ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ В БИБЛИОТЕКАХ

УДК 025.4.036

Ю. В. Смирнов
ГПНТБ России

Поисковые системы интернета и методы повышения качества обработки запросов при поиске научной информации в сети

Проанализированы современные поисковые системы интернета как глобальные, так и локальные, отмечены их достоинства и недостатки. Подробно рассмотрен популярный метод организации и систематизации информации интернет-сайтов – тегирование, перечислены его слабые и сильные стороны. Приведены примеры использования тегирования. Сделан вывод: каждая поисковая система уникальна и обладает своими алгоритмами поиска, совместное использование которых представляется оптимальным для пользователя и даёт ему возможность самостоятельно выбрать то, что подходит для решения конкретной задачи.

Ключевые слова: поисковые системы интернета, поиск, тегирование, каталогизация интернет-сайтов, Web of Science.

UDC 025.4.036

Yury Smirnov
Russian National Public Library for Science and Technology, Moscow, Russia

Internet search engines and ways to improve quality of inquiries fulfillment when searching for scientific information on the Internet

Existing Internet search engines are analyzed. Tagging, with its advantages and drawbacks, is examined as a popular method of Internet information organization and classification. The author concludes that every search engine is unique for its search algorithm, and combined use of many is seen and the most efficient for users.

Keywords: Internet search engines, tagging, www-site tagging, Web of Science.

There are more and more information produced (total for 2011 is 1,8 zettabyte). To perform searches in this ocean there are invented: global systems like Google, Bing, Yandex, etc., and local machines. All these systems have certain advantages: 1) simplicity and ease of use that enables an inexperienced user to proceed immediately to search; 2) sorting or ranking of search results from most relevant to least; 3) displaying the page title and a small excerptions. These systems have common disadvantages: 1) commercialization, the main purpose of their existence is to make money; 2) the vulnerability with respect to the fake keywords; 3) sorted by relevance only, not taking into account the date of creation of the page; 4) the number of relevant links sometimes exceeds several millions; 5) the absence of a clarification instruments; 6) lack of relevant links, which is also sometimes happens. One of the best query refinement systems has Web of Science. In the sidebar on the left there are all the available types of search refinements. The main type search engines are verbal, i.e. based on natural language. This leads to the fact that the relevant information is published in one language when searching for key words in another language does not get in the search results. This is one of the main disadvantages of verbal search. The libraries came to the conclusion that the use of coded information retrieval languages (IRL) is useful for the search of documents in different languages. One of the advantages of IRL is independence of the language of the document. Many Internet companies are developing their own system of organization and systematization of information. Some search engines offer the use the catalog, e.g. Yandex, which has annotated links to sites collected manually by editors of Yandex. One of these methods is the "tagging" – a class of verbal language. To select a tag from the message text one should use a special format, which is a combination of "#" sign (grid or octothorpe) followed by the word or phrase written without spaces (for example: #novyygod2016, #somuchfun). The active use of this method you can find at the Yarra Plenty Regional Library.

С каждым годом количество произведённой человечеством информации увеличивается. На 2011 г. объём данных составлял более 1,8 зеттабайт (1,8 трлн Гб) [**1**]. Как считают в международной исследовательской и консалтинговой компании *International Data Corporation* (IDC) [**2**], количество данных будет удваиваться каждые два года до 2020 г., а количество полезной информации будет составлять только 35% от общего объёма данных [Там же]. Однако и это очень много.

Для облегчения поиска и ориентирования в таком объёме информации создаются различные поисковые средства. В распоряжении пользователей интернета достаточно много поисковых систем, которые по охвату индексируемых сайтов можно разделить на две группы:

глобальные, осуществляющие поиск по всем сайтам сети (например *Google*, *Bing*, *Yandex* и т.д.);

локальные, встроенные в один или несколько родственных сайтов, которые ведут поиск только по ним.

Стоит отметить, что почти все глобальные поисковые системы могут использоваться и в качестве локальных, однако относить их к этой группе неправомерно, поскольку поиск по отдельному сайту для них является уточнением запроса.

