УДК 004.738.52
З-87

ИССЛЕДОВАНИЕ ТЕХНОЛОГИИ ПОВЫШЕНИЯ ЭФФЕКТИВНОСТИ ПОИСКА РЕЛЕВАНТНОЙ ИНФОРМАЦИИ В ИНТЕРНЕТЕ

В. В. Зосимов, асп.[1];
А. С. Булгакова, канд. техн. наук[2];
В. А. Поздеев, д-р физ.-мат. наук, проф.[3]

[1]*Отдел распределенных интеллектуальных систем МНУЦИТиС НАН и МОНМС Украины, г. Киев*

[2]*Отдел индуктивного моделирования МНУЦИТиС НАН и МОНМС Украины, г. Киев*

[3]*Николаевский национальный университет имени В.А. Сухомлинского, г. Николаев*

Аннотация. Описан процесс исследования технологии повышения эффективности поиска релевантной информации в Интернете. В ходе исследования было проведено сравнение полноты и точности поиска по результатам работы поисковой системы google.com.ua и предложенной технологии. В результате были получены данные, иллюстрирующие преимущества применения разработанной технологии перед современными поисковыми системами.

Ключевые слова: поиск информации, научно-техническая информация, поисковая система, повышение эффективности поиска.

Анотація. Описано процес дослідження технології підвищення ефективності пошуку релевантної інформації в Інтернеті. В ході дослідження було проведено порівняння повноти і точності пошуку за результатами роботи пошукової системи google.com.ua та запропонованої технології. В результаті були отримані дані, що ілюструють переваги застосування розробленої технології перед сучасними пошуковими системами.

Ключові слова: пошук інформації, науково-технічна інформація, пошукова система, підвищення ефективності пошуку.

Abstract. This article describes the process of researching technology of improving the efficiency of the relevant information search via the Internet. Comparison of search completeness and accuracy according to the results of google.com.ua search system and in accordance with proposed technology was carried out. As a result data, that shows advantages of the developed technology in comparison with moderns search systems, was obtained.

Keywords: information search, scientific-technical information, search system, improving of search efficiency.

ПОСТАНОВКА ПРОБЛЕМЫ

Современные тенденции развития Интернета привели к тому, что он все больше становится похож на огромную рекламную площадку, что значительно затрудняет поиск информации в сети. Алгоритмы работы поисковых систем развиваются и непрерывно совершенствуются вместе с Интернетом, но, несмотря на это, поиск релевантной

информации становится все более сложной задачей [1, 2]. Необходимость эффективного использования динамично растущего и изменяющегося объема информации обуславливает актуальность и значимость исследований в области информационного поиска.

В этой области отдельно выделяется задача поиска научно-технической информации, так как с развитием всемирной паутины и ежедневным добавлением в сеть Интернет тысяч веб-ресурсов с коммерческой информацией, информацию научно-технического характера все труднее найти.

Далее в работе 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. Примерами коммерческой информации являются:

предложения скачать выложенную на сайте информацию в обмен на просмотр рекламы или отправку SMS;

информация о предлагаемых товарах или услугах, условиях работы компании, проводимых акциях и т. д.;

объявления о вакансиях или покупке/продаже/обмене.

Поиск научно-технической информации — это целенаправленный поиск Интернет-документов, относящихся с той или иной степенью релевантности к определенной научно-технической теме, запрашиваемой пользователем. Одним из возможных вариантов осуществления поиска является разделение по некоторым признакам всего объема информации в Интернете на научно-техническую и коммерческую. То есть решение задачи отсеивания нерелевантной информации в Web, а также задачи классификации — уточнения (распределения) результатов поиска, полученных после работы этапа отсеивания с помощью классификатора, обученного на заранее заданной выборке.

АНАЛИЗ ПУБЛИКАЦИЙ И ПОСЛЕДНИХ ИССЛЕДОВАНИЙ

На сегодняшний день существует несколько методов, которые так или иначе оптимизируют информационный поиск. Весомый вклад в теорию и практику информационного поиска внесли М. Губин, И. Кураленок, А. Максаков, В. Рувинская, К. Манукян, A. Barfourosh, S. Chakrabarti, C. Manning, S. Meyer и другие украинские и зарубежные ученые. Однако все предложенные методы решали задачу информационного поиска как ряд отдельных задач, не связанных друг с другом, или при решении одних задач ухудшали показатели других.

Так, рядом ученых предлагались методы периодического тематического поиска, классификации результатов поиска по ключевому слову, сочетающие в себе методы поиска по ключевым словам и методы тематической фильтрации, основанные на машинном обучении. Однако данные методы имеют ряд недостатков, а именно:

разделение поиска на множество категорий (тем) приводит к постоянному обучению на этапе классификации, что ведет к очень низкой производительности. То есть пользователь по заданному запросу сможет получить ответ не через секунды или минуты, что еще приемлемо, а через часы работы системы (А. Максаков, S. Meyer)

подбор «правильных» ключевых запросов не всегда может удовлетворить потребности пользователя, что приводит к низкому качеству поисковой выдачи. Кроме того, задача подбора ключевых запросов с 2010 года успешно решается поисковыми системами (А. Максаков).

