УДК 025.4.03; 002.53:004.65

НОВЫЙ МЕТОД ПОИСКА НА ОСНОВЕ ИЕРАРХИЧЕСКОЙ КЛАСТЕРИЗАЦИИ ПО ОБЛАСТЯМ ТЕКСТОВЫХ ДОКУМЕНТОВ

Ф.В. Борисюк, В.И. Швецов

Нижегородский госуниверситет им. Н.И. Лобачевского

fedorvb@gmail.com

Поступила в редакцию 12.05.2009

Рассмотрены вопросы развития систем информационного поиска, усовершенствования классификационных и словарных информационно-поисковых систем (ИПС). Раскрыты общие аспекты построения системы нового поколения, в которой объединяются достоинства классификационных и словарных ИПС. Представлен новый подход к кластеризации текстовых документов, размещенных в сети Интернет, на основе предлагаемого в статье алгоритма «Иерархическая кластеризация по областям».

Ключевые слова: Интернет-поиск, кластеризация, дерево областей.

Введение

За последние годы потребность в информационных ресурсах сети Интернет значительно возросла. По данным исследования, опубликованного на конференции RIW-2008 (Russian Internet Week), к 2008 году количество пользователей сети Интернет составило 33% россиян в возрасте от 14 лет. Для сравнения: в 2000 году это число равнялось 3.6%. Объемы информации сети Интернет очень быстро растут. Большинство пользователей Интернета при поиске по сайтам используют поисковые системы. Послав запрос в поисковую систему, пользователь получает список результатов, соответствующих введенному запросу. Однако качество результатов поиска не всегда удовлетворяет пользователя, а количество выдаваемых результатов измеряется тысячами. Чем больше объем выдаваемой информации, тем труднее в ней разобраться. Таким образом, встает задача разработки более эффективных алгоритмов поиска в сети Интернет.

В настоящее время наиболее распространенными типами поисковых систем по сайтам сети Интернет являются словарные и классификационные ИПС. Эти системы обладают определенными достоинствами и недостатками.

Предлагаемая в данной работе модель поисковой системы объединяет достоинства классификационных и словарных ИПС. Основой предлагаемой системы является кластеризация текстовых документов, осуществляемая с помощью представленного в работе алгоритма иерархической кластеризации по областям (ИКО). Наибольшая степень эффективности алгоритма достигается на больших коллекциях текстовых документов. В статье представлены оценки сложности алгоритма ИКО и его сравнение с другими алгоритмами, показывающее его эффективность.

Структура представляемой системы предусматривает ее использование в качестве широкомасштабируемой поисковой системы, способной обрабатывать большие объемы данных и выдавать результаты с высоким качеством.

1. Обзор существующих типов поисковых систем

Все поисковые системы Интернета можно разбить на два класса. Рассмотрим общие характеристики каждого класса.

Классификационные информационно-поисковые системы. В классификационных ИПС используется иерархическая (древовидная) организация информации, которая называется классификатором. Разделы классификатора называются рубриками. Библиотечный аналог классификационной ИПС – систематический каталог. Классификатор разрабатывается и совершенствуется коллективом авторов. Затем его использует другой коллектив специалистов, называемых систематизаторами. Систематизаторы, зная классификатор, читают документы и приписывают им классификационные индексы, указывающие, каким разделам классификатора эти документы соответствуют.

Классическим примером классификационной ИПС является самый популярный во всем мире тематический каталог Yahoo (http://www.yahoo.com/). Каталог Yahoo представляет собой огромную базу данных Интернет-адресов сайтов самой различной тематики, представленную в виде иерархического дерева. Поиск документов осуществляется вручную путем навигации по каталогу.

Достоинства ИПС классификационного типа:

– Быстрый поиск сведений по определённой достаточно популярной и крупной теме. При выборе необходимой тематической секции каталога пользователь получает небольшой список адресов сайтов сети Интернет по выбранной теме.

– Каталог классификационной ИПС содержат качественную информацию, так как содержание сайтов проверено и оценено группой экспертов.

