFIDF. Что касается многословных признаков, теоретически они представляются более подходящими для описания смысла текста. Однако известные попытки их применения показывают, что качество автоматической классификации, если и увеличивается, то незначительно. Причём это достижимо только в том случае, когда уделяется особое внимание тщательному отбору «качественных» фраз, что связано с большими вычислительными затратами, иначе может наблюдаться даже снижение качества поиска или классификации [2, 8, 5]. Что касается расширения понятия признака документа информацией из тезаурусов и других внешних словарей, то

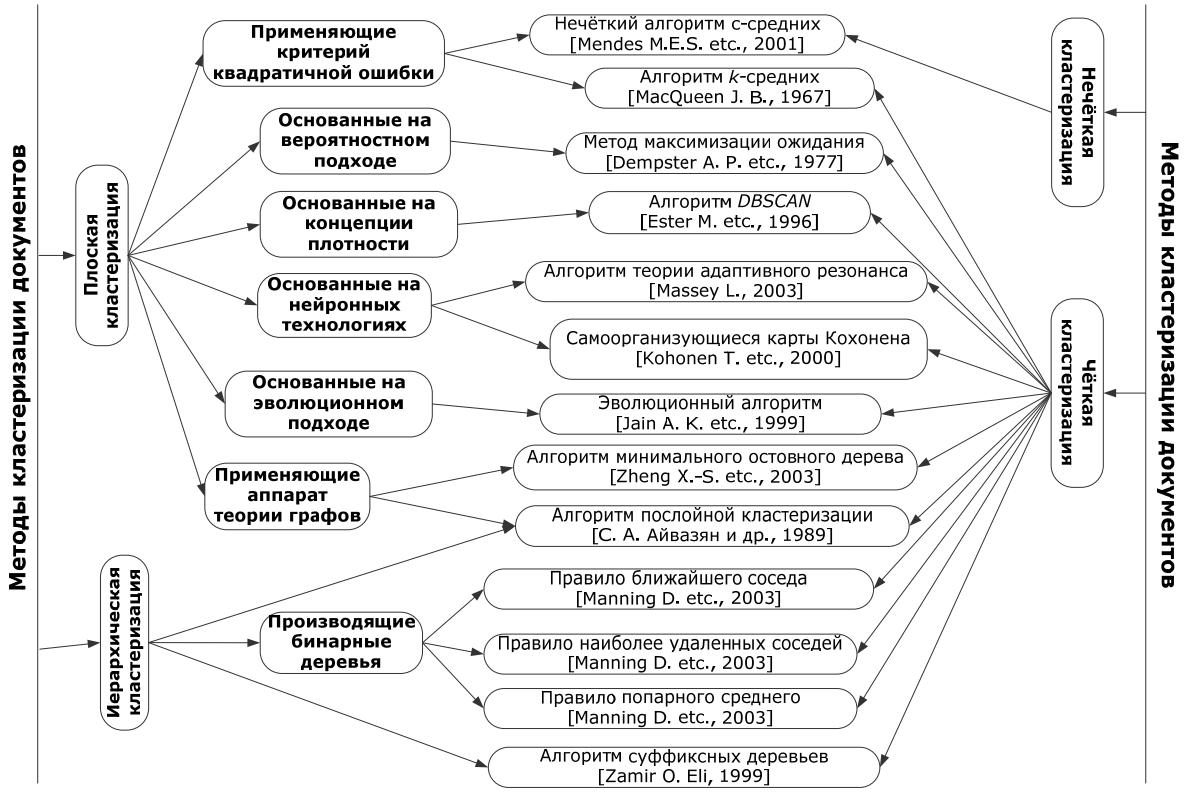

Рис. 2. Алгоритмы автоматической кластеризации документов

требование применимости метода классификации к политематическим коллекциям исключает возможность их использования.

Заметим, что количество однословных признаков может достигать десятков или сотен тысяч даже для коллекций небольшого размера. Процесс кластеризации в таких условиях связан не только с высокими вычислительными затратами, но и с низким качеством разбиения на кластеры. Поэтому остро стоит проблема отбора информативных признаков из всего первоначального множества слов. В работе сделана попытка решить данную проблему в условиях отсутствия априорной информации о массиве документов. Соответствующий алгоритм будет изложен ниже.

