Извлечение значимой информации из web-страниц для индексирования♣

© М.С. Агеев
Научно-исследовательский вычислительный центр МГУ им. М.В.Ломоносова, АНО Центр Информационных Исследований
ageev@mail.cir.ru

И.В. Вершинников
АНО Центр Информационных Исследований
vershin@mail.cir.ru

Б.В. Добров
Научно-исследовательский вычислительный центр МГУ им. М.В.Ломоносова, АНО Центр Информационных Исследований
dobroff@mail.cir.ru

Аннотация

В статье описываются разработанные нами методы разделения web-страниц на содержательную и навигационную части. Данные методы основаны на выделении одинаковых частей страниц с одного сайта. Основной целью данной работы является повышение качества информационного поиска за счет удаления навигационной части из web-страниц.

1 Введение

В настоящее время большинство сайтов интернет имеют сложное оформление и разрабатываются с использованием средств динамической генерации web-страниц. Оформление web-сайтов включает в себя единый дизайн страниц, средства навигации по сайту, логотипы и рекламную информацию.

Часть html-страницы, соответствующая оформлению (будем называть её *навигационной частью*) может существенно отличаться по тематике от основной (содержательной) части страницы. Это обстоятельство важно учитывать при разработке систем автоматической обработки текстов для информационного поиска.

Целью нашего исследования является разработка методов автоматического разделения html-страниц на навигационную и содержательную части. Мы полагаем, что исключение (понижение веса) навигационной части страниц из индекса поисковой машины может повысить качество информационного поиска.

В данной статье мы опишем разработанный нами алгоритм, методику практической оценки качества работы алгоритма и результаты оценки.

Труды 7ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2005, Ярославль, Россия, 2005.

2 Существующие подходы

Задача анализа структуры автоматически генерируемых html-страниц часто привлекает внимание исследователей в последние годы. Выделение структурной информации используется для различных задач обработки документов [6, 9], информационного поиска [7, 8] и для отслеживания изменений и кэширования web-страниц [10, 11].

Применяемые методы анализа можно разделить на

1) методы, основанные на анализе dom-дерева отдельной html-страницы [8, 9];
2) методы, основанные на выделении повторяющихся фрагментов страниц с одного сайта [10, 11, 7, 6, 9].

Стандартным способом формирования html-страниц является добавление навигационной части по краям страницы (содержательная часть находится посередине). Страница оформляется в виде таблицы или вложенных таблиц html, при этом отдельные элементы страницы вынесены в различные ячейки. Методы, основанные на анализе dom-дерева работают на основе анализа структуры html-страницы и выделяют содержательную часть страницы на основе эвристических предположений о "стандартных" способах оформления документов. Такие методы обычно сочетаются с методами, основанными на выделении повторяющихся фрагментов страниц одного сайта [9].

Методы, основанные на выделении повторяющихся фрагментов страниц с одного сайта работают на основе предположения о том, что у станиц одного сайта элементы оформления повторяются, а содержательная часть — разная.

Отметим также, что в работе [7] повторяющиеся фрагменты анализируются для разных сайтов с целью определения сообществ связанных между собой ресурсов.

Кроме методов выделения навигационной части web-страниц, для информационного поиска могут быть полезны методы ранжирования слов с учетом частоты встречаемости внутри одного сайта. Этот

метод предполагает понижение веса слов, которые встречаются на многих (большинстве) страниц сайта.

3 Предлагаемый алгоритм

В нашем исследовании мы использовали подход, основанный на выделении повторяющихся фрагментов страниц одного сайта. Предложен оригинальный алгоритм для решения данной задачи.

Описание алгоритма

На вход алгоритму подается директория с файлами, соответствующими страницам одного сайта. Алгоритм анализирует данные файлы и выделяет в них фрагменты, которые считаются навигационной частью. В зависимости от настроек, алгоритм либо удаляет навигационную часть из файлов, либо выделяет навигационную часть специальными тегами.

