

Извлечение значимой информации из web-страниц для задач информационного поиска*

М.С. Агеев^{1,2}, И.В. Вершинников², Б.В. Добров^{1,2}

¹ Научно-исследовательский вычислительный центр
МГУ им. М.В.Ломоносова

² АНО Центр информационных исследований
{ageev, vershin, 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.

Мы покажем, что разработанный алгоритм позволяет эффективно разделять страницы на содержательную и навигационную части.

С другой стороны, проведенный нами эксперимент по анализу эффективности информационного поиска в web на основе коллекции РОМИП-WEB-narod.ru с использованием разработанного алгоритма не привел к улучшению средней оценки точности поиска. В статье описан детальный анализ результатов информационного поиска с использованием описанного алгоритма.

* Работа поддержана компанией Яндекс, грант №102942

1 Введение

В настоящее время большинство сайтов Интернет имеют сложное оформление и разрабатываются с использованием средств динамической генерации web-страниц [7, 8]. 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*, или *обрамлением*), 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. Нашей гипотезой является предположение о том, что исключение (понижение веса) навигационной части страниц из индекса поисковой машины может быть полезно для решения задач информационного поиска.

В разделе 2 мы опишем существующие подходы к анализу шаблонов оформления web-страниц. В разделе 3 опишем предложенный нами алгоритм. В разделе 4 — методику практической оценки качества работы алгоритма и результаты оценки.

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].

В работе [7] показано, что существующая практика оформления сайтов с использованием средств динамической генерации web-страниц нарушает базовые предположения, на которых строятся современные поисковые системы в web. А именно, такими предположениями являются [7]:

- 1) предположение о релевантных ссылках (Relevant Linkage Principle) — гипертекстовые ссылки, установленные

- авторами страницы, указывают на авторитетные ресурсы по теме документа;
- 2) ресурсы, ссылающиеся друг на друга, тематически близки (Topical Unity Principle);
 - 3) текст, находящийся рядом с гиперссылкой, релевантен теме ресурса, на который идет ссылка (Lexical Affinity Principle).

В отличие от [7], мы рассматриваем, в первую очередь, тематическое содержание web-страниц и классический поиск документов без учета ссылочной структуры web. В этом случае становится возможным проведение оценки качества информационного поиска с использованием свободно доступных данных (РОМИП).

Применяемые методы анализа структуры web-страниц 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 [6, 7, 9, 10, 11].

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 страниц 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.

3 Предлагаемый алгоритм

В нашем исследовании мы использовали подход, основанный на выделении повторяющихся фрагментов страниц одного сайта. Нами предложен 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

- 1) отдельный тег html;
- 2) 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 (в экспериментах мы использовали `min_tokens_chain=6`). 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 Университетская Информационная Система РОССИЯ (<http://www.cir.ru> [1, 3]).

4.1 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 нами 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 с 22000 сайтов домена narod.ru.

Эксперт визуально анализировал вырезанные части страниц в стандартном web-браузере, используя подсветку вырезанной части страниц. На рис. 1 показан пример страницы с выделенным оформлением. Серым цветом показана часть страницы, отнесённая программой к оформлению.

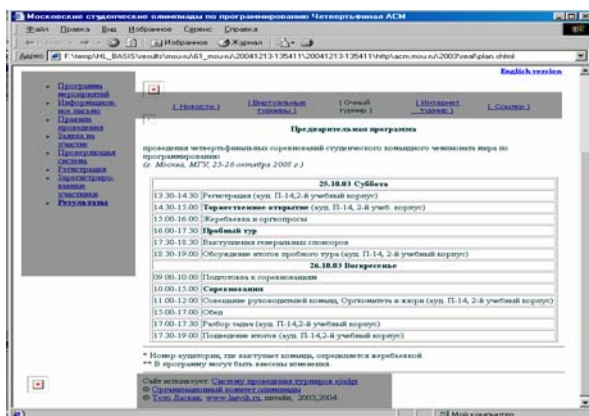


Рис. 1. Web-страница с выделенным оформлением. Программа выделила серым цветом навигационную часть. Исходная страница:

<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

4.2 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 РОМИП-WEB-narod.ru.

В рамках семинаров РОМИП 2003 и 2004 года [4, 5] 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, ассессоры РОМИП поставили экспертную оценку соответствия документов запросам.

Мы воспользовались данными матрицами релевантности для оценки качества работы поиска для разработанного метода. Мы использовали матрицу релевантности «og_relevant-minus_table.xml» 2004 года (то есть документ считается релевантным, если хотя бы один из ассессоров 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.