2.2 Алгоритмы кластеризации документов

Наибольшее распространение получили следующие алгоритмы кластеризации полнотекстовых документов (рис. 2): иерархические (основанные на правилах ближайшего соседа, наиболее удалённых соседей и попарного среднего, суффиксные деревья) [18, 26], квадратичной ошибки (алгоритм k-средних, нечёткий алгоритм c-средних) [17, 21], алгоритмов теории графов (алгоритм минимального остовного дерева) [27], вероятностные (алгоритм максимального ожидания) [11], основанные на концепции плотности (алгоритм DBSCAN) [12], нейросетевые (самоорганизующиеся карты Кохонена, алгоритмы теории адаптивного резонанса) [19, 15] и эволюционные (генетические) [14].

Для кластеризации документов в настоящей работе потребовался алгоритм, способный разбить документы на такие кластеры, которые пригодны для формирования рубрикатора в унаследованном от библиотечного предметного рубрикатора виде. Например, Предметный Рубрикатор в библиотеке МГТУ является иерархическим с глубиной иерархии не более 3 уровней, содержит очень «плотные» рубрики, это приводит к тому, что в рубрикаторе 25% всех рубрик (4520) являются корневыми, и в добавок между рубриками имеется родственная связь типа «смотри также». Следовательно, необходим алгоритм, позволяющий разбивать документы на иерархические кластеры с возможностью простого управления как количеством уровней, так и степенью детализации кластеров-рубрик на каждом.

Обобщение известных методов кластеризации показало, что традиционные иерархические и плоские алгоритмы для этого не пригодны. Поэтому предложен подход, называемый послойной кластеризацией, который взят из теоретического труда Айвазяна С. А [6]. Этот подход позволит получать необходимую структуру кластеров и для него характерны: (а) более низкие вычислительные затраты по сравнению с традиционными иерархическими алгоритмами, (б) отсутствие необходимости сложной настройки входных параметров и (в) независимость результата кластеризации от выбора начальных точек поиска разбиения данных. Однако, его применение к

текстовым документам неизвестно. Поэтому в работе была поставлена и решена задача на основе выбранного подхода создать алгоритм, обеспечивающий ещё и высокую достоверность автоматической классификации текстов.

2.3 Способы оценки кластерного решения

Оценка качества разбиения документов на кластеры различными алгоритмами выполняется путём вычисления значений мер качества полученного результата кластеризации. Выделяют следующие два вида мер качества кластеризации документов: *внешние* (а) и *внутренние меры* (б) [3, 9, 13 и др.]. Внешние меры основаны на сравнении автоматического разбиения с полученным от экспертов «эталонным» разбиением этих же данных. Идея, положенная в основу этих мер, заключается в том, чтобы для каждой пары документов автоматически сопоставить два решения о сходстве их тематиках. Одно решение получено от экспертов, второе – в процессе автоматической классификации. Примерами внешних мер являются традиционные для оценки систем поиска такие характеристики, как: полнота, точность, ошибка классификации, F_1-мера и другие. В данном случае их оценка производится на основе следующей таблицы:

Таблица 1.
Основные категории результата кластеризации документов и количество документов, принадлежащих каждой категории

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

Далее внешние меры вычисляются по известным формулам:

$$Recall = \frac{a}{a+b},$$

$$Precision = \frac{a}{a+c},$$

$$Error = \frac{b+c}{(a+b+c+d)},$$

$$F_1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall},$$

где *Recall* – это мера полноты, *Precision* – это мера точности, *Error* – ошибка автоматической классификации, F_1 – F_1-мера.

Внутренние меры основаны на анализе полученного разбиения без привлечения экспертов, они анализируют такие свойства как отделимость и компактность кластеров документов. т. е. оптимальным кластерным решением считается то, расстояние между кластерами которого максимально, и в то же время расстояние между документами внутри каждого кластера минимально. Приведём несколько примеров. Для краткости не будем описывать все формулы, по которым вычисляются перечисленные индексы, их можно найти в указанной литературе.

Мерой качества результата иерархической кластеризации является *кофенетический коэффициент корреляции (CPCC)*. Для его вычисления необходимо определить кофенетическую матрицу S_C, каждый элемент которой представляет собой уровень близости, на котором документы d_i и d_j впервые встретились в одном кластере, (номер уровня в иерархии кластеров). *CPCC* – это статистический индекс, показывающий степень сходства кофенетический матрицы S_C и действительной матрицы близости документов коллекции S, вычисляемый следующим образом:

$$CPCC(C) = \frac{(1/M)\sum_{i=1}^{N_D-1}\sum_{j=i+1}^{N_D}\left(s_{ij}sc_{ij} - \mu_S\mu_{S_C}\right)}{\sqrt{\left((1/M)\sum_{i=1}^{N_D-1}\sum_{j=i+1}^{N_D}\left(s_{ij}^2 - \mu_S\right)\right)\left((1/M)\sum_{i=1}^{N_D-1}\sum_{j=i+1}^{N_D}\left(sc_{ij}^2 - \mu_{S_C}\right)\right)}}$$