Алгоритм состоит из нескольких основных этапов:

1) Разбиение файлов на неделимые при сравнении последовательности символов — *токены*.
2) Поиск подмножества файлов (*кластер*), с одинаковым *набором цепочек токенов* (кластеры соответствуют подмножеству страниц с одинаковой навигационной частью). Набор цепочек токенов состоит из одной или нескольких цепочек (последовательностей подряд идущих) токенов. Желательно, чтобы набор цепочек был как можно длиннее и встречался в большом количестве файлов.
3) Удаление найденной последовательности цепочек токенов из всех файлов кластера.
4) Если кластер не покрывает всё множество файлов, то повторить шаги 2-4 для оставшихся документов.
5) Если остались нетронутые файлы (из которых ничего не было удалено), то в этих файлах производится поиск и удаление навигационных частей — цепочек токенов — из других кластеров.

Этап 1. Токенизация html-файла

Файл разбивается на неделимые при сравнении и выделении оформления последовательности символов — *токены*.

Токеном назовем

a) отдельный тег html;
b) текст между тегами.

Пробелы между тегом и текстом игнорируются.

Этап 2. Поиск кластера с одинаковой навигационной частью

Каждый файл представляется множеством *цепочек токенов*. Цепочка токенов — это последовательность подряд идущих токенов длины min_tokens_chain, где min_tokens_chain — параметр, задающий минимальную длину вырезаемого куска файла в токенах. Для каждой цепочки токенов длины min_tokens_chain вычисляется хэш-функция по алгоритму CRC32.

Таким образом, каждый файл отображается во множество чисел — хэш-значений цепочек токенов длины min_tokens_chain. Для каждой цепочки токенов длины min_tokens_chain запоминается длина исходного текста в байтах.

Построение кластера производится в несколько шагов:

1) Поиск первой пары документов в кластере.

Перебираются все пары документов сайта, для каждой пары вычисляется длина совпадающих цепочек токенов в байтах.

Если совпадение между файлами слишком большое (более 70% от общей длины файла), то эти файлы считаются дублями и не подходят для образования первой пары в кластере.

Пара файлов, для которых совпадение максимально, добавляются в кластер. Пересечение этих файлов (набор хэш-кодов цепочек токенов) — будем называть его *навигационной частью кластера* — запоминаем для дальнейших вычислений.

Определим порог min_nav_length равным 80% от длины навигационной части кластера в байтах. В дальнейшем при построении данного кластера навигационная часть будет уменьшаться, но не далее порога min_nav_length.

2) Поиск документов для добавления в кластер.

Производится поиск документа, который максимально пересекается с навигационной частью кластера. Файл добавляется в кластер если длина совпадения с навигационной частью кластера не меньше min_nav_length.

Навигационная часть кластера становится равной пересечению с добавленным кластером.

Этот шаг повторяется до тех пор, пока есть подходящие файлы для добавления в кластер.

Этап 3. Удаление навигационной части из всех документов кластера.

Если в построенном кластере оказалось не менее 4 документов, то кластер считается успешно построенным, из всех документов кластера удаляется "навигационная часть кластера" (общая часть всех файлов кластера).

Этап 4. Поиск других кластеров

Если после построения кластера осталось не менее 4 документов, то выполняются шаги 2-4 для оставшихся документов.

Если кластер построить не удаётся, то значение порога min_nav_length понижается до 40% с шагом 20% и выполняются шаги 2-4 с новым значением порога min_nav_length.

4 Оценка алгоритма

Для оценки качества работы алгоритма мы использовали следующие методы.

Во-первых, производился экспертный анализ результатов работы алгоритма. Эксперт анализировал документы, обработанные алгоритмом, с целью выяснить, действительно ли вырезанный фрагмент можно отнести к оформлению страницы.

Во-вторых, оценивалось влияние применения алгоритма на качество результатов поиска. Для этого использовалась коллекция и методика оценки Российского семинара по Оценке Методов Информационного Поиска 2004 года (РОМИП [4], http://romip.narod.ru) и информационно-поисковая система РОССИЯ [1, 3].