– Каждая секция каталога содержит сравнительно небольшое количество документов, что позволяет пользователю ориентироваться в них.

Недостатки ИПС классификационного типа:

– Каталог ИПС не может дать исчерпывающих сведений по определённой тематике в силу ограниченности по числу представленных в каталоге документов.

– Осуществляя поиск по каталогу, необходимо учитывать субъективность оценки разработчиков классификатора и систематизаторов, так как известно, что мнения различных экспертов могут не совпадать по одному и тому же вопросу.

– Из-за присутствия человеческого фактора при подготовке и поддержке каталога его обновление осуществляется несопоставимо медленно по сравнению со скоростью роста количества Интернет-ресурсов.

Словарные информационно-поисковые системы. В словарных ИПС используется поиск документов по ключевым словам. Алгоритм взаимодействия со словарной ИПС прост: пользователь на языке запросов выражает то, что он хочет найти, и, запустив запрос в поисковую систему, через несколько секунд получает список ссылок на документы, удовлетворяющие его запросу. Этот процесс выполняется быстро и не требует человеческого вмешательства, именно поэтому словарные ИПС приобрели широкое распространение.

С точки зрения внутренней организации словарная ИПС состоит из двух основных частей, как правило работающих параллельно [1]. Первая часть (индексирующий робот (*robot*) и паук (*spider*)) ответственна за индексирование Web-документов, а при помощи второй части (поисковая машина и Web-интерфейс) осуществляется поиск документов по индексу в соответствии с запросами пользователей. Словарная ИПС создает индекс (словарь из слов), встречающихся в документах Интернета. В этом индексе каждому слову (списку слов) будет соответствовать некоторый список документов, его содержащих.

Наиболее популярными словарными поисковыми системами в России сейчас являются Yandex.ru и Google.ru.

Достоинства ИПС словарного типа:

− Предоставляется широкий охват Web-ресурсов.

− При поддержке словарной ИПС не требуется дорогой ручной труд разработчиков классификатора и систематизаторов.

Недостатки ИПС словарного типа:

− Информация, представленная в результате ответа словарной ИПС на запрос пользователя, не всегда является достоверной и удовлетворяющей пользователя.

− Результат поиска содержит большое количество документов, в котором сложно ориентироваться.

Результаты анализа рассмотренных типов поисковых систем. Существующие классификационные ИПС за счет небольшого охвата Интернет-ресурсов теряют свою привлекательность. Влияние человеческого фактора при оценке принадлежности того или иного документа к определенной категории приводит к удорожанию поддержки классификационной ИПС. Точность поисковой выборки словарных ИПС не всегда является достоверной и удовлетворяющей пользователя. Также из-за большого количества документов в выборке пользователю сложно в ней ориентироваться.

Безусловным плюсом классификационных систем являются более качественные результаты поиска, а плюсом словарных ИПС, по сравнению с классификационными, – обширный охват сети Интернет и удешевление за счет отказа от использования труда систематизаторов.

Таким образом, возникает задача построения более совершенной поисковой системы, совмещающей в себе плюсы словарных и классификационных ИПС.

Предлагается следующий подход к построению новой поисковой системы:

1) Моделирование процессов кластеризации (группирование или скапливание однотипных

объектов) и классификации, выполняемых человеком, и внедрение их в поисковую систему. Применение кластеризации приведет к улучшению качества поисковых результатов.

2) Широкий охват Web-ресурсов, как и в словарных ИПС.

3. Модель предлагаемой поисковой системы

Рассмотрим основные аспекты представляемой поисковой системы.

Архитектура представляемой системы изображена на рис. 1.

При запуске системы ее компонент под названием «паук» получает URL-адрес Интернет-документа, который будет загружен, от компонента под название URL-контроллер.

Паук загружает документ и передает его в компонент под названием «индексатор», в котором производится выделение ключевых слов документа, их позиций в документе; также выделяются присутствующие в документе ссылки и пересылаются в URL контроллер. Индексатор передает выработанную информацию о документе в компонент кластеризации. Компонент кластеризации, используя алгоритм, который будет описан ниже, определяет область, в которую необходимо поместить документ, и сохраняет информацию о документе в хранилище. Хранилище представляет собой базу данных, содержащую информацию о характеристиках документов, а также характеристиках областей.