4.3 Результаты эксперимента

В качестве основной метрики оценки качества поиска мы использовали метрику «средняя точность» (AvgPrecision [2]). Данная метрика является комбинированной метрикой оценки полноты и точности запроса. Наглядно метрику AvgPrecision можно интерпретировать как площадь под графиком полноты/точности для данного запроса.

На рис. 2 показан график средней точности 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 выявил некоторые причины ухудшения качества поиска.

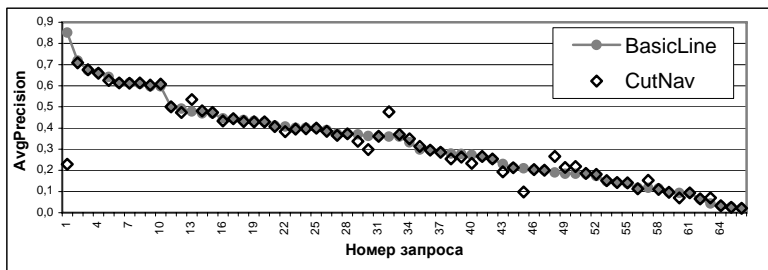


Рис. 2. Средняя точность для различных запросов для алгоритмов «BasicLine» и «CutNav»

4.4 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 «BasicLine» и «CutNav» наиболее высока.

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

- 1) возможно, текущая реализация алгоритма не учитывает некоторые варианты оформления web-страниц, что приводит к вырезанию содержательной части страниц;
- 2) возможно, принцип работы алгоритма — выделение одинаковых частей страниц — приводит к вырезанию содержательной части сайта;
- 3) возможно, обрамление сайта содержит полезную информацию, относящуюся к запросу пользователя;
- 4) возможно, вынесенное ассессором РОМИП решение о релевантности документа является спорным.

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

***Запрос №1 «телевизоры, домашний кинотеатр»
(BasicLine 85,16% / CutNav 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 ассессоров 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». Большинство пунктов этого списка не имеют никакого отношения к содержательной части страницы.

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 телевизоров и находятся по запросу благодаря списку «оптимизирующих слов». Ещё 25 документов содержат описания больших телевизоров (от 25 дюймов), поэтому их можно отнести к домашним кинотеатрам с большой натяжкой (на наш взгляд — скорее нерелевантны). Остальные 23 документа содержат описание обычных телевизоров небольшого размера — (на наш взгляд — однозначно нерелевантны).

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 ассессоров 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 ассессора РОМИП неправильна.

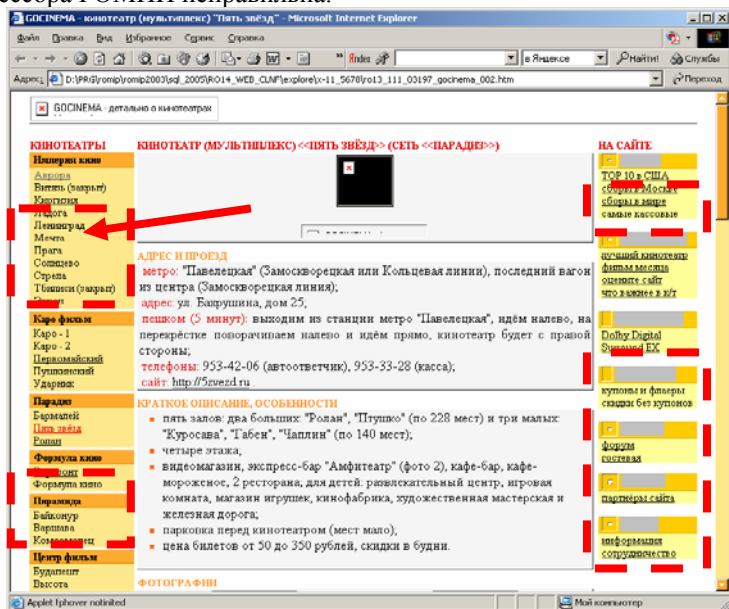


Рис. 4. Пример документа с сайта <http://gocinema.narod.ru>, найденного по запросу «кинотеатр мечта». Пунктиром отмечена вырезанная часть страницы.

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 документов признаны релевантными.

