УДК 004.738.52

ПОСТРОЕНИЕ И ПРИМЕНЕНИЕ МОДЕЛИ ОТСЕИВАНИЯ НЕРЕЛЕВАНТНЫХ ИСТОЧНИКОВ ПРИ ПОИСКЕ НАУЧНО-ТЕХНИЧЕСКОЙ ИНФОРМАЦИИ В ИНТЕРНЕТЕ

В.В. Зосимов

Николаевский национальный университет имени В.А. Сухомлинского

Описано построение и применение модели отсеивания нерелевантной информации в рамках решения задачи повышения эффективности поиска научно-технической информации в условиях присутствия в выдаче поисковых систем большого количества поискового спама и искусственно раскручиваемых сайтов. Рассмотрены основные группы сайтов, генерирующих поисковый спам. Показана эффективность построенной модели при отсеивании поискового спама.

Описано побудову і застосування моделі відсіювання нерелевантної інформації в рамках розв'язання задачі підвищення ефективності пошуку науково-технічної інформації в умовах присутності у видачі пошукових систем великої кількості пошукового спаму та штучно розкручуваних сайтів. Розглянуто основні групи сайтів, що генерують пошуковий спам. Показано ефективність побудованої моделі при відсіюванні пошукового спаму.

Введение

Одним из самых эффективных и востребованных видов рекламы в Интернете является поисковое продвижение сайтов. Активное продвижение коммерческих сайтов привело к тому, что пользователь, введя запрос в поисковой системе, получает в результатах выдачи большое количество коммерческих сайтов, которые релевантны введенному запросу, но не всегда релевантны потребностям пользователя. Наиболее сильно это затрудняет поиск научно-технической информации. Алгоритмы работы поисковых систем постоянно совершенствуются для борьбы с искусственным продвижением веб-ресурсов, но, несмотря на это, поиск релевантной информации становится все более сложной задачей.

Из всего объема информации, размещенной в сети Интернет, условно выделим два класса: коммерческую и научно-техническую информацию. Далее в статье под коммерческой информацией будем понимать информацию рекламного характера, представленную на сайте с целью привлечения новых покупателей, посетителей, подписчиков и т.д., и, как следствие, получение коммерческой выгоды.

Одним из возможных вариантов повышения эффективности поиска релевантной информации является разделение по некоторым заранее определенным признакам всего объема информации в Интернете на научно-техническую и коммерческую, выделение характерных признаков коммерческой информации и построение модели для ее эффективного отсеивания.

На сегодняшний день существуют различные методы, которые так или иначе пытаются оптимизировать информационный поиск. Весомый вклад в теорию и практику информационного поиска внесли М. Губин, И. Кураленок, А.В. Максаков, В.М. Рувинская, К.Л. Манукян, A. Barfourosh, S. Chakrabarti,

C. Manning, S. Meyer и другие украинские и зарубежные ученые. Однако все предложенные методы решают задачу информационного поиска как ряд не связанных друг с другом отдельных задач, и оказалось, что при решении одних задач ухудшаются показатели эффективности других.

Из сказанного следует, что разработка технологии повышения эффективности поиска научно-технической информации в сети Интернет на основе отсеивания нерелевантной информации в условиях постоянно изменяющихся характеристик всемирной паутины является актуальной проблемой.

Постановка задачи

В ходе анализа работы поисковых систем стало известно, что их алгоритмы, при поиске научно-технической информации, обладают низкими показателями точности. Причиной этому является большое количество поискового спама и искусственно продвигаемых сайтов в результатах поисковой выдачи.

Для устранения этой проблемы было принято решение разработать информационную технологию повышения эффективности поиска научно-технической информации путем отсеивания нерелевантной потребностям пользователя коммерческой информации на основе заранее определенных характерных признаков.

Предложенная в работе технология состоит из следующих этапов:

1. Поиск в Интернете веб-ресурсов, релевантных введенному поисковому запросу.

2. Отсеивание сайтов, содержащих нерелевантную ожиданиям пользователей коммерческую информацию. Отсеивание ведется на основе выделенных характерных признаков, позволяющих однозначно отнести сайт к категории коммерческих. Для ускорения работы необходимо создание в базе данных двух списков сайтов, содержащих научно-техническую и коммерческую информацию. В эти списки записываются все сайты согласно определенной на этапе отсеивания категории. Ускорение работы будет получено за счет того, что сайты перед анализом на наличие характерных признаков будут сверяться со списками из базы данных. Если сайт уже ранее был занесен в одни из списков, то он без дальнейшего анализа относится к указанной в списке категории.

3. Ранжирование результатов поиска, полученных в итоге работы на втором этапе с помощью классификатора, обученного на заранее заданной выборке [1]. Т.е. для научно-технической информации при поиске результатов по любому поисковому запросу, например «защита информации» или «кто такой аудитор» и т.п., веб-ресурсы будут ранжироваться согласно одному правилу классификации, найденному при помощи обобщенного интернационального алгоритма метода группового учета аргументов. Подобное разделение дает системе независимость алгоритма от процесса обучения по каждому запросу в отдельности, что значительно увеличивает быстродействие системы и дает возможность использовать ее в онлайн-режиме.