Экспертная оценка результатов работы алгоритма

Экспертная оценка алгоритма проводилась на двух коллекциях документов:

1) коллекции сайтов российских университетов;
2) коллекции РОМИП-WEB-narod.ru, используемой в рамках семинара РОМИП.

Коллекция сайтов российских университетов состоит из web-страниц, закачанных с сайтов нескольких российских университетов (МГУ, ГУ-ВШЭ, Амурский ГУ и др.) Коллекция была создана в рамках другого проекта.

Данная коллекция включает в себя более 35000 web-страниц, закачанных с сайтов различных подразделений (факультетов, кафедр, лабораторий и т.п.) данных университетов. Коллекция включает большое разнообразие вариантов оформления сайтов так как, несмотря на общность организаций, сайты подразделений создаются различными коллективами с большим разнообразием используемых технологий web-программирования.

Коллекция РОМИП-WEB-narod.ru (http://romip.narod.ru/ru/collections/narod.html) предоставлена компанией Яндекс (http://yandex.ru) для целей тестирования методов информационного поиска в рамках семинара РОМИП (http://romip.narod.ru). Коллекция состоит из более 700000 страниц, 22000 сайтов домена narod.ru.

Для наглядности отображения результатов программа выделения навигационной части выделяла вырезаемые части страниц цветом. Таким образом, эксперт мог визуально анализировать вырезанные части страниц в стандартном web-браузере. На рис. 1 показан пример страницы с выделенным оформлением. Серым цветом показана часть страницы, отнесённая программой к оформлению.

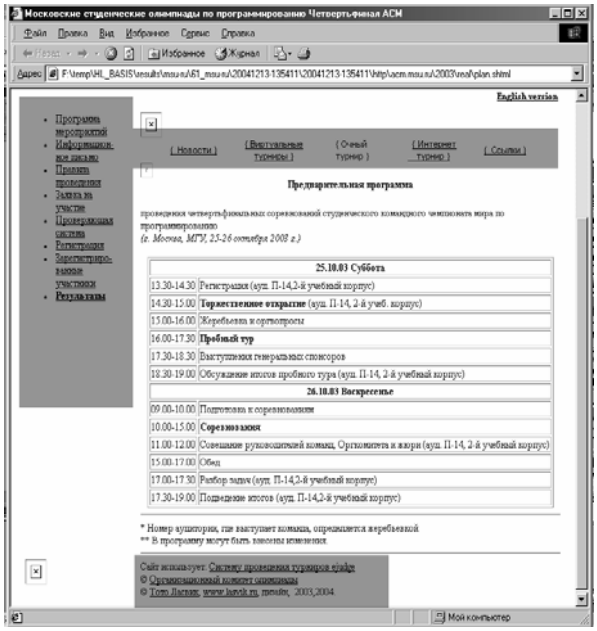

Рис. 1. Web-страница с выделенным оформлением. Программа выделила серым цветом навигационную часть. Исходная страница: http://acm.msu.ru/2003/real/

Для коллекции сайтов российских университетов экспертом было проанализировано 57 случайно выбранных сайтов и результатов обработки программой выделения навигационной части. Для коллекции РОМИП-WEB-narod.ru — 30 сайтов.

Для каждого проанализированного сайта эксперт выставлял оценку работы алгоритма от «2» (навигационная часть не выделена либо алгоритм «захватил» содержательную часть) до «5» (навигационная часть выделена правильно). Оценки «3» и «4» характеризовали промежуточные варианты, когда эксперт обнаруживал незначительные «промахи».

Результат оценки приведён в таблице 1.