Процесс обработки URL-адресов, поступающих из URL-контроллера, является циклическим. Обработка останавливается, когда все имеющиеся URL-адреса обработаны.

Взаимодействие с пользователем осуществляется через Web-интерфейс. Поступающий от пользователя запрос обрабатывается поисковым механизмом. Используя ключевые слова запроса, поисковый механизм формирует результирующий список документов и отправляет их пользователю.

Выделение ключевых слов в документе. Ключевым словом называется слово в тексте, способное в совокупности с другими ключевыми словами представлять текст.

Применяемые техники выделения ключевых слов в документе:

– используются частотные характеристики слов:
  - алгоритм TF-IDF [2];
  - законы Ципфа [3];

– выделенные элементы текста имеют больший вес;

– отбрасываются стоп-слова (предлоги, союзы и т.д.).

4. Кластеризация текстовых документов

Кластеризация документов – одна из важнейших задач информационного поиска. Целью кластеризации является автоматическое выявление групп семантически похожих документов среди заданного множества документов. Следует отметить, что группы формируются на основе попарной схожести описаний документов и никакие характеристики этих групп не задаются заранее, в отличие от классификации докумен-

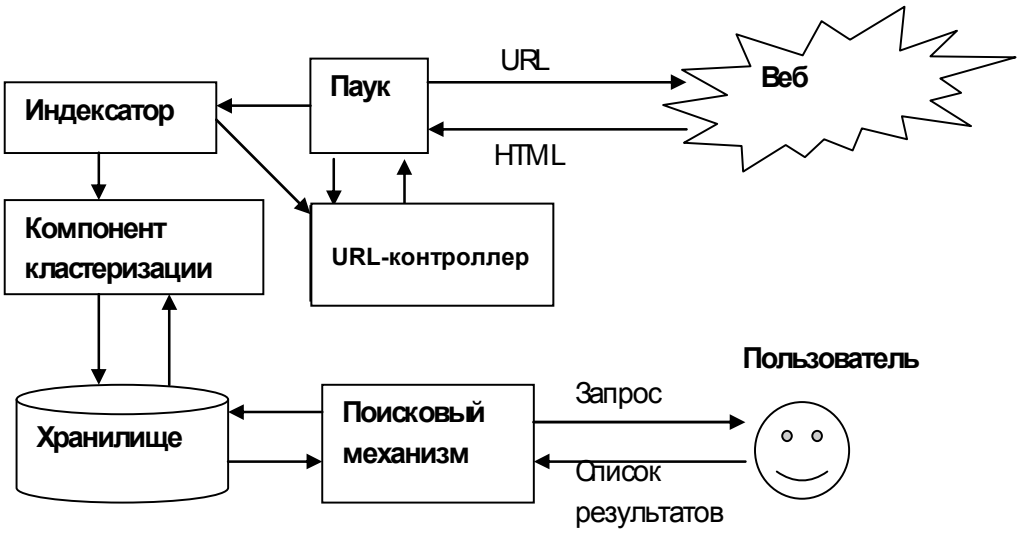

Рис. 1. Архитектура представляемой поисковой системы

тов, где категории задаются заранее. Объектами кластеризации являются текстовые документы. Каждый документ состоит из набора слов или словосочетаний, из которого можно выделить множество ключевых слов, представляющее документ. Каждое ключевое слово является элементарным признаком. Множество ключевых слов составляет пространство. В свою очередь, множество документов – это множество векторов этого пространства. Координатами вектора являются величины значимости (вес) каждого ключевого слова, рассчитанные для данного документа.