Цель данной работы: построение модели отсеивания нерелевантной коммерческой информации из результатов поиска по ключевым словам на

основе проведенного исследования по выявлению характерных признаков коммерческих сайтов. Построенная модель является частью технологии повышения эффективности поиска релевантной информации в Интернете.

Поисковый спам и искусственно «раскручиваемые» сайты

Поиско́вый спам (спамдексинг) — сайты и страницы в Интернете, созданные с целью манипуляции результатами поиска в поисковых машинах, в конечном счете, для обмана пользователя [2]. Для искусственной «раскрутки» сайтов, кроме поискового спама, который относится к так называемым «черным методам раскрутки» [3], используются еще и «белые» методы. Они основаны на том, чтобы максимально приблизить параметры сайта к идеальным с точки зрения поисковой системы. Обычно сайты оптимизируются под небольшое количество поисковых запросов, связанных с предлагаемыми на них товарами или услугами. В ходе «раскрутки» сайта общий вес его растет и он начинает появляться на первых строках в поисковой выдаче не только по тем запросам, под которые он оптимизирован, но и по тем, слова из которого встречаются в тексте сайта.

Поисковый спам используется для продвижения тех сайтов, которые, находясь на верхних позициях в поисковой выдаче, могут принести владельцу коммерческую выгоду. В основном это:

1. Сайты, с которых ведутся прямые продажи товаров, — интернет-магазины. Иногда в интернет-магазинах можно встретить разделы с информационными статьями. В основном эти статьи пишутся сторонними копирайтерами, которые не являются специалистами по теме статьи и всего лишь излагают материал, прочитанный в нескольких источниках. Иногда статьи просто копируются с сайтов производителей или сайтов конкурентов. Такие статьи не несут никакой новой информации, поэтому ими можно принебречь. То же касается и описания товаров.

2. Сайты фирм, предоставляющих товары и услуги. Здесь ситуация со статьями обстоит так же, как и в интернет-магазинах. Эти статьи пишутся с целью продвижения сайта и услуг, представленных на нем и носят ярко выраженный рекламный характер.

3. Различные сайты, позволяющие скачивать, что-либо за небольшую плату, будь то фильмы, музыка, книги или программное обеспечение.

4. Сайты, предоставляющие бесплатный доступ к информации, но взамен «заставляющие» просматривать рекламные объявления. (Информация на таких сайтах не бывает уникальной – она скопирована с каких либо других ресурсов).

5. Доски объявлений с предложениями о покупке, продаже, обмене, вакансиях и т.д.

До тех пор пока пользователь ищет информацию, представленную на вышеперечисленных сайтах, будь то товары или услуги, он не ощущает дискомфорта от присутствия в поисковой выдаче искусственно «раскрученных» сайтов. Это обусловлено тем, что при поиске коммерческой информации срабатывают стандартные правила бизнеса – чем больше средств вложено в рекламу, тем больше посетителей (потенциальных клиентов) получит сайт, что ведет к увеличению прибыли владельца сайта. Если же пользователь ищет научно-техническую информацию, то ему

приходится пересматривать и отфильтровывать десятки коммерческих сайтов в поисках релевантной информации.

Главная проблема, порождаемая поисковым спамом, заключается в том, что он генерирует множество мусорного контента, затрудняя эффективную работу поисковых серверов, искажает объективное ранжирование интернет-ресурсов и релевантность поисковых результатов. В итоге это во многом обесценивает Интернет как источник получения объективной информации.

Пути повышения релевантности поиска и борьбы с поисковым спамом

Основным способом борьбы поисковых систем с поисковым спамом на сегодняшний день является ряд фильтров, накладываемых поисковыми системами на сайты, использующие поисковый спам. Большая часть фильтров известна, и недобросовестные оптимизаторы неустанно ищут все новые способы их обойти.

Постоянное усовершенствование поисковых алгоритмов заставляет спаммеров искать новые обходные пути, на что поисковые системы отвечают очередным усовершенствованием алгоритмов. Такой подход является лишь временным решением и не позволяет решить проблему присутствия нерелевантной информации в поисковой выдаче.

Для повышения эффективности поиска научно-технической информации в статье предложен метод борьбы с поисковым спамом, основанный на отсеивании из выдачи поисковых систем сайтов, содержащих коммерческую информацию, как основной источник поискового спама. Для этого необходимо провести анализ коммерческих сайтов, по результатам которого будут выделены характерные признаки, для их идентификации.

После отсеивания необходимо провести повторное ранжирование оставшихся сайтов при помощи новой модели, полученной автоматически на основе обучающей выборки. Повторное ранжирование производится, потому что веса показателей релевантности, используемые в алгоритмах поисковых систем, выставлены с учетом наличия в ранжируемой выборке поискового спама. Ранжирование же результатов поиска, из которых был отфильтрован поисковый спам, требует новой модели с другими весами признаков релевантности [4].

Построение модели отсеивания

В ходе анализа работы поисковых систем было выявлено, что их алгоритмы, при поиске научно-технической информации, обладают низкими показателями точности. Причиной этого является большое количество поискового спама и искусственно продвигаемых сайтов в результатах поисковой выдачи.

Для устранения этой проблемы было принято решение разработать информационную технологию повышения эффективности поиска научно-технической информации путем отсеивания нерелевантной потребностям пользователя коммерческой информации на основе заранее определенных характерных признаков.