где $-1 \le CPCC \le 1$, $M = \frac{N_D(N_D-1)}{2}$, N_D – количество документов, s_{ij} и sc_{ij} – (i, j)-ые значения матриц S и S_C соответственно, $\mu_S = \frac{1}{M}\sum_{i=1}^{N_D}\sum_{j=i+1}^{N_D}s_{ij}$, $\mu_{S_C} = \frac{1}{M}\sum_{i=1}^{N_D}\sum_{j=i+1}^{N_D}sc_{ij}$ – здесь средние значения матриц S и S_C соответственно. Чем ближе значение *CPCC* к нулю, тем ниже сходство между матрицами.

Мерами качества плоского чёткого разбиения документов являются: *индекс Дана* (*DI*), модифицированный Бездеком для снижения чувствительности к шуму; *мера Дейвиса-Булдина* (*DB*), являющаяся функцией отношения суммы внутрикластерного разброса к межкластерному растоянию; индекс Калинского и Гарабача (Calinski и Harabasz) (*CH*), вычисляющий отношение следа матрицы межкластерного разброса к следу матрицы внутрикластерного разброса, *I*-индекс и другие [10, 20, 24 и др.]. Оптимальное разбиение данных достигается при минимизации значения второго индекса и максимизации значений остальных индексов.

Мерами качества плоского нечёткого разбиения документов являются, например:

модифицированный индекс коэффициента разделения (MPC), учитывающий только нечёткую принадлежность, и *индекс Хие и Бени* (*XB*), учитывающий не только нечёткую принадлежность, но и структуру данных [16]. Оптимальное разбиение данных достигается при максимизации значений этих индексов.

Заметим, что какими бы ни были надёжными меры обоснованности на одних данных, на других их поведение может существенно измениться. Дело в том, что данный подход к оценке обоснованности разбиения данных хорошо работает при условии, что достоверные кластеры являются компактными, плотными и хорошо разделимыми. В задаче кластеризации текстов документы в общем случае группируются на кластеры произвольной формы. Тогда традиционные основы критериев обоснованности (дисперсия, плотность, отделимость) не являются достаточными. В высокоразмерном пространстве признаков документов сходство между документом и центроидом кластера в общем случае является низким и, поэтому трудно ожидать компактных кластеров.

3. Предложенный метод автоматического формирования рубрикатора коллекции документов

Рассмотрим предложенные алгоритмы для каждого этапа формирования рубрикатора.

3.1 Алгоритм формирования образов

Алгоритм формирования образов на первом шаге строит множество признаков документов всей коллекции $P = \{p_i\}, i = \overline{1, N_P}$. Для этого из текстов документов извлекаются все слова, из них удаляются общеупотребительные, входящие в список стоп-слов, оставшиеся слова проходят морфологический анализ по алгоритму М. Портера [1], выделяющего псевдоосновы слов. В результате получается исходное множество признаков документов, которое задаёт многомерное пространство $\Pi^{N_P}(\vec{D})$. Исходные образы документов строятся в виде векторов в данном пространстве $\left(\vec{D} = \{\vec{d}_j\}, j = \overline{1, N_D}\right)$. Координатами вектора документа являются веса соответствующих признаков в данном документе, вычисленные по схеме TFIDF.

Следующий шаг – редукция признаков, целью которой является повышение качества представления содержания документов их образами. Предложенный алгоритм редукции заключается в последовательной обработке исходного множества признаков с использованием, во-первых, общеизвестного способа отсечения признаков (*принудительной редукции*), во-вторых, с применением нового подхода к выявлению малоинформативных признаков документов (*избирательной редукции*).

Принудительная редукция состоит из следующих операций:

1) Удалить из множества признаков P все признаки, частота встречаемости которых в коллекции менее $\tau_{\min}^{DF}$ или более $\tau_{\max}^{DF}$ документов, где $\tau_{\min}^{DF}$ и $\tau_{\max}^{DF}$ – заданные пороговые значения.

2) Удалить из образа каждого документа все признаки, имеющие наиболее низкие веса, доля низковесовых признаков задаётся общим параметром τ^{WP}.