Обзор существующих подходов к кластеризации. Большинство существующих информационно-поисковых систем, использующих кластеризацию текстовых документов, рассматривает проблему кластеризации на ограниченной коллекции документов. Наиболее популярным подходом в последнее время стал STC (Suffix Tree Clustering) [4], используемый, например, в поисковых системах nigma.ru, vivisimo.com для группировки результатов поиска [5]. Кластеры образуются в узлах суффиксного дерева, которое строится из слов и фраз входных документов. STC обладает высокой скоростью работы, пропорциональной количеству документов $O(n)$. К недостаткам метода кластеризации с помощью суффиксных деревьев можно отнести то, что для него важен порядок слов в документе, к тому же метод способен работать только на ограниченной коллекции документов.

Среди классических методов кластеризации можно выделить метод K-means (K-средних) [6]. В его основе лежит итеративный процесс стабилизирования центроидов кластеров. Центроид вычисляется как усредненный вектор от всех элементов кластера. Вычислительная сложность K-means – $O(knT)$, где n – число документов, k – число кластеров, T – количество итераций. Для достижения хорошего качества кластеризации T может быть достаточно большим.

Среди методов, работающих с большими базами данных, можно выделить DBSCAN [7]. В методе применяется алгоритм, использующий плотность расположения объектов кластеризации. В алгоритме задействованы два параметра: Eps – радиус окрестности вокруг объекта и MinPts – минимальное количество объектов в кластере. Для того чтобы построить кластер, DBSCAN стартует с некоторого объекта и извлекает все объекты, которые расположены на расстоянии не больше чем Eps, далее по аналогии рекурсивно обрабатываются все попавшие в окрестность объекты. Вычислительная сложность DBSCAN $O(n^2)$.

Иерархическая кластеризация по областям. Для выполнения текстовой кластеризации в данной работе предлагается подход с использованием алгоритма ИКО. Алгоритм ИКО строит иерархическое дерево областей, состоящих из документов коллекции. Характеристики областей вычисляются во время работы алгоритма. Итоговыми кластерами являются узлы дерева областей.

Рассмотрим основные понятия представляемого подхода. Введем понятие значащей области. *Значащая область* – это область человеческой деятельности, искусства, науки, человеческих интересов, увлечений, которая имеет отношение к сделанному пользователем запросу, формально содержит запрашиваемую им информацию. Как известно, области знаний взаимосвязаны. Взаимосвязь значащих областей можно представить в виде иерархического дерева (рис. 2).

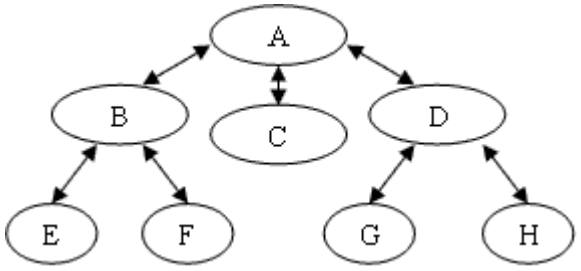

Рис. 2. Представление областей в виде иерархического дерева

Объекты, расположенные в узле иерархического дерева, являются наиболее близкими друг к другу. Буквами A, B, C, D, E, F, G, H обозначены значащие области дерева, в каждую из которых попали наиболее близкие друг к другу документы.

Перечислим характеристики значащей области:

1) Область описывают ключевые слова. Каждое ключевое слово имеет вес. Множество ключевых слов области состоит из ключевых слов документов, входящих в область.

2) К области принадлежат определенные документы.

3) Каждая область является узлом дерева и может иметь не более чем заданное число элементов. Обозначим его KMax.

4) Каждая область имеет не более чем заданное число потомков.

Под *близостью двух документов или областей* будем понимать сумму перемножений весов по всем ключевым словам, содержащимся в обоих документах или обеих областях. Каждое ключевое слово имеет вес, вычисляемый по алго-

ритму TF-IDF. Чем больше близость, тем более семантически похожими считаются объекты.

Инициализация алгоритма кластеризации. Пусть есть входящий поток документов, которые подлежат кластеризации. Каждый документ представляется вектором ключевых слов. Первоначально все поступающие документы попадают в корневой узел дерева областей до тех пор, пока количество документов не превысит заранее заданного предела KMax. При превышении предела корневая область разбивается на подобласти. Таким образом, на первом уровне дерева появляются потомки.