Выше были указаны четыре категории сайтов, которые генерируют основную часть поискового спама, а также искусственно «раскручиваются»:

1. Интернет-магазины.

2. Сайты фирм, предлагающих услуги либо товары.

3. Сайты, предоставляющие доступ к скачиванию информации за небольшую плату или просмотр рекламы.

4. Доски объявлений. В эту категорию входят как специализированные доски объявлений, предоставляющие возможность размещать предложения по определенной тематике, например автомобилей, вакансий или недвижимости, так и универсальные доски объявлений, разбитые на множество категорий и предоставляющие возможность размещать предложения в любую из них.

Для каждой из перечисленных выше категории сайтов в ходе анализа их содержимого были выделены присущие только ей характерные признаки. Они позволяют однозначно идентифицировать принадлежность сайта к той или иной категории. Подробнее ход исследования по выявлению характерных признаков коммерческих сайтов описан ниже.

Выявление характерных признаков коммерческих сайтов

В ходе анализа коммерческих сайтов для каждой из выше перечисленных категорий, генерирующих основную массу поискового спама экспериментальным путем был выявлен ряд характерных признаков, которые позволяют однозначно их идентифицировать.

Далее подробно рассмотрим эти признаки для каждой категории отдельно.

Характерные признаки интернет-магазинов. Одним из наиболее характерных признаков интернет-магазина является использование специализированной CMS (система управления содержимым сайта), разработанной для создания интернет-магазинов. Такие системы в большинстве своем платные и, в силу особенностей своей структуры, не подходят для создания информационного сайта. Поэтому все сайты, созданные на базе этих CMS, можно считать коммерческими (табл. 1).

Каждая из этих CMS оставляет свое название в мета-тегах, по которым их легко идентифицировать [5, 6]. Если же название по каким-либо причинам было удалено из мета-тегов, то можно использовать другие признаки. Каждая CMS использует свою уникальную структуру и названия папок, в которых хранятся их рабочие файлы. Поэтому для каждой CMS можно определить свой путь к Java-скриптам, а также к файлам, содержащим стили оформления дизайна. Изменение структуры готовой CMS для интернет-магазина очень трудоемкая задача, поэтому описанные выше признаки можно считать однозначным идентификатором интернет-магазинов.

Существует небольшой процент интернет-магазинов, разработанных на базе малоизвестных либо написанных под конкретный интернет-магазин CMS, которые не вошли в наше исследование. Для выделения характерных признаков таких магазинов было проведено дополнительное исследование.

Были отобраны 100 сайтов из каталога интернет-магазинов shop-list.com.ua. На основе анализа были выделены признаки, не зависящие от типа используемой CMS.

Анализу подлежали следующие структурные элементы: заголовки, метаописание, ключевые слова, текст на главной странице, элементы

навигации, наличие корзины покупателя, страницы каталога и страницы с товарами.

Таблица 1

Список готовых CMS, разработанных для создания интернет-магазинов

Платные CMS	Бесплатные CMS
1С-Битрикс: Управление сайтом	ShopCMS
Amiro.CMS	osCommerce
NetCat	Energine
OSG Интернет-магазин Enterprise	Virtuemart
Fast-Sales	TomatoCart
PHPShop	OpenCart
Melbis Shop	WP-shop
UMI.CMS	Drupal e-Commerce
HostCMS	Ubercart
Magento	PrestaShop
SiteEdit	Booot
Simpla CMS	
Smart Cart	
ABO CMS:Shop	
Shop-Script	
ImageCMS	
CreoShop	
VaM Shop	
АТИЛЕКТ.CMS	
Twilight CMS	
Astra.CMS	

Среди анализируемых сайтов у 94 % в заголовке, метаописании и ключевых словах присутствовали фразы с учетом словоформ: «интернет-магазин», «купить (название товара)», «продажа (название товара)» (рис. 1).

В текстах на главной странице у 93 % встречались фразы с учетом словоформ: «покупая (заказав, приобретая) в нашем магазине», «в нашем интернет-магазине», «наш интернет-магазин предлагает».

В элементах навигации (внутренние гиперссылки) у 95 % присутствовали пункты: «Доставка», «Оплата», «Доставка и оплата», «Оплата и доставка».

У 91 % имелась корзина покупателя.

На страницах каталога у 85 % рядом с картинкой товара встречаются слова: «Заказать», «В корзину», «Купить».

На страницах с информацией о товаре у 76 % рядом с большой картинкой встречаются описание товара (блок текста с не менее 100 символов) и его характеристики: цена, вес, объем, размер, артикул, номер товара.

Выявленные в результате исследования характерные признаки интернет-магазинов:

1) Использование специализированной CMS для создания интернет-магазина.

2) Наличие в заголовках, метаописании или ключевых словах фраз с учетом словоформ: «интернет-магазин», «купить (название товара)», «продажа (название товара)».

3) Присутствие среди элементов навигации пунктов: «Доставка», «Оплата», «Доставка и оплата», «Оплата и доставка».

4) Наличие корзины покупателя.

5) Присутствие в тексте на главной странице сайта фраз с учетом словоформ: «покупая (заказав, приобретая) в нашем магазине», «в нашем интернет-магазине», «наш интернет-магазин предлагает».