оценка	Коллекция сайтов российских университетов		Коллекция РОМИП-WEB-narod.ru	
	Количество сайтов	Количество сайтов в %	Количество сайтов	Количество сайтов в %
2	4	7%	3	10%
3	3	5%	3	10%
4	10	18%	9	30%
5	40	70%	15	50%
Количество оцененных сайтов	57		30	
Средний балл	4.5		4	

Табл. 1. Результаты экспертной оценки работы программы

Оценка влияния разработанного алгоритма на качество информационного поиска: постановка эксперимента

Мы провели эксперимент по оценке влияния выделения оформления сайта на качество информационного поиска. Для данного

эксперимента мы использовали коллекцию документов

В рамках семинаров РОМИП 2003 и 2004 года [5, 4] участниками РОМИП были созданы матрицы оценки соответствия найденных документов запросам (для рассматриваемой коллекции было оценено 66 запросов). Для каждого документа, который был найден хотя бы одной системой, участвовавшей в РОМИП, асессоры РОМИП проставили экспертную оценку соответствия документов запросам.

Мы воспользовались данными матрицами релевантности для оценки качества работы поиска для разработанного метода. Мы использовали матрицу релевантности «or_relevant-minus_table.xml» 2004 года (то есть документ считается релевантным, если хотя бы один из асессоров поставил оценку «скорее релевантен» для данного документа).

В качестве базового метода поиска документов («отправной точки») использовался алгоритм поиска информационно-поисковой системы РОССИЯ [1, 3]. Этот алгоритм осуществляет поиск документов по словам, с учетом морфологии русского языка, с ранжированием документов по классической формуле tf*idf и учетом взаимного расстояния между словами. Подробное описание алгоритма поиска есть в [1].

Сравнивались два метода поиска документов:

1) поиск документов без усечения навигационной части (будем называть его **алгоритм «BasicLine»**);
2) поиск документов, обработанных программой отсечения навигационной части (будем называть его **алгоритм «CutNav»**).

Программа отсечения обрамления сайта применялась отдельно для каждого сайта из коллекции РОМИП-WEB-narod.ru — домена 3-го уровня в домене narod.ru. Применялись те же настройки алгоритма, что и при применении к коллекции сайтов российских университетов, описанной в предыдущем разделе.

Результаты эксперимента

На рис. 2 показан график качества поиска (измеряемый метрикой AvgPrecision [2]) для различных запросов. Запросы упорядочены в порядке убывания AvgPrecision для алгоритма «BasicLine».

Среднее значение AvgPrecision по всем запросам составляет:

- для алгоритма «BasicLine» — 33,22%
- для алгоритма «CutNav» — 32,26%

То есть после применения алгоритма отсечения обрамления в среднем качество поиска понизилось.

Дальнейший анализ результатов обработки некоторые причины ухудшения качества поиска.

Анализ «провалов» алгоритма

График на рисунке 2 показывает, что для некоторых запросов можно наблюдать повышение качества поиска, а для некоторых — наоборот, понижение качества поиска. Мы провели анализ работы алгоритма для запросов, на которых применение алгоритма отсечения обрамления привело к резкому ухудшению качества поиска. Для этого мы выбрали 5 запросов, для которых разность метрик AvgPrecision для алгоритмов «CutNav» и «BasicLine» наиболее высока.

Исходными предпосылками для такого исследования были следующие гипотезы о возможных причинах ухудшения качества поиска:

a) возможно, реализация алгоритма содержит ошибки, приведшие к вырезанию содержательной части страниц;
b) возможно, принцип работы алгоритма — выделение одинаковых частей страниц — приводит к вырезанию содержательной части сайта;
c) возможно, обрамление сайта содержит полезную информацию, относящуюся к запросу пользователя;
d) возможно, вынесенное асессором РОМИП решение о релевантности документа является спорным.

