УДК 681.3

А. Г. Додонов[1], Д. В. Ландэ[1], В. В. Жигало[2]
[1]Институт проблем регистрации информации НАН Украины
ул. Н. Шпака, 2, 03113 Киев, Украина
[2]ИЦ «ЭЛВИСТИ»
ул. Максима Кривоноса, 2-А, 03037 Киев, Украина

Сетевые информационные потоки как содержательная составляющая информационно-аналитических систем

Представлены подходы к созданию средства мониторинга, адаптивного агрегирования и обобщения потоков информации из глобальных компьютерных сетей для обеспечения информационно-аналитической деятельности. 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.

Ключевые слова: *глобальная компьютерная сеть, информационный поток, информационные сетевые технологии, информационно-аналитическая среда, адаптивное агрегирование информации.*

Интенсивное развитие информационных сетевых технологий привело к резкому росту объемов документальной информации, публикуемой и сохраняемой в сетевой среде. Несмотря на то, что большое число аналитических материалов публикуется на «закрытых» информационных ресурсах (тех, которые требуют оплаты, регистрации, корпоративной принадлежности и т.п.), большая часть из них публикуется в веб-среде и в пиринговых сетях (на домашних страницах авторов, серверах пресс-релизов, торрентах, социальных сетях и др.). 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, в процессах принятия решений, управления, прогнозирования

динамики развития может быть достигнуто лишь путем внедрения новых теоретических и технологических решений. Сочетание средств мониторинга, формирования информационных хранилищ, документального информационного поиска с содержательным анализом данных в единой технологической цепочке позволит повысить качество обработки текущей информации и эффективность информационной поддержки научно-аналитической деятельности, процессов принятия решений [3, 4]. 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.

Для исследования и развития теоретических и технологических принципов, методов и программно-технических средств мониторинга, адаптивного агрегирования и обобщения информационных потоков необходимо:

— проведение анализа характеристик имеющихся сетевых информационно-поисковых систем с учетом количественных и качественных характеристик и динамики информации, которая охватывается ими;

— адаптация и модификация методов аналитической обработки и обобщения документальной информации, представленной в информационном хранилище сверхбольших объемов;

— разработка технологических принципов организации мониторинга, адаптивного агрегирования и обобщения информационных потоков в глобальных сетях;

— разработка технологических принципов организации сетевой метапоисковой системы в массивах документальной информации (статей, тезисов, диссертаций, научных отчетов и т.п.) в глобальных сетях;

— развитие теоретических принципов организации информационных ресурсов с целью формирования адаптивного документального хранилища для обеспечения научно-аналитической работы для широкого круга аналитиков.

Проблемы создания и развития методов и программно-технических средств мониторинга информационных потоков большого объема в компьютерных сетях, их адаптивного агрегирования и обобщения с последующим содержательным анализом данных сегодня решаются на основе разработки и внедрения специализированных программных приложений, предназначенных для применения в различных предметных областях, как крупными компаниями-производителями Hyperion Solutions Corporation, Oracle, Hewlett Packard, IBM, Microsoft, так и небольшими компаниями типа «third-party». Инструментальные средства разных производителей различаются по своим свойствам, функциональному составу, архитектуре базовых решений. Как правило, они имеют высокую себестоимость и для каждого отдельного случая внедрения нуждаются в существенной наладке и адаптации. Задачи мониторинга информационных потоков большого объема в компьютерных сетях, их адаптивного агрегирования и обобщения осложняются отсутствием типовых методик и решений, неполнотой существующих технологических подходов, невозможностью стандартизации процесса. В Украине, за редкими исключениями [1], исследования по проблемам анализа информационных

потоков большого объема в компьютерных сетях носят чаще всего узкоспециализированный характер [5]. Вместе с тем, опыт создания и внедрения корпоративных информационных систем свидетельствует о необходимости создания методологических принципов разработки и внедрения средств построения документальных информационных хранилищ для обеспечения проведения научных исследований, получения разнообразных аналитических сведений, навигации в документальных информационных потоках больших объемов.