Все эти системы обладают определёнными достоинствами, в числе которых простота и удобство использования, что позволяет неподготовленному пользователю сразу приступить к поиску информации; ранжирование или сортировка результатов поиска от наиболее релевантных к менее релевантным; отображение заголовка страницы и небольшого экстракта (обычно 2–3 строки) рядом со ссылкой на сайт, что позволяет составить первое впечатление о релевантности сайта или выданного результата.

Вместе с тем все эти системы обладают общими недостатками:

коммерциализованность: большинство этих систем коммерческие, основная их цель – приносить прибыль, поэтому они часто и не всегда к месту размещают рекламу, а также «продвигают» сайт, т.е. искусственно повышают его релевантность;

уязвимость: поскольку механизмы индексации поисковых систем автоматические, это позволяет создателям страниц вводить для повышения релевантности ключевые слова, которые не имеют отношения к содержанию страницы, но при этом видны только поисковым системам (т.е. при загрузке страницы они не отображаются);

сортировка только по релевантности: не учитывается дата создания страницы, поэтому очень часто на первых страницах результатов поиска идут ссылки на релевантные, но устаревшие материалы;

избыток релевантных ссылок, число которых иногда доходит до нескольких миллионов;

отсутствие уточнения запроса по интересующим областям;

иногда отсутствуют релевантные ссылки.

Каждая поисковая система старается улучшить результаты поиска и избавиться от перечисленных выше недостатков или хотя бы минимизировать их. Одни системы пытаются совершенствовать алгоритмы поиска, другие – предлагают пользователю уточнить поисковый запрос.

Многие поисковые системы реализовали функцию «подсказок», которая при наборе текста в поисковом поле выдаёт небольшой список наиболее часто встречающихся запросов. Большинство глобальных поисковиков

предлагают уточнить запрос по типу информации, например: *Yandex* – выбрать из небольшого списка (Поиск, Картинки, Видео, Карты, Маркет, Новости, Музыка, Диск, Перевод, Почта, Словари, Всё), что именно ищет пользователь.

Некоторые поисковики обеспечили пользователям возможность задать временные рамки запроса. Например, *Google* предлагает либо выбрать из списка период создания страниц, либо задать собственный временной интервал. Также некоторые поисковые системы для уточнения поиска предлагают воспользоваться специальными операторами и пунктуацией.

В качестве примера приведём некоторые операторы и знаки пунктуации, предлагаемые *Google* [**3**]:

«*» (звёздочка) служит для замены любого слова в запросе;

«-» (дефис) – для исключения слова из запроса;

«”текст”» (текст в кавычках) – для поиска полной фразы, заключённой в кавычки;

«OR» (оператор «ИЛИ») – для поиска одного из слов, разделённых этим оператором, и т.д.

Система уточнения запросов, несомненно, полезна, однако поиск научных статей в глобальных поисковиках по-прежнему затруднён, поскольку производится по всей сети. Для исключения сайтов, не содержащих научной информации, компания *Google* предлагает воспользоваться поисковой системой – *Академией Google* (*Google Scholar*) [**4**], которая ведёт поиск научных публикаций по статьям как со свободным, так и с ограниченным или платным доступом. Результаты поиска представляют собой ссылки либо на полный текст статьи, либо на страницу с кратким описанием.

Эта поисковая система имеет также небольшую систему уточнения запросов: уточнение времени публикации; выбор сортировки результатов поиска (по релевантности или по дате); возможность включить в результаты поиска либо исключить из них патенты, показывать либо скрывать цитаты.

Однако *Академия Google* обладает такими серьёзными недостатками, как недостаточность данных об охвате базы данных; неизвестная частота обновления; отсутствие опубликованного списка научных журналов, представленных в БД.

Таких недостатков нет у коммерческих поисковиков (например, *Web of Science* [**5**]), являющихся локальными.