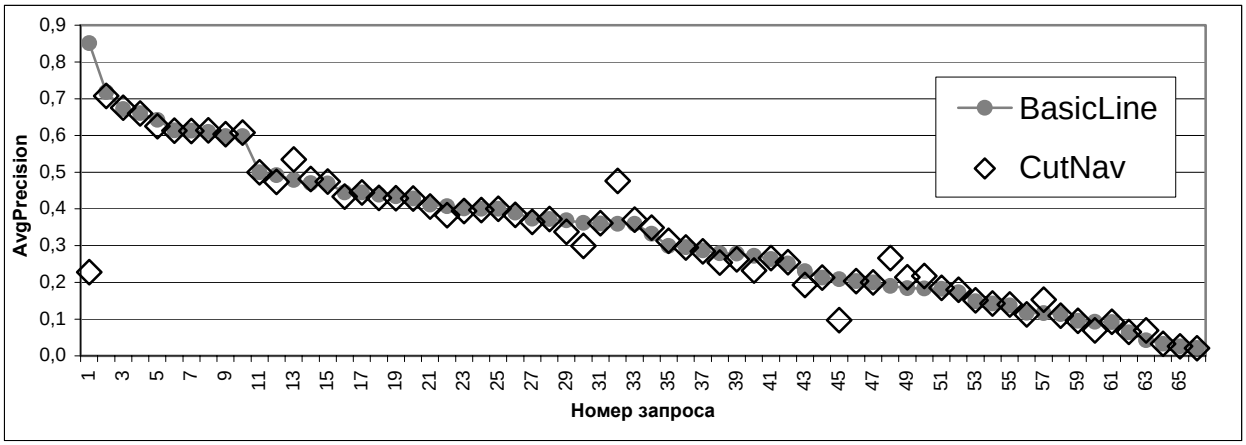

Рис. 2. Средняя точность для различных запросов для алгоритмов «BasicLine» и «CutNav».

Проанализируем запросы в порядке убывания разницы между результатом (средней точностью) алгоритма «BasicLine» и алгоритмом «CutNav».

Запрос №1 «телевизоры, домашний кинотеатр» (85,16%/ 22,82%)

По данному запросу средняя точность алгоритма «BasicLine» составляет 85,16%, алгоритма «CutNav» — 22,82%, 74 документа признаны релевантными. Основное расхождение результатов поиска произошло на документах сайта http://tvpanasonic.narod.ru. Алгоритм «BasicLine» нашел 51 релевантный (с точки зрения асессоров РОМИП) документ с этого сайта, а алгоритм «CutNav» эти документы не нашел.

Анализ показал, что документы имеют следующий вид (рис. 3): верхняя (основная часть) страницы содержит описание некоторой модели телевизора, а в нижней части находится длинный (3 экрана) список ключевых слов «оптимизирующих сайт», одинаковый для всех страниц сайта. Этот список ключевых слов содержит, в том числе, слова «домашний кинотеатр», «мелодии для телефонов panasonic», «автомагнитолы panasonic».

Алгоритм «CutNav» вырезал «список ключевых слов», так как он одинаковый для всех страниц сайта. Из-за этого документы не были найдены по словам «домашний кинотеатр».

На наш взгляд, лишь 3 из 51 документа соответствуют запросу «телевизоры, домашний кинотеатр». Три релевантных документа содержат описание широкоформатных плазменных панелей и находятся по запросу благодаря списку «оптимизирующих слов». Однако вопрос о том, насколько правильно вырезать такие списки «оптимизирующих слов», является спорным.

Рис. 3. Пример документа с сайта http://tvpanasonic.narod.ru, найденного по запросу «телевизоры, домашний кинотеатр». Пунктиром обозначена вырезанная часть страницы (вырезанный список продолжается вниз ещё на 3 экрана).

Запрос №2 «кинотеатр мечта» (20,91%/9,74%)

По данному запросу средняя точность алгоритма «BasicLine» составляет 20,91%, алгоритма «CutNav» — 9,74%, 32 документа признаны релевантными. Расхождение результатов поиска произошло на документах сайта http://gocinema.narod.ru/. Алгоритм «BasicLine» нашел 5 релевантных (с точки зрения асессоров РОМИП) документов с этого сайта, а алгоритм «CutNav» эти документы не нашел.

