Метапоиск доступных научно-технических документов в Интернете

© Д.В. Ландэ[1], А.А. Снарский[2], В.В. Жигало[1]

[1] Информационный центр «ЭЛВИСТИ», г. Киев

[2] НТУУ «КПИ», г. Киев

dwl@visti.net, asnarskii@gmail.com, vladlen@visti.net

Аннотация

Представлены подходы к созданию средства мониторинга, адаптивного агрегирования и обобщения потоков информации из интернета. Предложена концепция адаптивного агрегирования информации, дано краткое описание экспериментальной системы PDF Science Search (PDFSS). Практическая значимость работы заключается в обосновании подходов и средств создания общедоступной информационно-аналитической среды для проведения научно-аналитических исследований.

Интенсивное развитие информационных сетевых технологий привело к резкому росту объемов документальной информации в сетевой среде. Несмотря на то, что большое число аналитических материалов публикуется на «закрытых» информационных ресурсах (тех, которые требуют оплаты, регистрационных данных, корпоративной принадлежности и т. п.), большая часть из них публикуется в веб-среде (на домашних страницах авторов, серверах пресс-релизов, торрентах, социальных сетях и т. п.). Рост объема и динамики информационной среды сопровождается многократным дублированием информации, слабой ее структуризацией, ростом уровня информационного шума [1, 2].

Своевременное получение многоаспектной и объективной документальной информации с помощью средств мониторинга компьютерных сетей, современных поисковых и метапоисковых систем для последующего ее использования в научных исследованиях может быть достигнуто лишь путем внедрения новых теоретических и технологических решений. Поэтому особо актуальным является разработка теоретических и технологических принципов построения адаптивных информационных хранилищ, автоматизированных систем обработки и обобщения информации из документальных хранилищ сверхбольшого объема, которые должны стать основой для создания интеллектуальной среды решения аналитических междисциплинарных проблем.

Задачи мониторинга информационных потоков большого объема в компьютерных сетях, их адаптивного агрегирования и обобщения осложняются отсутствием типовых методик и решений, неполнотой существующих технологических подходов. В настоящее время исследования по проблемам анализа информационных потоков большого объема в компьютерных сетях носят чаще всего узко специализированный характер. Вместе с тем, опыт создания и внедрения корпоративных информационных систем свидетельствует о необходимости создания и внедрения документальных информационных хранилищ для обеспечения научных исследований, получения разнообразных аналитических сведений, навигации в документальных информационных потоках больших объемов.

Представляется очень важным, чтобы агрегирование информации, формирование информационного хранилища было адаптивным, т. е. ориентированным на информационные потребности реальных пользователей. Если учитывать динамику и объемы доступной информации в интернете (на сегодняшний день доступно свыше триллиона веб-страниц), то становится очевидным, что обеспечение эффективного доступа в режиме поиска к информации в отрыве от информационных потребностей является практически неразрешимой задачей. Основная идея адаптивного агрегирования информации заключается в сборе и хранении в информационном хранилище только той информации, которая соответствует информационным потребностям пользователей (существующих или потенциальных). Для этого предполагается, что по мере развития системы в ее информационное хранилище будут попадать актуальные документы из интернета, соответствующие текущим запросам пользователей. Естественно, с ростом количества пользователей объемы информационного хранилища (репозитария) будут также расти, что в некоторый момент потребует пересмотра его содержания по некоторым критериям, например, по времени в соответствии с формулой Бартона – Кеблера [3], или по содержанию с использованием методов Text Mining.

Труды 12[й] Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2010, Казань, Россия, 2010

В настоящее время ни одна из традиционных поисковых систем на достаточном уровне не помогает при поиске актуальной документальной информации, которая находится в динамической части интернета. Решение этой задачи требует применения системы-посредника между пользователем и сетью. Подобный посредник должен выполнять работу по сбору, селекции информации и осуществлять предварительную обработку данных для создания документального информационного хранилища.