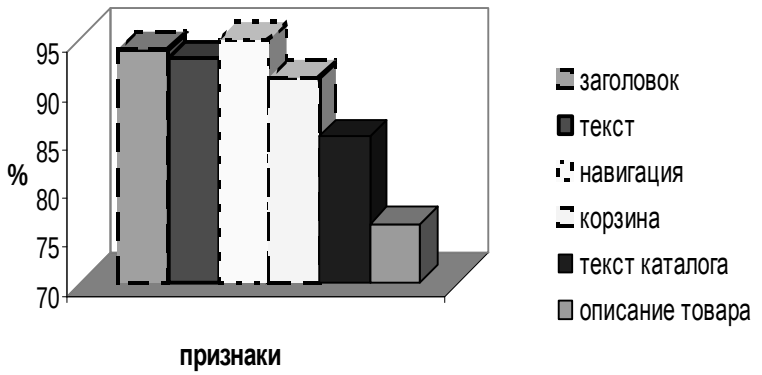

Рис. 1. Наличие у интернет-магазинов характерных признаков

6) Наличие на страницах рядом с картинкой слов: «Заказать», «В корзину», «Купить».

7) Наличие на странице информации о товаре рядом с большой картинкой блока описания товара (не менее 100 символов) и блока характеристик товара, таких как цена, вес, объем, размер, артикул, номер товара.

В предложенной технологии используются только пять первых признаков, так как они встречаются у большего процента интернет-магазинов и на их выявление требуется намного меньше времени, чем на выявление шестого и седьмого.

Характерные признаки сайтов для скачивания. Сайты, предлагающие скачивание информации, могут быть изготовлены на различных CMS, поэтому для нахождения характерных признаков необходимо анализировать их содержимое. Для исследования были отобраны 50 сайтов.

Анализу подлежали следующие элементы: заголовки сайта, метаописание, ключевые слова, элементы навигации, содержимое главной страницы (текстовая часть).

У 100 % сайтов в заголовках, метаописании, ключевых словах встречались словосочетания с учетом перестановки слов: «скачать бесплатно», «скачать фильмы», «скачать сериалы», «скачать софт», «скачать книги», «скачать музыку», «скачать игры», «скачать клипы». Такой высокий процент наличия этого признака обусловлен тем, что без него сайт будет трудно находим для посетителей и не принесет выгоды владельцу.

У 77 % в элементах навигации присутствовали словосочетания с учетом перестановки слов: «скачать фильмы», «скачать сериалы», «скачать софт», «скачать книги», «скачать музыку», «скачать игры», «скачать клипы».

У 85 % в тексте на главной странице присутствовали словосочетания с учетом перестановки слов: «скачать фильмы», «скачать софт», «скачать музыку» «скачать бесплатно», «скачать фильмы», «скачать сериалы»,

«скачать софт», «скачать книги», «скачать музыку», «скачать игры», «скачать клипы».

Рис. 2 иллюстрирует результаты исследования.

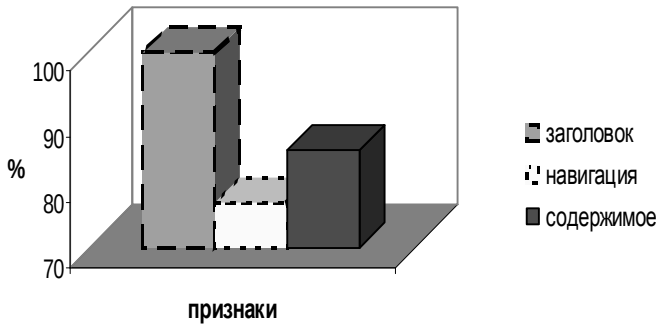

Рис. 2. Наличие характерных признаков на сайтах для скачивания

Выявленные в результате исследования характерные признаки сайтов для скачивания:

1) Наличие в заголовках, метаописании, ключевых словах сайта словосочетаний с учетом перестановки слов: «скачать бесплатно», «скачать фильмы», «скачать сериалы», «скачать софт», «скачать книги», «скачать музыку», «скачать игры», «скачать клипы».

2) Наличие тех же словосочетаний среди элементов навигации.

3) Наличие тех же словосочетаний в тексте на главной странице.

На рис. 2 видно, что наиболее точным признаком для идентификации сайтов для скачивания является первый – наличие характерных словосочетаний в заголовке, метаописании, ключевых словах. Кроме того, эти компоненты сайта легче всего анализировать, поэтому для ускорения идентификации в предложенной технологии будет использоваться только первый признак.

Характерные признаки сайтов фирм. Сайты фирм, как и сайты для скачивания, могут быть разработаны на различных CMS, поэтому для нахождения характерных признаков необходимо анализировать их содержимое.

В ходе исследования были проанализированы 100 сайтов различных фирм предлагающих услуги и товары, отобранных из каталога allkiev.com.ua.

Анализу подлежали следующие элементы сайтов: заголовки, метаописания, ключевые слова, элементы навигации, содержимое главной страницы.

У 64 % проанализированных сайтов в заголовках, мета-тегах, ключевых словах встречались следующие слова в сочетании с названием фирмы: «компания (название)», «фирма (название)», «предприятие (название)», «ООО («название»)».