Данные документы (рис. 4) содержат навигационную часть — левый тулбар — в котором содержится список кинотеатров, в том числе «Мечта» (без ссылки на страницу с описанием кинотеатра «Мечта»). В основной части документа содержится описание другого кинотеатра (не «Мечта»). На наш взгляд, эти пять страниц нерелевантны запросу «кинотеатр мечта» (на страницах нет никакой информации о кинотеатре «Мечта» или других объектах, которые назывались бы «Мечта»).

На наш взгляд, алгоритм «CutNav» улучшил результат, но оценка асессора РОМИП неправильна.

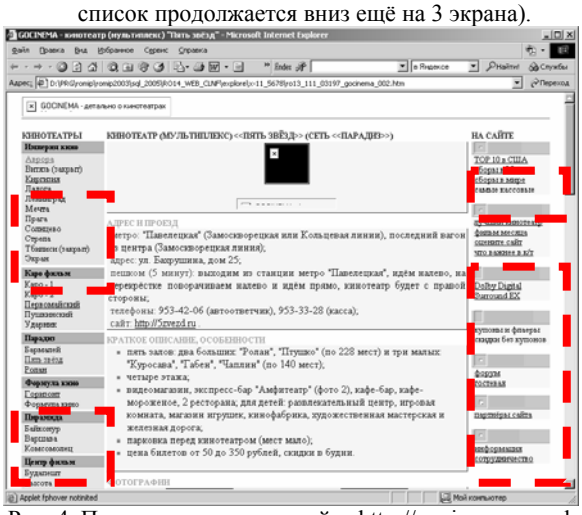

Рис. 4. Пример документа с сайта http://gocinema.narod.ru, найденного по запросу «кинотеатр мечта». Пунктиром обозначена вырезанная часть страницы.

Запрос №3 «что взять в роддом» (36,23%/29,91%)

По данному запросу средняя точность алгоритма «BasicLine» составляет *36,23%*, алгоритма «CutNav» — *29,91%*, 17 документа признаны релевантными.

Расхождение результатов (1 документ) произошло из-за ошибки обработки документов (алгоритм «BasicLine» некорректно обработал символы unicode, закодированные в виде Я).

Запрос №4 «с чего начать бизнес» (27,28%/23,27%)

По данному запросу средняя точность алгоритма «BasicLine» составляет *27,28%*, алгоритма «CutNav» — *23,27%*, 196 документов признаны релевантными.

Расхождение результатов произошло на семи документах с четырех различных сайтов.

Три релевантных документа с сайта http://business.narod.ru были найдены алгоритмом «BasicLine». Эти документы являются дублями, и поэтому алгоритм «CutNav» удалил всё содержимое документов (это ошибка алгоритма «CutNav»). Соответственно, документы не были найдены алгоритмом «CutNav».

Страница http://lolita.narod.ru/boys.htm содержит множество блоков со ссылками на другие страницы. В частности в левом тулбаре содержится (вполне релевантная) ссылка «Заработай деньги -> С чего начать». Данный документ найден алгоритмом «BasicLine». Алгоритм «CutNav» вырезал левый тулбар и данный документ не был найден.

Ещё три документа были найдены алгоритмом «BasicLine», но не были найдена алгоритмом «CutNav» из-за ошибки обработки документов (алгоритм «BasicLine» некорректно обработал символы unicode, закодированные в виде Я).

Запрос №5 «старинные русские праздники» (23,04%/19,31%)

По данному запросу средняя точность алгоритма «BasicLine» составляет *23,04%*, алгоритма «CutNav» — *19,31%*, 45 документов признаны релевантными. Расхождение результатов поиска произошло на двух документах сайта http://rusvarga.narod.ru.

Документ http://rusvarga.narod.ru/ts.htm (рис. 5) найден алгоритмом «BasicLine», но не найден алгоритмом «CutNav». Алгоритм «CutNav» правильно вырезал обрамление. Документ найден по словам, встретившимся в обрамлении. На наш взгляд, этот документ нерелевантен (содержимое документа ничего не говорит о праздниках).