Существенное повышение качества научно-аналитической деятельности отечественных научных работников, аналитиков, обеспечение высокого качества текущих и стратегических решений возможно лишь путем организации оперативного интеллектуального доступа пользователей к необходимой проблемно-ориентированной информации, которая, как показывает практика, в значительной мере представлена в глобальных компьютерных сетях, в частности, в Интернет. Для этого требуется создание теоретических и технологических принципов, развитие методов и программно-технических средств ориентации в сетевых информационных потоках сверхбольших объемов, методики обобщения информации, тематической фильтрации и редукции потоков по определенным критериям, средств содержательного анализа информации в рамках концепции глубинного анализа текстов (Text Mining).

На сегодня открытыми остаются еще множество вопросов, начиная от формальной постановки задачи, четкого определения понятий, таких как документальные информационные потоки в сетях, определения их основных характеристик, построения обобщенной схемы аналитической обработки и обобщения документальной информации, разработки методики тематической фильтрации и редукции информационных потоков на основе методов содержательного анализа информации и т.д.

Рассмотрим понятие информационного потока в контексте названной выше проблематики. Для однозначного понимания информационного потока, введем его формальное определение, которое корреспондируется с классическим из теории информации.

Рассмотрим интервал (a,τ) действительной оси (оси времени), где $\tau > a$. Допустим, что на этом отрезке времени в соответствии с некоторыми закономерностями в сети публикуется некоторое количество информационных документов — k. На оси времени их координаты обозначим как $\tau_1,\tau_2,...,\tau_k$ $(a \le \tau_1 \le \tau_2 \le ... \le \tau_k \le \tau)$. Информационным потоком будем называть процесс $N_\alpha(\tau)$, реализация которого характеризует количество точек (документов), которые появились на интервале (a,τ), как функцию правого конца интервала τ. В соответствии с этим определением реализация информационного потока является неубывающей ступенчатой целочисленной $N_\alpha(\tau)$.

Приведенное определение на локальных временных областях соответствует действительности, но не учитывает такой эффект, как старение информации, которое противоречит «накопительной» способности информационного потока $N_\alpha(\tau)$ на больших промежутках времени. Так определенный информационный поток учитывает лишь количество информационных сообщений, вне зависимости

от их содержания. Вообще, определение содержания, тематики отдельных документов является достаточно субъективным процессом. Для строгого моделирования тематических информационных потоков используют модели, которые различают документы по отдельным словам или словосочетаниям (обычно их называют термами, от англ. *Terms*). Обозначим множественное число документов как

$$D(\tau) = \{D_i, i = 1,..., N(\tau)\},$$

где D_i — документ с номером i; τ — время; $N(\tau)$ — количество документов в потоке в момент τ. $D_i = \{w_{ij}\}$, где w_{ij} — количество термов с номером j, которые входят в документ D_i.

Все Интернет-пространство можно условно разделить на две составляющие — стабильную и динамическую [2], которые имеют очень разные характеристики развития. В частности, процесс старения информации в известной модели Бартона-Кеблера описывается уравнением, которое состоит из двух компонент:

$$m(t) = 1 - ae^{-T} - b\,\mathrm{e}^{-2T},$$

где $m(t)$ — часть полезной информации в общем потоке через время T; первое вычитаемое соответствует стабильным ресурсам, а второе — динамическим. Как оказалось, это уравнение также в полной мере соответствует объемам информации, которые публикуются в Интернет по определенным тематикам. Стабильная составляющая Интернета содержит информацию «долгосрочного» плана, в то время как динамическая составляющая содержит ресурсы, которые постоянно обновляются. Некоторая часть последней составляющей впоследствии вливается в стабильную, однако большая часть «исчезает» из Интернет или попадает в сегмент так называемого «скрытого» веб, не доступного пользователям с помощью обычных информационно-поисковых систем (ИПС).