В элементах навигации у сайтов фирм в 87 % встречались следующие пункты: «Услуги», «О компании», «О фирме», «Прайс лист», «Для дилеров», «Деятельность компании», «Работа в компании», «Заказать услугу», «Вакансии», «Цены», «Наши клиенты».

В тексте на главной странице сайта у 91 % встречаются фразы,

получаемые различными комбинациями сочетания слов из групп, представленных в табл. 2.

Таблица 2

Группы слов

Группа 1	Группа 2	Группа 3
компания	предлагает	услуги
фирма	предоставляет	обслуживание
корпорация	оказывает	*название товара*
ООО	осуществляет	
мы		
предприятие		

Примеры получаемых фраз: «компания предоставляет услуги», «фирма предлагает следующие услуги», «мы предлагаем холодильники», «предприятие осуществляет сервисное».

Рис. 3 иллюстрирует результаты исследования.

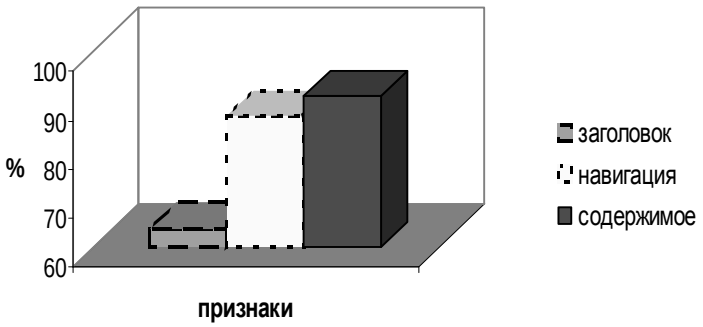

Рис. 3. Наличие характерных признаков на сайтах фирм

В ходе эксперимента выявлены следующие характерные признаки сайтов фирм:

1) Наличие в заголовках, метаописании, ключевых словах, словосочетаний следующего вида с учетом словоформ: «компания (название)», «фирма (название)», «предприятие (название)», «ООО («название»)», «корпорация (название)».

2) Наличие в элементах навигации следующих пунктов с учетом словоформ: «Услуги», «О компании», «О фирме», «Прайс лист», «Для дилеров», «Деятельность компании», «Работа в компании», «Заказать услугу», «Вакансии», «Цены», «Наши клиенты».

3) Наличие в тексте на главной странице фраз, получаемых сочетанием слов из вышеперечисленных групп с учетом словоформ и перестановки слов.

Для идентификации сайта фирмы необходимо анализировать все три элемента сайта. Выявив сходство по одному из них, можно относить сайт к коммерческим.

Характерные признаки досок объявлений. Доски объявлений могут быть разработаны на различных CMS, поэтому для нахождения характерных

признаков необходимо анализировать их содержимое.

В ходе исследования были проанализированы 100 сайтов досок объявлений, отобранных из каталога link.7do.ru.

Анализу подлежали следующие элементы сайтов: заголовки, метаописания, ключевые слова, элементы навигации, содержимое страниц (текстовое).

У 92 % проанализированных сайтов в заголовках, мета-тегах и ключевых словах встречались следующие словосочетания: «доска объявлений», «бесплатные объявления», «бизнес объявления», «вакансии (название города)» и, кроме того, фразы, получаемые сочетанием слов из представленных в табл. 3 двух групп с учетом словоформ.

Таблица 3

Группы слов

Группа 1	**Группа 2**
база	вакансия
каталог	резюме
создать	объявление
разместить	
загрузить	
опубликовать	
добавить	

В элементах навигации в 94 % встречались следующие пункты: «бесплатные объявления», «бизнес-объявления», «вакансии (название города)», а также фразы, получаемые сочетанием слов из представленных в табл. 3 двух групп с учетом словоформ.

В тексте на главной странице сайта у 93 % встречаются словосочетания: «доска объявлений», «бесплатные объявления», «бизнес-объявления», «вакансии (название города)» и, кроме того, фразы, получаемые различными комбинациями сочетания слов из представленных в табл. 3 двух групп с учетом словоформ.

Примеры получаемых фраз: «база вакансий», «создать резюме», «опубликовать объявление», «добавить вакансию».

Рис. 4 иллюстрирует результаты исследования.

В ходе эксперимента выявлены следующие характерные признаки досок объявлений:

1) Наличие в заголовках, метаописаниях, ключевых словах словосочетаний: «доска объявлений», «бесплатные объявления», «бизнес-объявления», «вакансии (название города)» и, кроме того, фразы, получаемые сочетанием слов из двух представленных в табл. 3 групп с учетом словоформ.

2) Наличие в элементах навигации следующих пунктов с учетом словоформ: «бесплатные объявления», «бизнес объявления», «вакансии (название города)», а также фразы, получаемые из представленных в табл. 3 двух групп слов с учетом словоформ.

3) Наличие на главной странице словосочетаний с учетом словоформ: «доска объявлений», «бесплатные объявления», «бизнес объявления»,

«вакансии (название города)» и, кроме того, фразы, получаемые различными комбинациями сочетания слов из представленных в табл. 3 двух групп с учетом словоформ.