Одна из лучших систем уточнения запросов создана для поисковой системы сайта *Web of Science* [Там же], представляющего собой реферативную БД публикаций в научных журналах (компания *Thomson Reuters*). В боковой

панели слева расположены все доступные типы уточнения поиска, например: базы данных; направления исследования; авторы; годы публикаций; языки; страны/территории и т.д. В каждом из этих типов есть небольшой список наиболее часто встречающихся вариантов во всех документах основного запроса. Например, на запрос «*folksonomy*» и для типа уточнения «Годы публикаций» предлагается следующий список: 2009, 2011, 2010, 2013, 2008.

Это означает, что наибольшее количество публикаций по основному запросу «*folksonomy*» пришлось на 2009 г.

Система также предлагает воспользоваться операторами поиска (например: «AND» для поиска записей, содержащих все условия) и символами усечения (например: «?» (знак вопроса) для замены одного символа).

При всех достоинствах эта поисковая система обладает одним существенным недостатком, особенно для русскоязычных пользователей. Несмотря на то, что сайт русифицирован, запрос в основной БД – *Web of Science Core Collection* – вводится только латинскими символами, а значит возникают сложности с транслитерацией. Зачастую автор транслитерирует свою фамилию и имя по-разному в разных публикациях, поэтому сотрудники компании *Thomson Reuters* предлагают пользоваться символами усечения, однако это не уменьшает количество результатов, а наоборот увеличивает.

Например, результаты поиска при транслитерации, согласно ГОСТу 7.0.34–2014 «Правила упрощённой транслитерации русского письма латинским алфавитом» [**6**], фамилия автора «Полежаев» в базе данных *Web of Science Core Collection* выглядят следующим образом:

«Polezhaev» – 287 публикаций;

«Polejaev» – 5 публикаций;

«Poleshaev» – 5 публикаций;

При использовании символа усечения («Pole*aev») система выдает 598 публикаций, включая авторов не только с этой фамилией, но и с фамилией «Полетаев».

Необходимо обратить внимание и на то, что основной вид поиска всех упомянутых нами поисковых систем – вербальный, т.е. базируется на естественном языке. Поэтому релевантная информация, опубликованная на одном языке, при поиске по ключевым словам на другом языке не попадает в результаты поиска. Это один из главных недостатков вербального поиска.

Компания *Thomson Reuters* попыталась обойти это ограничение и приняла решение вести основную БД *Web of Science Core Collection* на английском языке. Несомненно, английский является языком международного общения, однако не все люди хорошо владеют им, поэтому предпочитают ис-

кать информацию на родном языке.

Конечно, можно встроить в поисковую систему автоматический перевод запроса на разные языки, но это не так просто, поскольку естественные языки обладают рядом особенностей, которые затрудняют поиск и увеличивают объём нерелевантных документов. Профессор Н. И. Гендина в своей книге «Лингвистические средства библиотечно-информационных технологий» [**7**. С. 38] сформулировала эти особенности:

избыточность, т.е. наличие слов с небольшой смысловой нагрузкой (союзы, предлоги и т.д.), которыми можно пренебречь;

синонимичность, т.е. наличие синонимов, в том числе сокращений (например: «Соединённые Штаты Америки» и «США»);

многозначность, которая проявляется в омонимии и полисемии (например: «гусеница (личинка насекомого)» и «гусеница (танка)»).

С потребностью организации большого объёма информации, в том числе на различных языках, и улучшения последующего поиска библиотеки столкнулись ещё до появления интернета. За столетия развития библиотеки пришли к выводу: использование кодированных информационно-поисковых языков (ИПЯ) полезно не только для расстановки фонда, но и для поиска документов на разных языках. Одно из достоинств таких ИПЯ – независимость от языка составления документа. С течением времени они стали использоваться почти во всех библиотеках и также полезны при поиске по электронному каталогу.

Для подтверждения приведём один из тезисов учебника «Аналитико-синтетическая переработка информации»: «Отечественные библиотековеды считают, что именно систематический поиск естественен для читателей, так как вся система образования построена по систематическому принципу и обучение ведется не по “ключевым словам”, а по “дисциплинам”, отраслям знания, областям науки и практической деятельности» [**8**. С. 178].