В традиционной сетевой информационно-поисковой системе информационное пространство, которое состоит из стабильной и динамической частей, и индексируется с помощью этой ИПС, изменяет свое наполнение в течение определенного количества дней: некоторые новые документы переходят в стабильную часть в виде архивов, а другие исчезают. В этом случае пользователь при обращении к традиционной ИПС находит релевантные запросу документы из стабильной части, ссылки из динамической части, которые устарели, и ничего не находит из обновленной динамической части.

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

Представляется очень важным, чтобы агрегирование информации, формиро-

вание информационного хранилища было адаптивным, т.е. ориентированным на информационные потребности пользователей. Если учитывать динамику и объемы доступной информации в Интернет (на сегодняшний день доступно свыше триллиона документов), 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 сохранении в информационном хранилище только той информации, которая соответствует информационным потребностям пользователей (существующих или потенциальных). Для этого предполагается, что по мере развития системы (и приобретения ей популярности) в ее информационное хранилище будут попадать актуальные документы из Интернет, соответствующие текущим запросам пользователей. Естественно, с ростом количества пользователей, объемы информационного хранилища (репозитария) будут также расти, что в конкретный момент потребует пересмотра его содержания по некоторым критериям, например, по времени в соответствии с формулой Бартона-Кеблера, или по содержанию, используя методы Text Mining.

Авторами была построена модель технологии агрегирования документальных информационных потоков, реализованная в виде метапоисковой системы PDF Science Search (PDFSS), доступной в настоящее время по адресу http://choos.in.ua.

В настоящее время в Интернет-пространстве содержится большое количество документальных ресурсов, представленных в формате PDF [7]. 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 и пр., выводят в список результатов информацию о найденных PDF-файлах [8], вместе с тем они часто дают ссылки на несуществующие PDF-файлы, или ссылки на сайты, где PDF-файлы находятся в закрытом доступе. В указанных поисковых системах нет возможности отсортировать или отфильтровать результаты поиска, или просто поискать в базе данных с уже сохраненными PDF-документами.

Основная идея предложенной метапоисковой системы состояла в том, чтобы находить в Сети PDF-файлы без сопровождающего их информационного шума или рекламы (до настоящего времени такой системы не существовало). 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.

Система PDFSS состоит из трех основных модулей (рис. 1):

— метапоисковой системы;

— модуля кэширования информации (информационного прокси-сервера [6]);

— внутренней поисковой системы, работающей как с информационным про-

кси-сервером, так и с репозитарием.

Рис. 1. Модель системы адаптивного агрегирования информации

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 [9], которые обращаются за помощью сразу к нескольким поисковым системам. Каждая из метапоисковых систем имеет свой язык запросов. Мультипоисковая 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.

На рис. 2. показана общая схема работы местапоисковой системы PDFSS. После того, как пользователь задает запрос метапоисковой системе, этот запрос обрабатывается, создаются запросы для каждой поисковой системы в ее специфическом формате. 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). 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, 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.

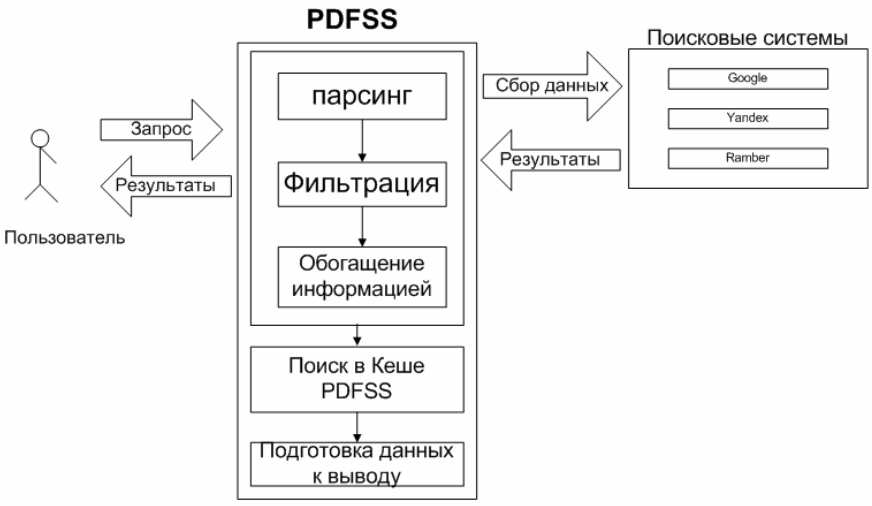

Рис. 2. Модель работы метапоисковой системы PDFSS

Главная задача модуля кэширования — 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, размер файла.