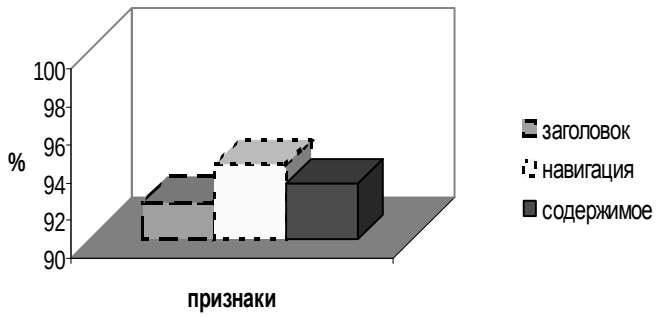

Рис. 4. Наличие характерных признаков на досках объявлений

Для идентификации доски объявлений необходимо анализировать все три элемента сайта. Выявив сходство по одному из них, можно относить сайт к коммерческим.

Отсеивание коммерческой информации

На этапе отсеивания коммерческой информации модель классификации на выбранном множестве признаков представлена в виде набора правил принятия решений. Для отсеивания коммерческой информации применяется ДНФ классификатор, в котором для категории C «коммерческая информация» в ходе исследования был выделен ряд характерных признаков $\{a_1^C,\ldots,a_n^C\}$ (где n — количество признаков) и набор структурных элементов сайта $\{b_1^C,\ldots,b_m^C\}$ (где m — количество структурных элементов сайта), содержащих эти признаки.

Все признаки $\{a_1^C,\ldots,a_n^C\}$ описываются с помощью регулярных выражений [7].

Классификатор строится следующим образом:

ЕСЛИ ((a_1^C И b_1^C) **ИЛИ**

(a_2^C И b_1^C) **ИЛИ**

…

(a_n^C И b_1^C) **ИЛИ**

...

(a_1^C И b_m^C) **ИЛИ**

(a_2^C И b_m^C) **ИЛИ**

…

(a_n^C И b_m^C))

ТО Коммерческая информация

ИНАЧЕ НЕ Коммерческая

Во время проведения экспериментов по выявлению характерных признаков сайтов, генерирующих поисковый спам, анализировался ряд структурных элементов сайтов. Эти же элементы анализируются при отсеивании коммерческой информации из полученного на этапе сбора данных списка сайтов.

Список анализируемых структурных элементов сайтов $\{b_1^C, \ldots, b_m^C\}$:

— мета-теги, пути к Java-скриптам и стилям оформления;

— заголовки, метаописание, ключевые слова;

— текст на главной странице;

— элементы навигации.

Представленные выше структурные элементы определяются в ходе анализа главной страницы сайта по характерным HTML-тегам:

— значения мета-тегов извлекается из атрибута content=" " тега <meta>;

— пути к Java-скриптам содержатся в атрибуте src=" " тега <script>;

— пути к файлам стилей содержатся в атрибуте href=" " тега <link>;

— заголовки сайта заключены в теги <title> … </title>;

— метаописания содержатся в атрибуте content=" " тега <meta name="description">;

— ключевые слова содержатся в атрибуте content=" " тега <meta name="keywords">;

— для анализа текста главной страницы берется весь текст, заключенный между тегами <body> … </body> за исключением текста, заключенного в теги <a> … (элементы навигации);

— элементы навигации заключены в тегах <a> … .

Для анализа из представленных выше тегов извлекается текст и сохраняется в виде строк в отдельный массив, который принимает в качестве параметров пары ключ => значение. Ключами являются названия структурных элементов, а значениями — их текстовое содержимое. В табл. 4 приведен пример содержимого такого массива для сайта seyf.dp.ua, найденного на первой странице поисковой выдачи Google по запросу «Защита информации».

Все выделенные признаки коммерческих сайтов разбиваются на группы соответственно структурным элементам, в которых они могут находиться. Каждая группа признаков записывается в массив, ключами которого являются натуральные числа, начиная с нуля, а значениями – регулярные выражения, описывающие эти признаки. В табл. 5 представлен массив признаков для структурного элемента *title*.

Анализ сайтов происходит в следующем порядке:

1) Проверка доменного имени на наличие в списках с коммерческой и научно-технической информацией.

2) Мета-теги анализируются на наличие в них названий распространенных CMS для разработки интернет-магазинов.

3) Проверяются пути к Java-скриптам и файлам со стилями оформления на соответствие с CMS интернет-магазинов.

4) Проверяется наличие корзины покупателя.

5) Проверяются заголовок, метаописание, ключевые слова на наличие

характерных признаков для всех категорий коммерческих сайтов.

6) Проверяются элементы навигации на наличие пунктов, характерных для всех категорий коммерческих сайтов.

7) Проверяются тексты главной страницы на наличие характерных признаков для всех категорий коммерческих сайтов.