ЦЕЛЬЮ РАБОТЫ является повышение эффективности и качества

поиска релевантной научно-технической информации в веб с применением индуктивных алгоритмов классификации.

ИЗЛОЖЕНИЕ ОСНОВНОГО МАТЕРИАЛА

В ходе анализа работы поисковых систем стало известно, что их алгоритмы при поиске научно-технической информации обладают низкими показателями точности [3].

Для повышения релевантности поиска научно-технической информации необходимо отфильтровать всю информацию, заведомо нерелевантную потребностям пользователя. Работа предложенной технологии состоит из трех этапов: сбор данных, отсеивание нерелевантной информации, ранжирование результатов с учетом их релевантности введенному запросу.

В ходе исследований были выявлены четыре категории сайтов, которые генерируют основную часть поискового спама, а также искусственно «раскручиваются»: интернет-магазины; сайты фирм, предлагающих услуги либо товары (не магазины); сайты, предоставляющие скачивание информации за просмотр рекламы; доски объявлений.

В последнюю категорию входят как специализированные доски объявлений, предоставляющие возможность размещать предложения по определенной тематике, например автомобили, вакансии или недвижимость, так и универсальные доски объявлений, разбитые на множество категорий и предоставляющие возможность размещать предложения в любую из категорий.

Для каждой из перечисленных выше категорий сайтов в ходе анализа их содержимого были выделены характерные только ей признаки. Они позволяют однозначно идентифицировать принадлежность сайта к той или иной категории.

Предложенная в работе технология повышения релевантности поиска научно-технической информации в Интернете состоит из трех этапов [3].

1. Поиск в Интернете веб-ресурсов, релевантных введенному поисковому запросу. Данный этап может быть реализован тремя различными способами:

написание собственных программ для поиска информации. Такой способ обеспечит высокое быстродействие, но потребует огромного количества ресурсов как материальных, так и временных;

использование данных из поисковой выдачи одной или нескольких поисковых систем, например Яндекс, Rambler и т. д. Такое решение даст намного меньшую скорость работы за счет того, что при каждом введенном пользователем запросе будет происходить анализ поисковой выдачи, а затем каждого отдельного сайта для получения данных, необходимых для последующего ранжирования результатов. Здесь низкое быстродействие компенсируется низкими затратами;

получение информации напрямую из базы данных поисковой системы. Сегодня такую возможность предоставляет только поисковая система Google. Данный способ обеспечивает показатели скорости и полноты поиска не ниже, чем при написании собственных программ, а также требует меньше материальных и временных затрат, чем при использовании данных из поисковой выдачи. Как видно, третий вариант реализации этапа поиска сочетает в себе все плюсы двух предыдущих, при этом исключая их минусы. Еще одним аргументом в пользу выбора поисковой системы Google является то, что она наиболее популярна среди украинских пользователей и имеет высокие показатели полноты поиска информации.

Полученный из базы данных Google список сайтов сохраняется

и передается на обработку фильтру, который на основе описанных выше признаков отделяет коммерческую информацию. Параметры, полученные вместе со списком сайтов из базы данных Google, не используются на этапе фильтрации и сохраняются до момента начала ранжирования оставшихся после фильтрации сайтов.

2. Отсеивание сайтов, содержащих нерелевантную ожиданиям пользователей коммерческую информацию. Отсеивание коммерческих сайтов 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 один из списков, то он без дальнейшего анализа относится к указанной в списке категории.

Во время проведения экспериментов по выявлению характерных признаков сайтов, порождающих поисковый спам, анализировался ряд элементов сайтов. Эти же элементы анализируются при проведении фильтрации полученного на этапе сбора данных списка сайтов.

Список анализируемых элементов сайтов:

мета-теги, пути к Java-скриптам и картинкам дизайна;

заголовки;

наличие корзины покупателя;

приветственный текст на Главной странице;

элементы навигации.

Анализ этих элементов происходит в следующем порядке:

1) проверяется доменное имя на наличие в списках с коммерческой и научно-технической информацией;

2) мета-теги анализируются на наличие в них названий распространенных CMS, предназначенных для разработки Интернет-магазинов;

3) проверяются пути к Java-скриптам и картинкам из дизайна на соответствие с CMS Интернет-магазинов;

4) проверяется наличие корзины покупателя;

5) проверяются заголовок, мета-описание, ключевые слова;

6) проверяются элементы навигации;

7) проверяются заглавные (приветственные) тексты с главной страницы.

Представленные выше элементы сайтов 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.