Избирательная редукция в отличие от принудительной признаёт малоинформативными не одни и те же слова для всех документов сразу. Она мотивирована тем, что один и тот же признак может являться значимым для описания одной предметной области и незначимым для другой, и при этом иметь достаточно высокую частоту встречаемости в документах обеих областей. Тогда анализировать его информативность следует по отдельности для каждой тематики, представленной в коллекции. Таким образом, избирательная редукция состоит из следующих операций:

1) Выполнить первоначальную кластеризацию документов с высоким значением меры близости документов τ^{sim}, где $0 \le \tau^{sim} \le 1$ – входной параметр алгоритма (алгоритм кластеризации как в п. 3.2). В результате выделяется множество кластеров «очевидно родственных» документов $C^0 = \{C_1^0, ..., C_{|c^0|}^0\}$, т. е. документов, принадлежность которых к одному тематическому классу ярко выражена в использовании большого количества одинаковых слов. Полагается, что все признаки, встретившиеся в «очевидно родственных» документах, являются представительными для этого тематического класса.

2) Среди всех признаков «очевидно родственных» документов отобрать признаки, обладающие наиболее высокой различительной способностью. Благодаря таким признакам в группу данного тематического класса не попали документы, являющиеся менее близкими по смыслу с остальными документами этой группы. Для выполнения данного отбора следует:

а) для каждой группы документов

$$C_i^0 = (\vec{d}_1^{(i)}, ..., \vec{d}_{N_i}^{(i)}) = \begin{pmatrix} w_{1(i)}^{(1)} & \cdots & w_{N_i(i)}^{(1)} \\ & \cdots & \\ w_{1(i)}^{(N_P)} & \cdots & w_{N_i(i)}^{(N_P)} \end{pmatrix},$$

где N_i – количество документов в C_i^0; $w_{j(i)}^k$ – вес k-ого признака в j-ом документе, принадлежащем i-ому кластеру C_i^0, $k = \overline{1, N_P}$, $j = \overline{1, N_i}$,

перейти в пространство документов данной группы $\prod^{C_i^0}(P)$, где объектами наблюдения являются признаки

$$\begin{pmatrix} \vec{p}_1 \\ \dots \\ \vec{p}_{N_P} \end{pmatrix} = \begin{pmatrix} w_{1(i)}^{(1)} & \cdots & w_{N_i(i)}^{(1)} \\ & \cdots & \\ w_{1(i)}^{(N_P)} & \cdots & w_{N_i(i)}^{(N_P)} \end{pmatrix};$$

б) в каждом полученном подпространстве $\prod^{C_i^0}(P)$ выполнить кластеризацию признаков со значением τ^{sim}=1 (алгоритм кластеризации как в п. 3.2). Этот шаг основан на предположении, что в подпространстве родственных документов информативные признаки окажутся слабо схожими между собой, а малоинформативные признаки будут в векторном виде лежать на одной прямой;

б) считать информативными только те признаки, которые образовали одноточечные кластеры, остальные удалить из образов документов, принадлежащих группе C_i^0.

3.2 Модифицированный метод послойной кластеризации

Итоговые образы документов разбиваются на кластеры с применением послойного подхода, суть которого заключается в представлении информации о документах в виде графа, вершины которого соответствуют документам, а рёбра помечены значениями меры близости этих документов. Мера близости вычисляется как косинус угла между векторами документов. Тогда для заданной последовательности порогов меры близости $\{\tau_1^{sim}, \tau_m^{sim}\}$ определяются подграфы $G^1 \subseteq \dots \subseteq G^m$ путём удаления из исходного графа рёбер, помеченных значением меры близости, меньшим порога τ_t^{sim}. Затем для каждого подграфа выделяются компоненты связности. Множество выделенных компонент t-ого подграфа считается разбиением коллекции документов на уровне близости τ_t^{sim}, иначе называемое *t*-ым слоём классификации документов. На выходе алгоритма получается *m* вложенных разбиений, которые отражают иерархические связи между кластерами документов.

Первые испытания данного подхода показали необходимость его модификации, вызванной склонностью некоторых документов являться «узкими» перемычками между кластерами, что приводило к ошибочному объединению кластеров на более высоких уровнях. Для уменьшения влияния таких документов на результат кластеризации, в работе предложено заменять все документы, образовавшие кластеры, центроидами данных кластеров, т. е. их средними элементами, для вычисления кластеров на последующих уровнях.

Рис. 3. Модифицированный алгоритм послойной кластеризации

На заключительном этапе иерархический набор кластеров преобразуется в рубрикатор следующим образом:

– выявляются родственные связи между кластерами одного и того же уровня путём вычисления меры близости между их центроидами;

– формируется описание рубрик, состоящее из краткого названия и списка ключевых слов, детально раскрывающих тематику рубрики. Список ключевых слов – это наиболее весомые слова центроида кластера. А название – это самое весомое слово, встречающееся и в заглавиях документов кластера, и в его центроиде.