В настоящее время в интернет-пространстве содержится большое количество документальных ресурсов, представленных в формате PDF [4]. Популярность данного формата вызвана тем что он является компактным и удобным для хранения информации, представленной изначально в различных видах: простого текста, векторных и растровых изображений, страниц веб-сайтов, форм и мультимедийных файлов. Вместе с тем, при поиске необходимой документации в формате PDF с помощью традиционных сетевых информационно-поисковых систем пользователь постоянно сталкивается с проблемами, связанными с плохой доступностью целевой информации (условиями платного доступа, отсутствием необходимых файлов по указанным адресам или неверными гиперссылками). Хотя большинство поисковых систем, таких, как Google, Yandex, Rambler, Yahoo, выводит в список результатов информацию о найденных PDF-файлах, вместе с тем они часто дают ссылки на несуществующие PDF-файлы или ссылки на сайты, где PDF-файлы находятся в закрытом доступе. Например, указывая в строке адреса название PDF-файла (полученное с помощью Google Scholar) 36W622113036P357.pdf на сервере такого популярного издания, как Springer, пользователь получает не искомый документ, а его описание и регистрационную форму. Сказанное относится и к специализированным поисковым системам, ориентированным на поиск документов в формате PDF (например, OSUN – http://www.osun.org, PDFGod, http://www.pdfgod.com, http://pdf-search-engine.com/ и др.). В указанных поисковых систем нет возможности отсортировать или отфильтровать результаты поиска или просто поискать в базе данных с уже сохраненными PDF-документами. Все перечисленные системы поиска PDF-документов основаны на поиске информации в других поисковых системах. В основном они направлены на англоязычный сегмент пользователей и используют для получения информации в основном систему Google, что ограничивает выдаваемые результаты. Кроме того, лишь одна из специализированных поисковых систем может выдавать PDF-файлы в HTML-виде (это удобно для оперативного ознакомления с содержанием документов) – это pdf-search-engine.com.

С участием авторов была построена модель технологии агрегирования документальных информационных потоков, реализованная в виде метапоисковой системы PDF Science Search (PDFSS), доступной в настоящее время по адресу http://weblib.in.ua (рис. 1).

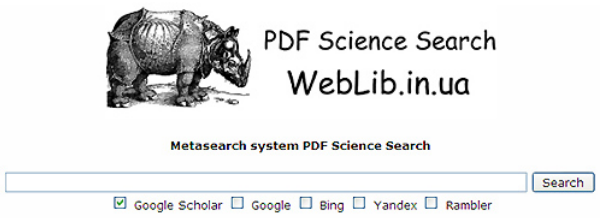

Рис. 1. Строка поиска на сайте WebLib.in.ua

Любая поисковая система в процессе работы просматривает определенный набор серверов и отбирает документы в соответствии с заданными критериями. Сегодня поиск с помощью разных систем по одним и тем же ключевым словам дает различные результаты. Это привело к идее создания так называемых метапоисковых (или мультипоисковых) систем [5], которые обращаются за помощью сразу к нескольким поисковым системам. Каждая из метапоисковых систем имеет свой язык запросов. Метапоисковая система переводит сформулированный на ее языке запрос на языки, используемые каждой машиной поиска. Далее, результаты поиска всеми системами объединяются и представляются в соответствующей форме. Естественно, поиск с помощью метапоисковых систем занимает больше времени по сравнению с обычными ИПС.

С помощью метапоисковой системы PDFSS можно искать PDF-файлы в таких поисковых системах, как Google, Bing, Yandex, Rambler, а также в ее собственной базе данных (кэше PDFSS). Поиск в кэше производится при любом запросе по умолчанию и выводится списком ниже результатов, полученных от других ИПС.

Особенностью PDFSS является то, что она полностью направлена на поиск доступных пользователю PDF-файлов, с возможностью фильтрации платных ресурсов, текстовых описаний, любой информации, кроме самих файлов, без сопровождающих их информационного шума или рекламы.

Общая схема работы местапоисковой системы PDFSS охватывает ряд этапов. После того, как пользователь задает запрос метапоисковой системе, с ее помощью создаются запросы для каждой поисковой системы, учитывающие уникальные возможности их синтаксиса. Затем модифицированные запросы пересылаются поисковым системам, которые возвращают результаты поиска. После этого метапоисковая система разбирает полученные результаты на отдельные документы и проверяет их доступность. Например, если в пути к документу присутствует доменное имя, присутствующее в стоп-списке, то документ отбрасывается и не используется в дальнейшей обработке. Это лишь один из критериев фильтрации. Те документы, которые прошли этап фильтрации, преобразуются для вывода результатов пользователю. Также производится поиск во внутренней базе данных файлов (в информационном кэше на прокси-сервере, содержащем найденные

ранее документы [6]). Если такие файлы были найдены, то вывод документа дополняется информацией о возможной доступности этого файла по обнаруженной ссылке. Если данный файл отсутствует по указанному адресу в интернете, то выводится сообщение, что данный файл может отсутствовать. Если же информация о данном файле присутствует в информационном кэше и он предположительно существует, то вывод дополняется информацией, такой, как размер файла, а также создается HTML-версия этого файла. После подсчета количества найденных документов подготовленные результаты выводятся пользователю через стандартный веб-интерфейс [7].