Документ http://rusvarga.narod.ru/modernlit.htm найден алгоритмом «BasicLine», но не найден алгоритмом «CutNav», аналогично предыдущему документу.

На наш взгляд, алгоритм «CutNav» улучшил результат, но оценка асессора РОМИП неправильна, документы нерелевантны.

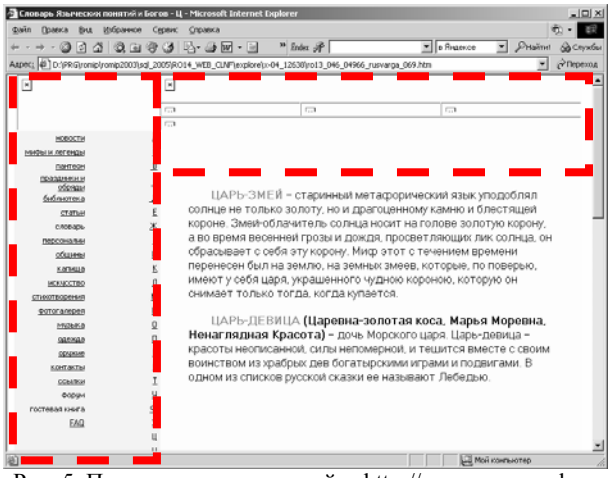

Рис. 5. Пример документа с сайта http://rusvarga.narod.ru, найденного по запросу «старинные русские праздники». Пунктиром обозначена вырезанная часть страницы.

Анализ «провалов» алгоритма: выводы

Мы проанализировали причины понижения оценок качества работы алгоритма на пяти сайтах, для которых ухудшение наиболее сильно выражено. Для остальных сайтов разница результатов несущественна (менее 3%), либо алгоритм «CutNav» улучшил результаты поиска.

Всего было проанализировано 66 страниц, которые были найдены алгоритмом «BasicLine» и признаны релевантными асессорами РОМИП, но не были найдены алгоритмом «CutNav». В табл. 2 сведена информация о причинах «провалов» алгоритма для проанализированных страниц. Причины провалов алгоритма «CutNav» классифицированы в соответствии с гипотезами, которые мы сформулировали вначале эксперимента (см. выше). Некоторые страницы отнесены сразу к нескольким категориям, поэтому сумма чисел во втором столбце больше 100%.

Кол-во страниц	**Кол-во страниц в %**	**Причина**
7	11%	Ошибка реализации алгоритма
5	8%	Обрамление невозможно точно выделить на основе анализа совпадающих частей страниц сайта
4	6%	Обрамление содержит релевантную информацию
55	83%	Мы не согласны с решением о релевантности, вынесенным асессором РОМИП

Табл. 2. Классификация причин «провалов» алгоритма для проанализированных страниц

Стоит отметить, что доля спорных оценок релевантности так высока (83%) потому, что выбранные для анализа документы являются «наиболее спорной частью» из всех найденных документов — эти документы были найдены по словам, находящимся в обрамлении документа.

Проведенный анализ позволяет сделать следующие выводы:

1) понижение оценки качества поиска для алгоритма «CutNav» на один процент связано, в основном, с проблемами оценки релевантности документов;
2) даже после осуществления тщательной проверки оценок релевантности не стоит ожидать существенного (более 5%) повышения оценок качества поиска.

5 Заключение

Мы представили алгоритм разделения web-страницы на навигационную и содержательную части. Экспертная оценка результатов работы алгоритма показала, что алгоритм в целом выполняет поставленную задачу.

Проведен эксперимент по оценке влияния отсечения обрамления страниц на качество информационного поиска в web. В связи с выявленными проблемами оценки релевантности страниц, точно оценить результаты эксперимента невозможно, однако можно утверждать, что даже после решения данных проблем не стоит ожидать существенного (более 5%) повышения оценок качества поиска.

Проведенный эксперимент выявил, что имеются случаи, когда оформление сайта содержит информацию, важную для поиска. Кроме того, есть случаи, когда оформление невозможно выявить только на основе анализа совпадающих частей страниц.