Таблица 4

Содержимое массива структурных элементов

Ключ элемента массива	**Значение элемента массива**
meta	Romanchuk Larisa Интернет-магазин сейфов в Днепропетровске index,follow 48.479671;35.025798 48.479671,35.025798
script	/misc/drupal.js /modules/!thickbox-5.x-1.x-dev/thickbox/thickbox_login.js /modules/!thickbox-5.x-1.x-dev/thickbox/thickbox.js /modules/jquery_update-5.x-2.0/jquery_update/compat.js /modules/ubercart-5.x-1.6/ubercart/uc_cart/uc_cart_block.js?02476 /sites/all/modules/seyf_project/js/user_search.js /sites/all/modules/seyf_project/js/user_login.js
link	/files/favicon.ico
title	Каталог товаров (Сейфы) | интернет-магазин сейфов в Днепропетровске
description	Купить сейфы в Днепропетровске. Большой выбор сейфов с описаниями, актуальными ценами, фото. Доставка по Украине к двери клиента недорого и быстро. На всю продукцию мы даем гарантию, осуществляем гарантийное и послегарантийное обслуживание
keywords	Сейф, сейфы, сейфы днепропетровск, сейфы в днепропетровске, магазин сейфов в днепропетровске, сейфи, купить сейф, купить сейф днепропетровск, купить сейф в днепропетровске, сейфы украина
body	Интернет-магазин сейфов в Днепропетровске (056) 7166650 (067) 5399504 Доставка по украине! Гарантия. Каталог товаров (Сейфы) Добро пожаловать в наш каталог Сейфы Днепропетровск! Вы можете купить сейф в Днепропетровске у нас в офисе или забрать со склада. Также можете купить сейф с доставкой по Украине к Двери. Звоните, спрашивайте, заказывайте! Чтобы узнать больше о том, как выбрать сейф прочтите статью. Инструкции и сертификаты находятся в разделе Карта проезда и режим работы в разделе. Не забудьте оставить отзыв о работе менеджеров и службы доставки. В 21 веке под определением сейфы мы понимаем незаменимый и надежный уголок интерьера, который нужен для надежной охраны от доступа к нашим ценностям в каждом доме. У наших клиентов ценность приобретает собственное значение. Сейф может быть оригинальным подарком на день рожденье или знаменательное событие, такой подарок не изнашивается, не стареет и хранит память о празднике и тех кто такой подарок подарил. Информация по доставке. Вы знаете, что можно с доставкой к двери в и по областям в Украине. Подписка на новости. Корзина покупок. Ваша корзина пуста. Каталог товаров (Сейфы). Навигация. Рекомендуем. Сейфы по брендам. Вход в систему. Подписка на новости. Будьте в курсе последних новостей нашего сайта!
a	Металлическая Мебель, Теплый пол, Биокамины, Доставка и услуги, Информация, Контакты

Представленные выше элементы анализируются на наличие признаков сразу всех выявленных категорий коммерческих сайтов. При выявлении совпадения на каком-либо этапе, проверка прекращается и сайт помечается, как коммерческий, без определения к какой именно категории коммерческих сайтов он принадлежит. Разделение коммерческих сайтов на категории было

введено для облегчения поиска характерных признаков, а для отсева достаточно определить коммерческий сайт или нет.

Сайты, при проверке которых не было выявлено соответствий ни на одном этапе, сохраняются в отдельный список для дальнейшего ранжирования.

Для ускорения дальнейшей работы программы, сайты, помеченные как коммерческие, добавляются в список коммерческих сайтов. Сайты, помеченные как некоммерческие, помещаются в список, содержащий сайты с научно-технической информацией. В дальнейшем, при проведении отсеивания, сайты, помещенные в один из списков, не подлежат последующему анализу и либо сразу отбрасываются, либо проходят на этап ранжирования. С ростом объема этих списков возрастает и скорость работы программы, так как все больше сайтов проходят этап отсеивания без анализа содержимого.

Таблица 5

Содержимое массива характерных признаков для заголовков сайта

Ключ элемента массива	Значение элемента массива
0	интернет-магазин*
1	купить /b*
2	продажа /b*
…	…
8	скачать книг*
9	скачать музыку
10	скачать игр*
…	…
14	предприятие /b*
15	ООО /b*
…	…
19	доск* объявлений
20	вакансии /b*
…	…
26	баз* резюме
27	загруз* объявлени*
28	каталог* ваканси*
29	добав* резюме

Результаты тестирования эффективности отсеивания нерелевантной информации в двух государственных учреждениях

В Николаевском национальном университете им. В.А. Сухомлинского и Украинском радиотехническом институте был внедрен программный комплекс разработанный для повышения эффективности поиска научно-технической информации для обучения и ведения исследований. В нем использована модель отсеивания нерелевантной информации. На основе результатов внедрения анализировались эффективность отсеивания нерелевантной информации и ранжирования результатов поиска.

Результаты применения разработанной технологии для отсеивания нерелевантной информации из поисковой выдачи представлены в табл. 6.

Из табл. 6 видно, что построенная модель отсеивания нерелевантной информации позволяет значительно повысить эффективность поиска целевой информации в Интернете.

Таблица 6

Результаты работы модели отсеивания [1]

Название учреждения	Показатели точности поиска, %		Неотсеянные коммерческие сайты, %	Ошибочно отсеянные релевантные сайты, %
	Поисковая система	Программный комплекс		
УРТИ	83	97	3	0,4
ННУ	54	85	8	0,5

Выводы

В ходе анализа содержимого коммерческих сайтов был выявлен ряд характерных признаков, позволяющий однозначно их идентифицировать. На основе выявленных признаков построена новая модель автоматической классификации информации на релевантную и коммерческую.