Расхождение результатов (1 документ) произошло из-за того, что алгоритм «BasicLine» некорректно обработал html-файл, закодированный в виде символов unicode вида 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://busi-ness.narod.ru> были найдены алгоритмом «BasicLine». Эти документы являются дублями, и поэтому алгоритм «CutNav» удалил всё содержимое документов. 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 найдены алгоритмом «CutNav» из-за того, что алгоритм «BasicLine» некорректно обработал html-файлы, закодированные в виде символов unicode вида 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>.

Документы <http://rusvarga.narod.ru/ts.htm> (рис. 5) и <http://rusvarga.narod.ru/modernlit.htm> найдены алгоритмом «BasicLine», но не найдены алгоритмом «CutNav». Алгоритм «CutNav» правильно вырезал обрамление. Документы найдены по словам, встретившимся в обрамлении. На наш взгляд, эти документы нерелевантны (содержимое документа ничего не говорит о праздниках).

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

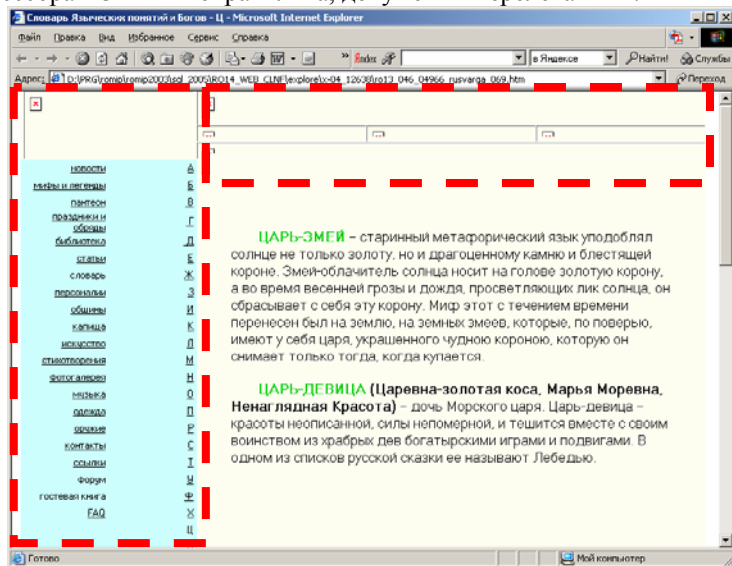


Рис. 5. Пример документа с сайта <http://rusvarga.narod.ru>, найденного по запросу «старинные русские праздники». Пунктиром обозначена вырезанная часть страницы.

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 незначительна (менее 3%), либо алгоритм «CutNav» улучшил результаты поиска.

Всего было проанализировано 66 страниц, которые были найдены алгоритмом «BasicLine» и признаны релевантными ассессорами 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%	Текущая реализация алгоритма не учитывает некоторые варианты оформления web-страниц, что приводит к вырезанию содержательной части страниц
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 ассессором 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 на данной коллекции документов.

Отметим, что в статье [7] был получен противоположный результат — отсечение оформления привело к улучшению качества информационного поиска. В отличие от нашего исследования, в [7] использовались:

- 1) другая задача поиска — поиск наиболее авторитетных страниц в web;
- 2) другой алгоритм поиска — использовался алгоритм ссылочного ранжирования HITS;
- 3) другая коллекция документов — документы собраны в web по списку из 37 запросов.

Действительно, оформление документов зачастую содержит множество ссылок, нерелевантных теме документа (зачастую большая часть ссылок документа находится в обрамлении). В нашем же случае текстовая часть обрамления вносит меньше шума, чем ссылки. На наш взгляд, данное различие результатов обусловлено условиями эксперимента.

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 на коллекции РОМИП-WEB-narod.ru. В связи с выявленными проблемами оценки релевантности страниц, точно оценить результаты эксперимента оказалось невозможно. Можно предположить, что даже после решения данных проблем не стоит ожидать существенного (более 5%) повышения оценок качества поиска в условиях отсутствия ограничений на тематику запросов.

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. В частности, такими задачами могут быть:

- 1) периодический мониторинг фиксированного списка сайтов;
- 2) поиск информации на сайтах форумов, блогов, web-конференций, которые имеют стандартную структуру.

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) — Пушино, 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) — Пушино, 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] Lakshmi Ramaswamy, Arun Iyengar, Ling Liu, and Fred Douglass. 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] Lakshmith Ramaswamy, Arun Iyengar, Ling Liu, and Fred Douglass. 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.

We show that our method successfully extract navigational and main parts from web-pages.

On the other hand, cutting of navigational part do not improve quality of web information retrieval.