4 Испытания

Разработанные алгоритмы реализованы в программном комплексе, с которого помощью исследован предложенный подход к кластеризации полнотекстовых документов. Тестовыми данными являлись две коллекции русскоязычных полнотекстовых документов:

1) документы on-line библиотеки по информационным технологиям CITFORUM [7] (1572 документа, тематика различных областей информационных технологий, документы различного объема (от обзорных статей до полных учебников по программированию), общий объём – 81,35 МБ текста) – коллекция CL1572;

2) электронные ресурсы библиотеки МГТУ им. Н. Э. Баумана − авторефераты диссертаций (234 документа, тематика различных научных специальностей, общий объём – 18,14 МБ текста) – коллекция TAL234.

Пример экранной копии интерфейса навигации по подмножеству коллекции CL1572 (200 случайных документов) с помощью автоматически

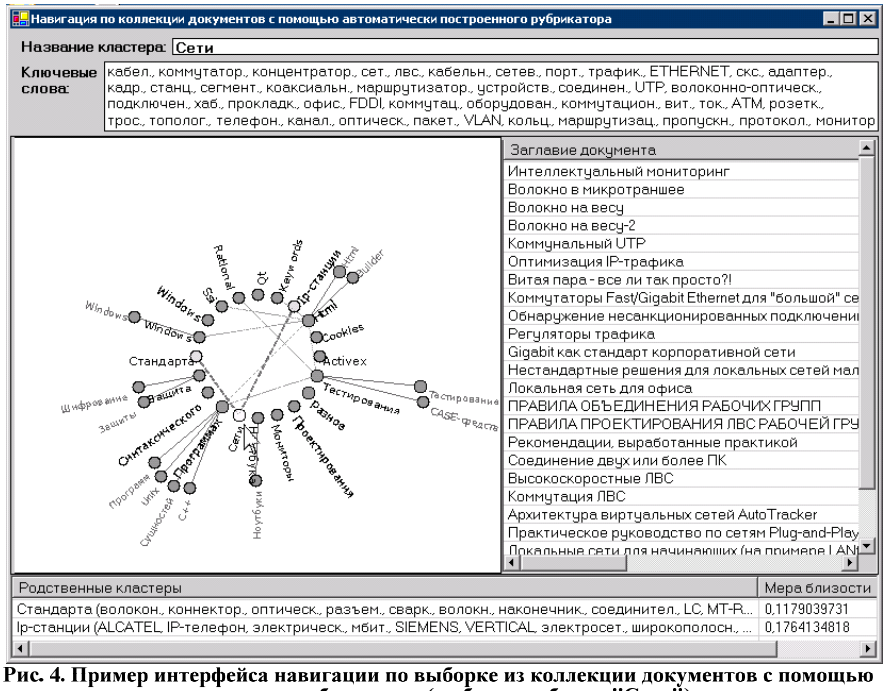

Рис. 4. Пример интерфейса навигации по выборке из коллекции документов с помощью автоматически построенного рубрикатора (выбрана рубрика "Сети")

построенных рубрикаторов представлен на рис. 4. У каждой рубрики имеется однословное название, подписанное на графе, при выборе рубрики отображается список её ключевых слов, список родственных рубрик и список документов, ей принадлежащих. Родственные связи также отражены на самом графе рубрикатора. В зависимости от выбора порогов близости документов для алгоритма кластеризации может быть получен рубрикатор с разной степенью укрупнения рубрик.

4.1 Испытание предложенного алгоритма формирования образов документов

Испытания алгоритма формирования образов документов проведены на коллекции CITFORUM, содержащей разнородные по объёму и содержанию документы. Сравнивался результат кластеризации с применением предложенного подхода к отбору информативных признаков и без него.

Таблица 2.
Оценка способа формирования образов. (1) – без редукции, (2) – с принудительной редукцией, (3) – с принудительной и избирательной редукцией

	(1)	(2)	(3)
Внешние меры качества кластеризации			
MicroF$_1$-мера	0,190	0,466	**0,505**
Error	0,506	0,101	**0,084**
Внутренние меры качества кластеризации			
CPCC	-0,188	-0,508	**-0,580**
DI	0,539	**0,598**	0,577
DB	0,196	0,258	**0,180**
CH	3,086	6,3687	**7,226**
I-Index	0,0014	0,0016756	**0,0018**
Скорость кластеризации			
Время (с)	1783	584	**85**

Анализ показал, что применение разработанного алгоритма редукции, сократило количество признаков в 3,5 раза и связей типа «признак-документ» в 5,7 раз, что, с одной стороны, снизило вычислительные затраты алгоритма кластеризации, а с другой стороны, повысило качество разбиения данных.