Таким образом, система PDFSS состоит из трех основных модулей (рис. 2):

- метапоисковая система;
- модуль кэширования информации (информационный прокси-сервер);
- внутренняя поисковая система, работающая как с информационным прокси-сервером, так и репозитарием.

Рис. 2. Модель системы адаптивного агрегирования информации

Основным критерием ранжирования информации в системе PDFSS является рейтинг поисковых систем. Так, например, у поисковой системы Google рейтинг выше, чем у системы Bing (в Google больший охват ресурсов, более релевантные результаты). В PDFSS происходит фильтрация неинформативных сайтов или сайтов с недоступными первоисточниками (так называемый «черный список»).

Если ссылка на один и тот же PDF-документ была получена из различных поисковых систем, то выбирается та из них, которая содержит более полное описание.

Результаты представляются пользователю в виде списков результатов различных поисковых систем, которые следуют друг за другом.

В системе PDFSS используется модуль кэширования, основная задача которого – сбор ссылок на PDF-документы, которые получены в процессе работы с пользователем метапоисковой системы, чтобы в дальнейшем сохранить в информационном хранилище (кэше PDFSS) файлы, а также сопутствующую им информацию, такую, как доступность файла по данной ссылке и размер файла.

Поиск в кэше PDFSS и ранжирование полученных результатов происходят по иному принципу. Так как в системе уже загружены тексты pdf-файлов, то строятся собственные таблицы релевантности с учетом частоты встречаемости ключевых слов, их позиции (если ключевое слово встречается в названии, то данный документ более релевантен, чем тот, в котором ключевое слово встречается в середине текста).

Система периодически обновляет информацию о тех файлах, которые были сохранены в базе данных PDFSS. Если файл не был ранее доступен, но доступен в тот момент, когда производится вторичное сканирование, информация в базе данных PDFSS обновляется; если же он становится недоступным, то в базу данных записывается информация о недоступности данного файла, чтобы в дальнейшем предложить пользователю получить этот файл из кэша. Далее PDF-файл кэшируется, конвертируется в текст, затем строится поисковый индекс этого файла.

Во внутреннем формате для каждого файла присутствует такая информация, как текстовый вариант PDF-файла, размер файла, ссылка, по которой был сохранен файл, ссылки на похожие файлы с других сайтов.

Внутренняя информационно-поисковая система позволяет пользователю искать в кэше системы PDFSS документы, которые динамически накапливаются. Каждый документ во внутренней поисковой системе ранжируется по релевантности. Критериями релевантности документа являются: количество вхождений ключевых слов (по которым пользователь ищет документ), размер документа, а также наличие подобных документов в базе данных метапоисковой системы. Результатом поиска информации в кэше PDFSS является аннотированный список найденных документов. Аннотации (сниппеты) документов – строки с первыми вхождениями ключевых слов введенных пользователем.

Метапоисковая система PDFSS изначально была создана как система метапоиска научно-технической документации и использовалась пользователями, которые искали именно такие документы. Соответственно в адаптивном кэше PDFSS присутствуют преимущественно научно-технические документы (их количество превышает 120 000) с более чем 40 тысяч источников. Лидируют среди источников сайт nbuv.gov.ua (Национальная библиотека Украины им. В.И. Вернадского – 5517 файлов), ioffe.ru (Физико-технический институт имени А.Ф. Иоффе – 1814 файлов), window.edu.ru (Единое окно, доступ к образовательным ресурсам – 1268 файлов) и др. Большинство из источников – это сайты университетов, институтов, а также научных журналов и электронных библиотек (таблица 1).

Таблица 1. Количество охваченных системой PDFSS PDF-файлов для различных источников

№ п.п.	Название источника	Количество файлов
1	nbuv.gov.ua	5517
2	www.ioffe.ru	1814
3	window.edu.ru	1268
4	ebiblioteka.lt	670
5	vestnik.udsu.ru	460
6	ecsocman.edu.ru	439
7	eprints.ksame.kharkov.ua	420
8	ict.edu.ru	393
9	rrc.dgu.ru	329
10	tstu.ru	314
11	unn.ru	297
12	rae.ru	297
13	library.iapm.edu.ua	290
14	library.tane.edu.ua	276
15	science.ncstu.ru	264
16	iai.dn.ua	264
17	sun.tsu.ru	262
18	lib.csu.ru	235
19	elar.usu.ru	232
20	isras.ru	229
21	eprints.zu.edu.ua	213
22	dtic.mil	211
23	vestnik.vsu.ru	210
24	lomonosov-msu.ru	208
25	jetpletters.ac.ru	200
26	vant.kipt.kharkov.ua	191
27	vak.ed.gov.ru	188
28	ogbus.ru	177
29	zhurnal.gpi.ru	174
30	ej.kubagro.ru	172
	…	
39	dialog-21.ru	138
	…	
49	arxiv.org	123
	…	
122	rcdl.ru	66

Именно благодаря эффекту адаптивности, наличия большого количества информации, уже загруженной с научных сайтов, журналов, серверов препринтов и т. д., можно констатировать, что сегодня система PDFSS лучше всего настроена на поиск научно-технической информации.