Фаза обработки входящего потока документов. В качестве исходных данных на этом этапе алгоритма выступают документ, представленный множеством ключевых слов, и дерево областей.

На первом шаге 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 установленного предела, который определяется как минимум близости между уже обработанными документами, то документ остается в первой области и временно не может быть встроен в дерево. Обозначим минимум близости *MinimumProximity*. 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 нему область. Документ помещается в найденную область. При превышении размером области определенного ограничения происходит разбиение ее на подобласти. Если количество подобластей превзошло определенное ограничение, то выполняется операция интеграции подобластей. Операция интеграции подобластей состоит из двух основных операций:

а) разбиение подобластей на две группы наиболее близких друг к другу;

б) объединение под единым началом одного из элементов группы других подобластей.

Рассмотрим алгоритм вставки документа в дерево областей. Введем обозначения:

а) RootArea – корневая область;

б) proximity(A,B) – вычисление близости между объектом A и B;

в) divide(Область) – разделение области на подобласти;

г) getChildren(Область) – построить список потомков области.

```
1 шаг. Поступил документ Doc.
2 шаг. areaList=getChildren(RootArea);
3 шаг. FOR EACH area IN areaList:
   Найти область Area максимально
   близкую к Doc.
4 шаг. Проверить можно ли вставить
   документ в дерево:
   IF (proximity (Area, Doc) <
      MinimumProximity)) { Root
      Area.Add (Doc);
   IF (RootArea.size() > KMax)
      Divide (RootArea);
   Конец алгоритма;
   }
5 шаг. areaList=getChildren(Area); IF
     areaList.size() == 0) GOTO
     8 шаг.
6 шаг. FOR EACH area IN areaList:
     Найти область NArea –
     максимально близкую к Doc.
7 шаг. IF (proximity (Area, Doc)
     <  proximity (NArea, Doc)) {
   Area = NArea; GOTO 5 шаг;
   } ELSE {  Area.add (Doc);
   IF (Area.size() > KMax) {
      divide (Area);
      IF (количество потомков Area
         превысило предел) {
         произвести интеграцию
         потомков;
         } GOTO 8 шаг.
  }
8 шаг. Обновить набор ключевых
   слов областей, которые
   составляют путь до
   результирующей области.
```

Анализ алгоритма кластеризации. Кластеризация документа заключается в том, что вновь поступивший документ встраивается в дерево областей. Особенностью алгоритма является введение проверки возможности корректной вставки в дерево обрабатываемого документа. Этот аспект алгоритма можно назвать «инкубатором», когда те документы, которые не соответствуют областям, составляющим дерево, остаются в первой области. Когда объем документов первой области перерастает установленный предел, то на первом уровне дерева появляются новые области, которые увеличивают диверсификацию областей дерева на первом уровне, что улучшает качество кластеризации.

Поиск. Пользователь посылает запрос в поисковую систему. Поисковая система заранее

проиндексировала (выделила ключевые слова и сохранила их в базе данных) документы и построила дерево областей. Запрос от пользователя может состоять как из нескольких ключевых слов, так и из целых предложений. Запрос пользователя преобразуется в вектор ключевых слов, из запроса удаляются стоп-слова. Поиск происходит по построенному дереву областей. Результирующие документы выдаются из области, наиболее близкой к запросу, и ее подобластей. Исходными данными для алгоритма поиска является вектор – elem.