3. Уточнение (распределение) результатов поиска, полученных после работы второго этапа с помощью классификатора, обученного на заранее заданной выборке. То есть 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, най-

денному при помощи ОИА МГУА для всей категории «научно-техническая информация». Подобное разделение дает системе независимость алгоритма от процесса обучения по каждому запросу в отдельности, что значительно увеличивает скорость системы и возможность ее работы в онлайн-режиме. Однако стоит учесть ситуацию, в которой пользователь самостоятельно сможет формировать поисковую выдачу для собственных категорий.

В работе не используются данные ранжирования Google, так как его алгоритмы настроены на ранжирование сайтов с учетом наличия среди них большого количества поискового спама. В алгоритме Google большой вес имеют внешние показатели сайта (количество внешних ссылок, возраст домена), так как их, по мнению Google, труднее подделать, чем внешние. И соответственно меньше веса полагается внутренним параметрам (заголовки, количество ключевых слов на странице). Алгоритм Google неплохо ранжирует сайты с учетом поискового спама, но на этапе фильтрации поисковый спам был отсеян, поэтому формула ранжирования должна быть скорректирована с учетом этого. В предложенной в работе технологии используется своя формула для ранжирования, в которой веса для параметров определяются автоматически при помощи ОИА МГУА на основе обучающей выборки [4].

В ходе исследования был проведен ряд экспериментов по определению эффективности работы предложенной технологии. Приведены результаты трех наиболее значимых из них. Для каждого эксперимента был выбран один поисковый запрос. Выбранный запрос сначала обрабатывался поисковой системой Google, а затем отдельно обрабатывался по представленной в работе технологии. Для упрощения описания эксперимента анализу подлежали только первые сто сайтов из поисковой выдачи Google. Далее сравнивались показатели точности поиска поисковой системы Google и поиска по представленной в работе технологии.

Результаты проведения экспериментов представлены в табл. 1.

Таблица 1. Результаты экспериментов по исследованию эффективности предложенной технологии

Поисковый запрос	Показатель точности поиска, %	
	Google.com.ua	Предложенная технология
Защита информации	34	89
Программирование 1С	37	92
Устройство двигателя bobcat	54	83

Из табл. 1 видно, что применение предложенной технологии позволяет значительно повысить показатель точности поиска научно-технической информации. Важно отметить, что кроме повышения показателей точности поиска предложенная технология дает показатели полноты поиска не меньше, чем у современных поисковых систем. Об этом свидетельствует тот факт, что ни один сайт, содержащий научно-техническую информацию, не был ошибочно отсеян на втором этапе предложенной технологии.

ВЫВОДЫ

В рамках предложенной технологии разработан метод фильтрации коммерческой информации (поискового спама). Данный метод позволяет, как было выявлено в ходе экспериментов, повысить точность поиска научно-технической

информации до 83…91 %. Также в рамках технологии экспериментально на основе анализов экспертов получена новая формула для ранжирования результатов поиска.

Предложенная технология не просто повышает точность поиска научно-технической информации, но и может выступать в роли фильтра, который отсеивает все коммерческие (рекламные) сайты. Благодаря этому ее можно применять в высших учебных заведениях, НИИ, коммерческих организациях, словом, везде, где идет процесс обучения и ведутся научные исследования. Эти процессы в настоящее время не обходятся без поиска научно-техничской информации в сети Интернет, а наличие огромного количества поискового спама существенно его замедляет, а порой даже отвлекает и побуждает задуматься над покупкой рекламируемого товара или услуги.

Предложенная технология также выступает и в роли фильтра, не позволяющего искать товары и услуги, то есть предотвращает нецелевое использования рабочего Интернета в личных целях. Особенно полезно это свойство технологии будет в коммерческих структурах (в офисах), где для борьбы с нецелевым использованием Интернета существуют отдельные администраторы.

СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

[1] **Байков, В. Д.** Интернет. Поиск информации. Продвижение сайтов [Текст] / В. Д. Байков. — С.Пб. : БХВ-Петербург, 2000. — 288 с.

[2] **Колисниченко, Д. Н.** Поисковые системы и продвижение сайтов в Интернете [Текст] / Д. Н. Колисниченко. — М. : Диалектика, 2007. — 272 с.

[3] **Зосимов, В. В.** Проектирование программного модуля для исследования работы алгоритмов поисковых систем [Текст] / В. В. Зосимов, А. С. Булгакова // Материалы междунар. науч. конф. «Интеллектуальные системы принятия решений и проблем компьютерного интеллекта» ISDMCI 2011. — С. 39–42.

[4] **Stepashko, V.** Modified multilayered GMDH algorithm with combinatorial optimization of partial descriptions complexity [Text] / V. Stepashko, O. Bulgakova, V. Zosimov // Proceedings of the International Workshop on Inductive Modelling IWIM-2010, Ukraine. — Yevpatoria, 2010. — P. 24–30.