Значения параметров алгоритма редукции подобраны на той же тестовой коллекции путём оценки качества кластеризации для различных значений. В результате выработаны следующие рекомендации: $\tau_{\min}^{DF} = 1$ и $\tau_{\max}^{DF} = (75\%$ *от количества документов*); $\tau^{WP} = 0{,}60$; $\tau^{sim} = 0{,}40$.

4.2 Испытание модифицированного алгоритма послойной кластеризации

Испытание алгоритма кластеризации проведено как на коллекции CITFORUM, так и на коллекции авторефератов.

На коллекции CITFORUM кластеризация выполнена тремя алгоритмами:

1) иерархическим агломеративным алгоритмом по правилу попарного среднего; усечение полученного дерева выполнено при минимальном пороге меры близости документов внутри кластеров, равном $\tau^{sim} = 0{,}20$;

2) исходным алгоритмом послойной кластеризации при $\tau_1^{sim} = 0{,}40$ и $\tau_2^{sim} = 0{,}20$.

3) модифицированным алгоритмом послойной кластеризации при $\tau_1^{sim} = 0{,}40$ и $\tau_2^{sim} = 0{,}20$.

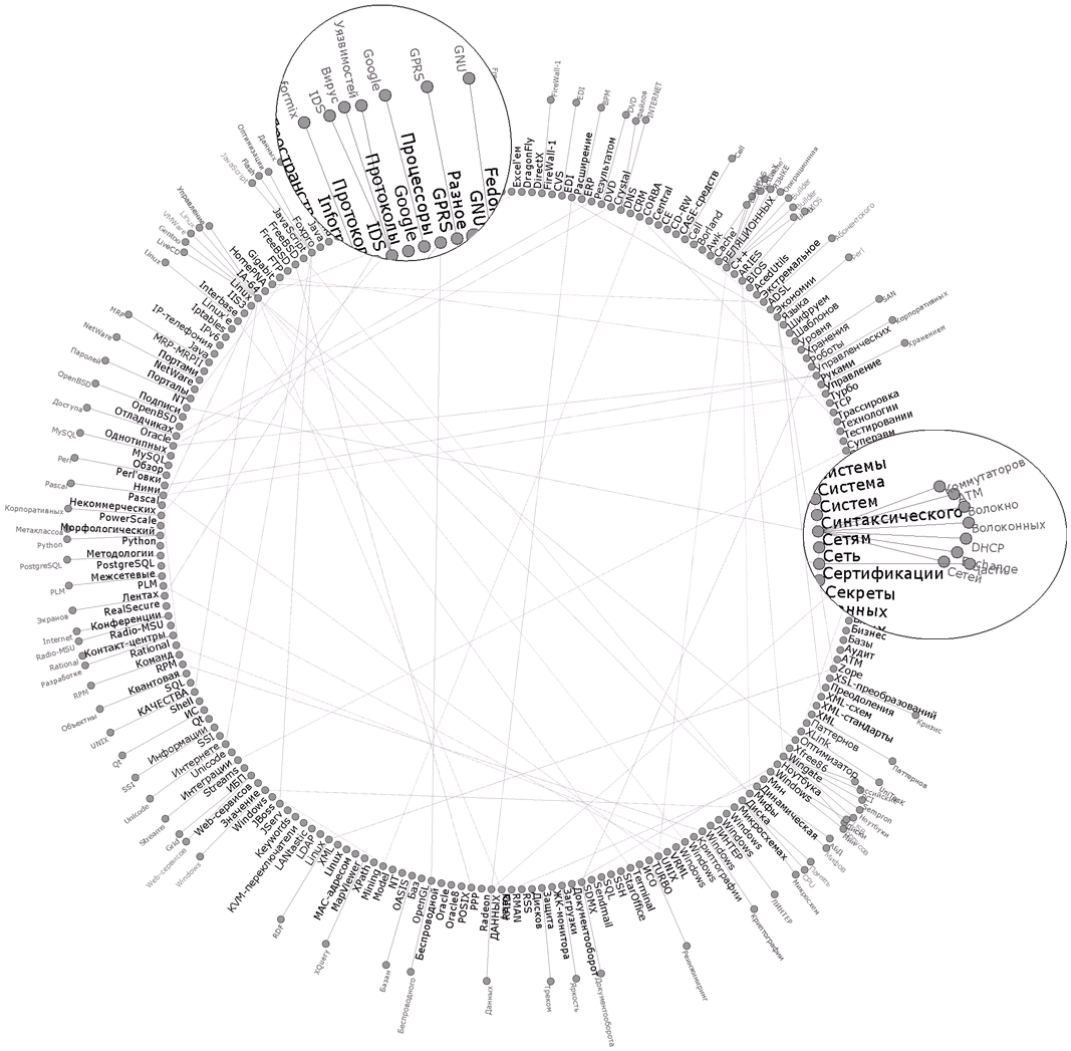

Рис. 5. Рубрикатор коллекции CL1572

Рубрикатор, полученный для полной коллекции CITFORUM, представлен на рис. 5.