Разработанная новая модель автоматической классификации информации на релевантную и коммерческую по ряду установленных характерных признаков позволяет повысить процент содержания релевантной информации в поисковой выдаче до 83–92 %.

Описанная в статье модель может быть применена не только для повышения эффективности поиска путем отсеивания нерелевантной информации, но и в качестве фильтра, блокирующего поиск товаров и услуг, т.е. предотвращающего нецелевое использования рабочего времени Интернета в личных целях. Особенно полезным это свойство технологии будет в офисах коммерческих структур, где для борьбы с нецелевым использованием Интернета существуют отдельные администраторы.

1. Zosimov V., Stepashko V., Bulgakova O. Enhanced technology of efficient Internet retrieval for relevant information using inductive processing of search results. *Artificial Intelligence Methods and Techniques for Business and Engineering Applications*. Rzeszow, Poland; Sofia, Bulgaria, ITHEA Publ., 2012, 345, pp. 99–112.
2. Энж Э. SEO — искусство раскрутки сайтов / Э. Энж, С. Спенсер, Р. Фишкин, Д. Стрикчиола. — СПб. : БХВ-Петербург, 2011. — 592 с.
   Enzh E., Spencer S., Fishkin R., Strikchiola D. *SEO — the art of site promotion*. St. Petersburg, BHV-Petersburg, 2011. 592 p.
3. Евдокимов Н. Раскрутка веб-сайта. Практическое руководство / Н. Евдокимов, И. Лебединский. – М. : Вильямс, 2011. – 288 р.
   Evdokimov N., Lebedinsky I. *Website promotion. A practical guide*. Moscow, Williams Publ., 2011. 288 p.
4. Зосимов В.В. Построение и сравнительный анализ моделей ранжирования результатов работы поисковых систем : Google и Яндекс / В.В. Зосимов, В.С. Степашко, О.С. Булгакова // Індуктивне моделювання складних систем. зб. праць. Вип. 3. — К. : МННЦ ІТС, 2011. — С. 90–95.
   Zosimov V., Stepashko V., Bulgakov O. Construction and comparative analysis of models ranking for the results of Google and Yandex search ngines. Inductive modeling of complex systems, Issue 3, Kyiv, pp. 90–95.
5. Рейтинг CMS Интернет-магазинов [Электронный ресурс]. — Режим доступа: http://www.ratingruneta.ru/cms/shop/.
   Rating of online stores CMS. Available at: http://www.ratingruneta.ru/cms/shop/
6. Гусев В.С. *Яндекс. Эффективный поиск* / В.С. Гусев. — М. : Вильямс, 2007. — 224 с.
   Gusev V. *Yandex. Efficient search*. Moscow, Williams Publ., 2007. 224 p.
7. Фридл Д. Регулярные выражения / Д. Фридл. — М. : Символ-Плюс, 2008. — 608 с.
   Friedl D. *Regular expressions*. Moscow, Symbol-Plus Publ., 2008. 608 p.

V.V. Zosimov

CONSTRUCTION AND APPLICATION OF A MODEL FOR SIFTING OF IRRELEVANT SOURCES AT THE RETRIEVAL OF THE SCIENTIFIC AND TECHNICAL INFORMATION ON THE INTERNET

Introduction: One of the most effective and popular forms of advertising on the Internet is the search engines optimization. Active promotion of commercial sites led to the fact that the user by typing a query into a search engine gets among the search results a large number of commercial websites that are relevant to his query, but not always relevant to the user's needs. Most strongly it complicates the search for scientific and technical information. The algorithms of the search engines are constantly being improved to deal with the artificial promotion of web resources, but despite this searching for relevant information becomes more and more difficult task.

Purpose: Construction of the model for sifting irrelevant commercial information from the results of web search on the basis of research for identifying the characteristic features of commercial sites. Constructed model is a part of technology for improving efficiency of the relevant information Internet search.

Problem statement: One of possible options for improving the efficiency of relevant information search is the division of all the volume of Internet information by some pre-defined characteristics to scientific-technical and commercial, the selection of characteristic features for commercial information and to build a model for its effective sifting. So, it was decided to develop an information technology for improving the efficiency of scientific and technical information search by sifting irrelevant to the user's needs commercial information based on pre-defined characteristic features.

The proposed technology consists of the following steps:

1. Search the Internet for web resources relevant to the search query.
2. Sifting of sites containing irrelevant to user expectations commercial information.
3. Ranking of search results obtained from the second stage using the classifier trained on a predefined selection.

Main results: During the analysis of commercial sites content there were identified a number of characteristic features, allowing uniquely identify them. Based on the identified features it was built a new model of automatic information classification to relevant and commercial.

A new model of automatic information classification to the relevant and commercial by the set of characteristic features was developed. This technology allows to increase the percentage of relevant information in search results to 83-92%.

Conclusion: The described model can be applied not only to improve the efficiency of web search by sifting irrelevant information, but also as a filter that blocks the search for goods and services to prevent improper use of working time. This feature will be especially useful at the offices of commercial structures, where there are separate administrators to deal with misuse of the Internet.

Keywords: information search, scientific and technical information, search engine, improving the search relevance.

Получено 11.02.2013