Алгоритм поиска в дереве областей:

```
1. Построить список потомков
     корневой области:
     areaList=getChildren(RootArea);
2. FOR EACH area IN areaList:
     Положить в стек максимально
     близкие к elem области.
3. Извлечь элемент-область из
     стека, если он имеет потомков
     – перейти к шагу 1,
     в противном случае занести
     извлеченную из стека область
     в результат.
```

5. Практические результаты

Прототип представленной поисковой системы был реализован на языке Java с использованием системы управления базами данных MySQL. Алгоритм кластеризации был апробирован на сайтах сети Интернет с целью проверки точности кластеризации и поискового механизма. Ниже в таблице приведены результаты одного из экспериментов, проведенного на сайтах algolist.manual.ru и alglib.sources.ru, которые имеют близкую тематику. При тестировании сначала производилась обработка сайта algolist.manual.ru, на его основании происходило построение дерева областей. На втором шаге производилась обработка документов сайта alglib.sources.ru, документы сайта alglib.sources.ru встроились в подобласти по схожей тематике сайта algolist.manual.ru. Алгоритм ИКО показал более эффективные по количеству операций результаты по сравнению с методами K-means и DBSCAN. Алгоритм ИКО выполнил кластеризацию в 7.44 раза быстрее по сравнению с K-means, и в 23.141 раза быстрее по сравнению с DBSCAN (табл.). Практические результаты испытаний алгоритма ИКО показали, что теоретическая оценка соответствует количеству выполняемых операций. В качестве учитываемых операций выступают: измерения близости между двумя документами (или областями), количество операций, необходимое для обновления дерева областей. Количество обработанных документов: 705, KMax = = 20, максимальное количество подобластей – 6.

Таблица

Сравнение эффективности ИКО и алгоритмов K-means, DBSCAN

Алгоритм	Количество операций	Теоретическая оценка
ИКО	36512	$O(\mathrm{KMax}\cdot n\cdot \log_{\mathrm{KMax}}(n))$
K-means	271810	$O(kTn)$
DBSCAN	753032	$O(n^2)$

Выводы

В данной статье представлена модель новой поисковой системы, основанной на кластеризации текстовых документов. Предложен оригинальный алгоритм текстовой кластеризации ИКО. Приведенные результаты численного анализа алгоритма ИКО показывают, что он представляет эффективное по скорости и качеству решение задачи кластеризации документов для больших коллекций текстовых данных. Представленная модель ИПС совмещает в себе плюсы классификационных и словарных поисковых систем:

- предоставляет возможность широкого охвата Web-ресурсов;
- для поддержки системы не требуется дорогой ручной труд разработчиков классификатора и систематизаторов;
- система выдает в результатах поиска интересующую пользователя информацию.

Предполагается использование системы на широком круге Web-ресурсов.

Список литературы

1. Arasu A., Cho J., Garcia-Molina H., Paepcke A., Raghavan S. Searching the Web // ACM Transactions on Internet Technology (TOIT). 2001. V. 1. P. 2–43.

2. Kelleher D., Luz S. Automatic Hypertext Keyphrase Detection // Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence, Edinburgh, Scotland, UK. 2005. P. 1608–1610.

3. Li W. Random texts exhibit Zipf's-law-like word frequency distribution // Information Theory, IEEE Transactions. 1992. V. 38. P. 1842–1845.

4. Zamir O., Etzioni O. Grouper: a dynamic clustering interface to Web search results // Computer Networks. 1999. V. 31. N. 11–16. P. 1361–1374.

5. Eissen S.M., Stein B., Potthast M. The Suffix Tree Document Model Revisited // Proceedings of the 5th International Conference on Knowledge Management (I-KNOW 05). 2005. P. 596–603.

6. Arthur D., Vassilvitskii S. How Slow is the k-Means Method? // Proceedings of the Twenty-second annual symposium on Computational geometry table of contents, Sedona, Arizona, USA. 2006. P. 144–153.

7. Ester M., Kriegel H.-P., Jörg S., Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise // Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96). 1996. P. 226–231.

NEW SEARCH METHOD BASED ON HIERARCHICAL CLUSTERING BY AREAS OF TEXT DOCUMENTS

F.V. Borisyuk, V.I. Shvetsov

The development of information retrieval systems (IRS) and ways for improvement of classification and subject word IRSs are considered. General aspects are revealed in the design of a new generation IRS which combines the advantages of classification and subject word IRSs. A new approach to the clustering of the Internet text documents is proposed which is based on the algorithm «Hierarchical Clustering by Areas» presented in this article.

Keywords: internet-search, clustering, area tree.