Таблица 3.
Оценка алгоритма кластеризации. (1) – иерархический агломеративный алгоритм, (2) – исходный алгоритм послойной кластеризации, (3) – модифицированный алгоритм послойной кластеризации

	(1)	(2)	(3)
Внешние меры качества кластеризации			
MicroF$_1$-мера	0,21838	0,11321	**0,25823**
MacroF$_1$-мера	0,10998	0,06513	**0,14383**
Error	0,05741	0,30799	**0,03439**
Внутренние меры качества кластеризации			
CPCC	**-0,4642**	-0,1399	-0,3553
DI	**0,598**	0,378	0,500
DB	0,394	**0,053**	0,076
CH	**13,505**	2,746	6,278
I-Index	**0,000041**	0,000012	0,000022
Скорость кластеризации			
Время (с)	12503	1216	**624**

Сравнение результатов показало, что модифицированный алгоритм разбил тестовую коллекцию с наименьшей ошибкой автоматической классификации *Error* за значительно меньшее время.

Оценка алгоритма кластеризации на коллекции авторефератов диссертаций TAL234 проведена путём анализа ошибки автоматической классификации:

– в сравнении с классификацией авторефератов по УДК погрешность составила 3,2%;

– в сравнении с областью знания по номенклатуре ВАК – 13,6%, что объясняется тематическим перекрытием укрупнённых направлений, по которым осуществляется подготовка и защита диссертаций.

Вопрос устойчивости алгоритма оценен путём половинного деления тестовых коллекций и измерения ошибки классификации как для полной коллекции, так и для каждой из половин в отдельности. Для полной коллекции авторефератов погрешность составила 3,2%, для первой и второй половин коллекции – 3,2% и 5,0% соответственно. Для полной коллекции CITFORUM погрешность

классификации составила 3,4%, а для первой и второй половин – 5,1% и 5,6% соответственно.

Посмотреть примеры автоматически построенных рубрикаторов для некоторых подмножеств коллекции CITFORUM и осуществить навигацию по ним возможно по адресу http://demo.peskova.ru.

5 Заключение

Проведённые эксперименты показали эффективность предложенного в работе метода автоматического формирования рубрикатора документов и положенного в его основу алгоритма кластеризации полных текстов. Программная реализация метода предназначена для электронных библиотек как элемент их поисковых систем. Такой элемент способен являться как самостоятельным поисковым механизмом, так и служить средством повышения качества работы других поисковых механизмов, например, поиска по ключевым словам. Также предложенный метод может быть использован при разработке средств анализа динамики развития научно-технического знания в фондах электронных документов.

К сожалению, на момент написания статьи не удалось провести испытание метода на тестовых коллекциях, содержащих десятки или сотни тысяч документов, это планируется сделать в дальнейшем.

Литература

[1] [Алгоритм выделения псевдооснов М. Портера] [Электронный ресурс]. – Режим доступа: http://snowball.sourceforge.net, свободный.

[2] Губин М. В. Модели и методы представления текстового документа в системах информационного поиска / М. В. Губин // Научно-техническая информация. Сер. 1. – 2004. – №12. – С. 12-24.

[3] Гусарова Л. Проверка обоснованности кластерного решения / Л. Гусарова, И. Яцкив // Reliability and statistics in transportation and communication (RelStat'03). – Рига, 2004. – Т. 5, №2. – С.49-56.

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

[5] Поляков П.Ю., Плешко В.В. RCO на РОМИП 2006 // Труды четвертого российского семинара РОМИП'2006. (Суздаль, 19 октября 2006г.). – Санкт-Петербург: НУ ЦСИ, – 2006.

[6] Прикладная статистика: Классификация и снижение размерности: Справ. изд. / С. А. Айвазян, В. М. Бухштабер, И. С. Енюков, Л. Д. Мешалкин; Под. ред. С. А. Айвазяна. – М.: Финансы и статистика, 1989. – 607с.: ил.

[7] [Электронная библиотека по информационным технологиям CITFORUM] [Электронный ресурс]. – Режим доступа: http://www.citforum.ru, свободный.