Небольшая часть как глобальных, так и локальных поисковых систем интернета также предлагает воспользоваться каталогом или рубрикатором. Например, каталогом Яндекса [**9**], который представляет собой аннотированные ссылки на сайты, собранные вручную редакторами компании Яндекс.

Развитие таких каталогов в интернете не всегда оправдывает вложенные в их создание усилия, поскольку каталогизация сайтов – это исключительно интеллектуальный труд, требующий затрат на содержание штата работников. А так как объёмы информации – огромные, штат таких сотрудников должен быть очень большим. Поэтому поисковые системы интернета пока не готовы обратиться к опыту библиотек и использовать кодированный ИПЯ.

Однако многие интернет-компании разрабатывают собственные си-

стемы организации и систематизации информации, одна из них, уже завоевавшая популярность, – *тегирование*. «Тегирование – это подход, основанный на простоте действий при создании требуемого контента (теги – простые поисковые термины). Термин *фолксономия*, сейчас чаще говорят *таксономия*, был придуман для описания этого подхода по восходящему принципу для представления метаданных. Он в корне отличается от традиционного подхода, при котором квалифицированные каталогизаторы (индексаторы) присваивают соответствующие ключевые слова, взятые из общеизвестных списков нормализованной лексики» [**10**. С. 38, 39].

Такой подход характерен для локальных поисковых систем (социальные сети для публичного обмена сообщениями *Twitter*, ВКонтакте и т.д.), однако уже сейчас глобальный поисковик *Google* способен выдать список страниц с популярными тегами.

Для выделения тега из текста сообщения используется специальный формат, который представляет собой сочетание знака «#» (решётка, или октоторп) с последующим словом или фразой, написанной без пробелов (например: #новыйгод2016, #somuchfun).

Язык тегов относится к классу вербальных языков, для него характерны такие особенности естественных языков, как избыточность, синонимичность и многозначность. Также необходимо отметить, что теги создаются авторами публикаций произвольно, что избавляет интернет-компании от проблемы финансирования развития каталогов или рубрикаторов. Однако это достоинство может обернуться одним или несколькими недостатками для пользователя поисковой системы, которых лишены кодированные ИПЯ, например:

написание тегов: чтобы найти публикации, помеченные данным тегом, надо знать, как он точно пишется, а для этого необходимо ознакомиться со списком тегов, созданных на этом сайте (например: «#облако» или «#облачныевычисления»);

орфографические ошибки (в том числе намеренные) и опечатки, т.е. в запросе будут отсутствовать документы, помеченные ошибочно написанным тегом.

Несмотря на выявленные недостатки, тегирование становится всё популярнее как среди пользователей, так и среди разработчиков интернет-сайтов и поисковиков. В качестве примера активного использования этого метода в библиотеке можно привести Региональную библиотеку Ярра Пленти (*Yarra Plenty Regional Library*) [**11**] (Мельбурн, Австралия), поисковая система электронного каталога которой имеет не только систему уточнения запросов, но и возможность поиска по тегам, проставленным самими пользователями.

Каждая поисковая система, как глобальная, так и локальная, является уникальной и обладает своими алгоритмами поиска или их сочетанием. Подходы при формировании поискового запроса также отличаются, каждый из них имеет как достоинства, так и недостатки. Однако выявленные недостатки не должны быть причиной отказа от их дальнейшего использования.

Совместное использование разных подходов поиска представляется оптимальным для пользователя, который сам сделает выбор в зависимости от решаемых им задач. Например, электронный каталог ГПНТБ России, как было сказано выше, предлагает восемь видов поиска (стандартный, расширенный, профессиональный, по словарю, по УДК-навигатору и т.д.) [**12**], каждый из них – для решения определённого круга задач. Конечно, это усложняет поисковый интерфейс, однако позволяет пользователю самостоятельно решать, каким поисковым алгоритмом воспользоваться. Перефразируя всем известное выражение, можно сказать: «Определение вида поиска – это уже половина решения поискового запроса».