Мы полагаем, что применение алгоритмов, подобных описанному, может быть эффективно для узких задач информационного поиска, когда априори известно, что элементы оформления группы индексируемых ресурсов мешают информационному поиску.

Литература

[1] Агеев М.С., Добров Б.В., Лукашевич Н.В., Сидоров А.В. Экспериментальные алгоритмы поиска/классификации и сравнение с «basic line». // Российский семинар по Оценке Методов Информационного Поиска (РОМИП 2004) — Пущино, 2004. — стр. 62-89 http://romip.narod.ru/romip2004/05_uis_russia.pdf

[2] Агеев М.С., Кураленок И.Е. Официальные метрики РОМИП'2004. // Российский семинар по Оценке Методов Информационного Поиска (РОМИП 2004) — Пущино, 2004. http://romip.narod.ru/romip2004/appendix_a_metrics.pdf

[3] Журавлев С.В., Юдина Т.Н., Информационная система РОССИЯ // НТИ. Сер.2. — 1995. — № 3. — С.18-20.

[4] Кураленок И.Е., Некрестьянов И.С., Павлова Е.Ю. РОМИП 2003: Опыт организации. // Труды РОМИП'2003, октябрь 2003, — СПб: НИИ Химии СПбГУ — стр. 9-30. http://romip.narod.ru/romip2003/1_romip_overview.pdf

[5] Кураленок И.Е., Некрестьянов И.С. РОМИП'2004: отчет организаторов // Российский семинар по Оценке Методов Информационного Поиска (РОМИП 2004) — Пущино, 2004. http://romip.narod.ru/romip2004/01_romip_overview.pdf

[6] И. Некрестьянов, Е. Павлова. Обнаружение структурного подобия HTML-документов. // Труды четвертой всероссийской конференции RCDL'2002, 38-54, Дубна, Россия, 2002. http://meta.math.spbu.ru/~igor/papers/rcdl02-structure/rcdl02-structure.html

[7] Ziv Bar-Yossef, Sridhar Rajagopalan Template Detection via Data Mining and its Applications // In Proceedings of WWW2002, May 7-11, 2002, Honolulu, Hawaii, USA. http://www2002.org/CDROM/refereed/579/

[8] Soumen Chakrabarti. Integrating the Document Object Model with Hyperlinks for Enhanced Topic Distillation and Information Extraction // In Proceedings of WWW10, May 1-5, 2001, Hong Kong. http://www10.org/cdrom/papers/489/

[9] Suhit Gupta, Gail E Kaiser, Peter Grimm, Michael Chiang, Justin Starren, Automating Content Extraction of HTML Documents // World Wide Web Journal, January 2005 http://www.psl.cs.columbia.edu/crunch/WWWJ.pdf

[10] Lakshmish Ramaswamy, Arun Iyengar, Ling Liu, and Fred Douglis. Automatic Detection of Fragments in Dynamically Generated Web Pages // In Proceedings of the 13th International World Wide Web Conference (WWW2004), New York City, May 2004. http://www.research.ibm.com/people/i/iyengar/www2004.pdf

[11] Lakshmish Ramaswamy, Arun Iyengar, Ling Liu, and Fred Douglis. Automatic Fragment Detection in Dynamic Web Pages and its Impact on Caching // To appear in IEEE Transactions on Knowledge and Data Engineering (TKDE'05). http://www.research.ibm.com/people/i/iyengar/TKDEFragmentDetection.pdf

Automatic Extraction of Significant Part of Web Pages for Informational Retrieval

Mikhail S. Ageev, Igor V. Vershinnikov, and Boris V. Dobrov

We describe a new algorithm for automatic partition of web page onto navigational and main parts. The algorithm is based on extraction of common parts in web-pages from one web-site. Our basic supposition is that we can improve quality of information retrieval system by purging navigational part of web-pages.

♣ [i] Работа поддержана компанией Яндекс, грант №10294