[8] Bekkerman R., Allan J. Using Bigrams in Text Categorization [Electronic resource]. – 2003. – Access mode: www.cs.umass.edu/~ronb/papers/bigrams.pdf.

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

[10] Boutin F., Hascoët M. Cluster Validity Indices for Graph Partitioning // Proceedings of the Eight International Conference on Information Visualization (IV'04). IEEE – 2004.

[11] Dempster A. P. Maximum likelihood from incomplete data via the EM algorithm / A. P. Dempster, N. M. Laird, D. B. Rubin // Journal of the Royal Statistical Society. Series B. – 1977. – Vol. 39, No. 1. – P. 1-38.

[12] Ester M. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise / M. Ester, H.-P .Kriegel, J. Sander, X. Xu // Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96). – Portland, 1996. – P. 226-231.

[13] Halkidi M. On Clustering Validation Techniques / M. Halkidi, V. Batistakis, M. Vazirgiannis // Journal of Intelligent Information Systems, Kluwer Academic Publishers. Manufactured in The Netherlands. – 2001. – 17:2/3. – P. 107-145.

[14] Jain A. K. Data Clustering: A Review / A. K. Jain, M. N. Murty, P. J. Flynn // ACM Computing Surveys. – 1999. – Vol. 31, No. 3. – P. 264-323.

[15] Kohonen T. Self organization of a massive document collection / T. Kohonen, S. Kaski, K. Lagus, J. Salojärvi, J. Honkela, V. Paatero, A. Saarela // IEEE Transactions on neural networks. – 2000. – Vol. 11, No. 3. – P. 574-585.

[16] Kuo-Lung W., Miin-Shen Y. A cluster validity index for fuzzy clustering // Pattern Recognition Letters. – 2005. – Vol. 26. – P. 1275-1291.

[17] MacQueen J. B. Some Methods for classification and Analysis of Multivariate Observations // Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability. – Berkeley, 1967. – Vol. 1. – P. 281-297.

[18] Manning C. D., Schutze H. Foundations of statistical natural language processing. – Cambridge: MIT Press, 1999. – 620 p.

[19] Massey L. Evaluating quality of text clustering with ART1 // Proceedings of the International Joint Conference on Neural Networks. – Portland, 2003. – Vol. 2. – P. 1402-1407.

[20] Maulik U., Bandyopadhyay S. Performance Evaluation of Some Clustering Algorithms and Validity Indices // IEEE Transactions On Pattern Analysis And Machine Intelligence. – 2002. – Vol. 24, No. 12. – P. 1650-1654.

[21] Mendes M.E.S., Sacks L. Dynamic Knowledge Representation for e-Learning Applications // Proc. of the 2001 BISC International Workshop on Fuzzy

Logic and the Internet, FLINT'2001. – Berkeley, 2001. – P. 176-181.
[22] Salton G., Buckley C. Weighting approaches in automatic text retrieval // Information Processing and Management. – 1988. – Vol. 24(5). – P. 513-523.
[23] Sebastiani F. Machine Learning in Automated Text Categorization // ACM Computing Surveys. – 2002. – Vol. 34, No. 1. – 47 p.
[24] Stein B. On Cluster Validity and the Information Need of Users / B. Stein, S. M. zu Eissen, F. Wißbrock // 3rd IASTED Int. Conference on Artificial Intelligence and Applications: Proceedings of AIA 03. – Benalmadena, 2003. – P. 216-221.
[25] Yang Y., Pedersen J. O. A Comparative Study on Feature Selection in Text Categorization // The Fourteenth International Conference on Machine Learning: Proceedings of ICML'97. – San Francisco, 1997. – P. 412-420.
[26] Zamir O. E. Clustering Web Documents: A Phrase-Based Method for Grouping Search Engine Results [Electronic resource]. – 1999. – Access mode: http://turing.cs.washington.edu/papers/zamir_thesis.pdf.
[27] Zheng Xiao-Shen Algorithm of documents clustering based on minimum spanning tree / Zheng Xiao-Shen, He Pi-Lian, Tian Mei, Yuan Fu-Yong // International Conference on Machine Learning and Cybernetics. – Xi-an, 2003. – Vol. 1. – P. 199-203.

Automatic Full-text Documents Classifier Building

Peskova Olga V.

Developing information retrieval systems based on full-text document classification mechanism is discussed. Text representation approaches, text clustering methods and clustering validation techniques are described. Automatic full-text documents subjects heading building method is proposed. Method effectiveness is evaluated on two Russian document collections that contain different size and different subject documents.