СПИСОК ИСТОЧНИКОВ

1. **BigData** шагает по планете / Виталий Постолатий. – Режим доступа: http://www.rg.ru/2013/05/14/infa-site.html (Дата обращения: 11.01.2016)
***BigData** shagaet po planete / Vitaliy Postolatiy*.

2. **IDC.** – Режим доступа: http://www.idc.com/home.jsp (Дата обращения: 11.01.2016).

3. **Операторы** в поисковых запросах. – Режим доступа: https://support.google.com /websearch/answer/2466433?hl=ru (Дата обращения: 11.01.2016).
***Operatory** v poiskovyh zaprosah.*

4. **Академия** Google. – Режим доступа: https://scholar.google.ru/ (Дата обращения: 11.01.2016).
***Akademiya** Google*.

5. **Web** of Science. – Режим доступа: http://apps.webofknowledge.com/ (Дата обращения: 11.01.2016).

6. **ГОСТ Р 7.0.34–2014**. Правила упрощённой транслитерации русского письма латинским алфавитом [Текст] / Федер. агентство по техн. регулированию и метрологии. – Москва : Стандартинформ, 2015. – 5 с.
***GOST R 7.0.34–2014.** Pravila uproshchennoy transliteratsii russkogo pisma latinskim alfavitom [Tekst] / Feder. agentstvo po tehn. regulirovaniyu i metrologii. – Moskva : Standartinform, 2015. – 5 s.*

7. **Гендина Н. И.** Лингвистические средства библиотечно-информационных технологий : учеб. / Н. И. Гендина. – Санкт-Петербург : Профессия, 2015. – 440 с.

Gendina N. I. *Leengvisticheskie sredstva bibliotechno-informatsionnyh tehnologiy : ucheb. / N. I. Gendina. – Sankt-Peterburg : Professiya, 2015. – 440 s.*

8. **Аналитико-**синтетическая переработка информации : учеб. / Н. И. Гендина и др. ; науч. ред. А. В. Соколов ; координатор проекта Л. В. Трапезникова. – Санкт-Петербург : Профессия, 2013. – 336 с.

Analitiko-*sinteticheskaya pererabotka informatsii : ucheb. / N. I. Gendina i dr. ; nauch. red. A. V. Sokolov ; koordinator proekta L. V. Trapeznikova. – Sankt-Peterburg : Professiya, 2013. – 336 s.*

9. **Яндекс** каталог. – Режим доступа: https://yaca.yandex.ru/ (Дата обращения: 18.01.2016).

Yandex katalog.

10. **Шрайберг Я. Л.** Электронная информация, библиотеки и общество: что нам ждать от нового десятилетия информационного века : Ежегод. докл. конф. «Крым», год 2011. – Судак / Я. Л. Шрайберг. – Москва : ГПНТБ России, 2011. – 80 с.

Shrayberg Ya. L. *Elektronnaya informatsiya, biblioteki i obshchestvo: chto nam zhdat ot novogo desyatiletiya informatsionnogo veka : Ezhegod. dokl. konf. «Crimea», god 2011. – Sudak / Ya. L. Shrayberg. – Moskva : GPNTB Rossii, 2011. – 80 s.*

11. **Yarra** Plenty Regional Library. – Режим доступа: http://www.yprl.vic.gov.au/ (Дата обращения: 18.01.2016).

12. **Электронный** каталог ГПНТБ России. – Режим доступа: http://library2.gpntb.ru/cgi/irbis64r_simplesite/cgiirbis_64.exe?C21COM=F&I21DBN=IBIS&P21DBN=IBIS&S21CNR=&Z21ID= (Дата обращения: 18.01.2016).

Elektronnyy *katalog GPNTB Rossii.*

Yury Smirnov, *researcher, Russian National Public Library for Science and Technology;*

yu.smirnoff@gmail.com

17, 3rd Khoroshevskaya st., 123436 Moscow, Russia