Сравнения результатов эксплуатации системы PDFSS (http://weblib.in.ua/) с другими подобными системами позволяет сделать заключение не только о том, что эта система лучше отфильтровывает недоступные пользователю документы, но и о ее лучшей ориентации на русский язык. Так, по запросам «персистентность фрактал» (запрос 1) и «persistence fractal» (запрос 2) различными документальными поисковыми системами было выдано соответственно документов:

http://www.pdf-search-engine.com/ – 10 и 100;
http://www.pdfgeni.com/ – 0 и 52;
http://pdfdatabase.com/ – 0 и 307;
http://ebookee.com/ – 0 и 86;
http://www.osun.org/ – 19 и 64;
http://weblib.in.ua/ – 131 и 184.

Данный оценочный пример свидетельствует о том, что при поиске по русскоязычному запросу система PDFSS является абсолютным лидером по полноте. Вместе с тем, она уступает по количеству выданных документов системе pdfdatabase.com. Однако, анализируя выдачу последней, можно сделать вывод, что pdfdatabase.com попросту не реализует операции конъюнкции и не всегда обеспечивает получение пользователями оригиналов документов (около 20% документов недоступны). Так, по слову persistence в этой системе находится всего 139 документов, а по слову fractal – 168.

Рассмотренная модель, реализованная в виде метапоисковой системы PDFSS, в настоящее время уже нашла своих пользователей и позволила сформулировать более сложные задачи, которые должны быть решены в рамках отдельной научно-исследовательской работы.

Предполагается, что результаты данной работы должны составить теоретическую базу для разработки автоматизированных систем мониторинга, адаптивного агрегирования и обобщения информационных потоков, построения и ведения информационных ресурсов сверхбольших объемов и разнообразной тематической направленности. Ожидаемые результаты позволят совместить в единой технологической цепочке мониторинг, информационный поиск, агрегирование информации с содержательным анализом данных, их обобщением, что повысит качество обработки сетевой информации, соответственно, эффективность информационно-аналитической поддержки научно-аналитической деятельности отечественных ученых и специалистов.

Литература

[1] Брайчевский С.М., Ландэ Д.В. Современные информационные потоки: актуальная проблема-

тика // Научно-техническая информация. Сер. 1. – 2005. – № 11. – С. 21-33.
[2] Lande D., Braichevski S., Busch D. Informationsfluesse im Internet // IWP – Information Wissenschaft & Praxis. – 2007. – V. 5, No 59. – P. 277-284.
[3] Bruton R., Kebler R. The half-life of some scientific and technical literature. // Am. Document. – 1960. – V. 11, No 1. – P. 18-22.
[4] Document management – Portable document format – Part 1: PDF 1.7 // Adobe Systems Inc. – 2008. – 756 p. – http://www.adobe.com/devnet/ acrobat/PDFs/PDF32000_2008.PDF.
[5] Meng W., Yu C, Liu K.L. Building efficient and effective metasearch engines // ACM Comput. Surv. – 2002. – V. 34, No 1. – P. 48-89.
[6] Додонов А.Г., Ландэ Д.В. Организация сети информационных прокси-серверов // Регистрация, хранение и обработка данных. – 2006. – Т. 8, № 3. – С. 24-31. – http://dwl.visti.net/art/infproxy/.
[7] Ландэ Д.В., Снарский А.А. Возможности системы мультипоиска доступных научно-технических документов в Интернет на примере тематики неразрушающего контроля и технической диагностики // Материалы 15-й межд. науч.-техн. конф. «Электромагнитные и акустические методы неразрушающего контроля материалов и изделий», 15 – 20 февраля 2010 г., Славское Львовской обл. – С. 105-107. – http://dwl.visti.net/art/slv/.

Metasearch of accessible scientific and technical documents in the Internet

D.V. Lande, A.A. Snarskii, V.V. Zhygalo

The article describes creation of means for monitoring, adaptive aggregation and generalization of streams of the information from the Internet. The concept of adaptive aggregation of the information is offered, the short description of experimental system PDF Science Search (PDFSS) is given.

The practical importance of work consists in a substantiation of approaches and means of creation of the popular information-analytical environment for carrying out of scientifically-analytical researches.