Система периодически обновляет информацию о тех файлах, которые 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.

Внутренняя поисковая система представляет собой поисковую систему, построенную на базе системы InfoStream [1]. 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.

Поисковая система позволяет пользователю искать в кэше системы PDFSS документы, которые динамически накапливаются. Каждый документ при поиске информации 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. В данном случае результатом поиска информации в системе является аннотированный список найденных в кэше PDFSS документов (рис. 3). Аннотации документов — строки с первыми вхождениями ключевых слов, введенных пользователем.

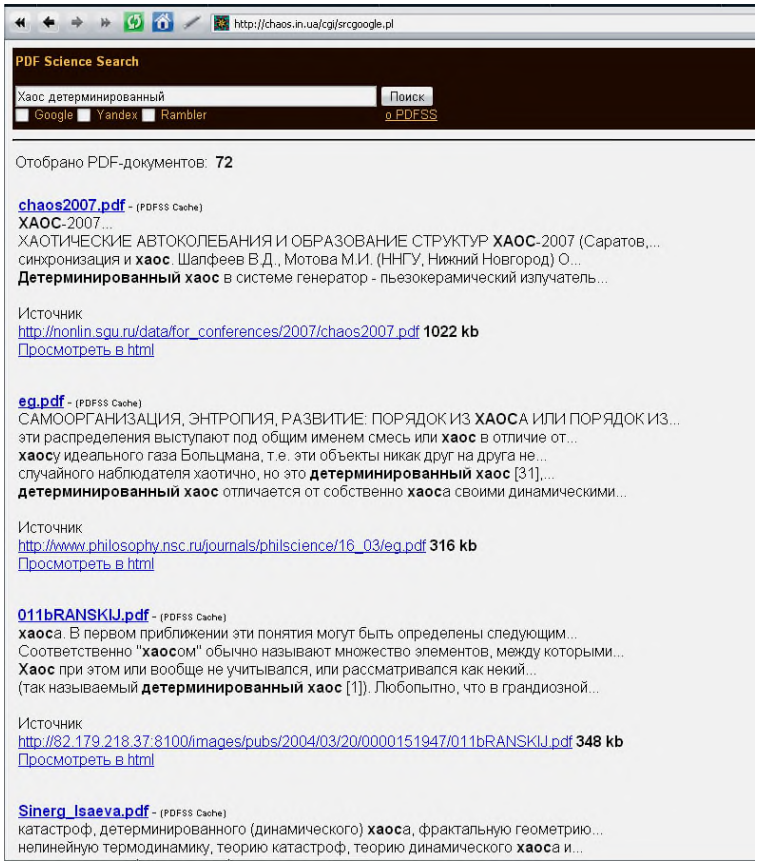

Рис. 3. Поиск в кэше PDFSS

На данный момент поисковая система PDFSS находится в бесплатном онлайн доступе на сайте «Хаос. Нелинейная динамика». С помощью поисковой системы PDFSS можно искать PDF-файлы в таких поисковых системах как Google, Yandex, Rambler, а также в ее собственной базе данных (кэше PDFSS) (рис. 4). Поиск в кэше производится при любом запросе по умолчанию и выводится списком ниже результатов, полученных от других ИПС.

Рис. 4. PDFSS на сайте chaos.in.ua

На рис. 5 показана страница с поисковыми результатами, полученными при поиске документов по запросу «Хаос&детерминированный».

Рис. 5. Поиск информации в таких ИПС как Google, Yandex, Rambler

Рассмотренная модель уже в настоящее время нашла своих пользователей и позволила сформулировать более сложные задачи, которые должны быть решены в рамках отдельной научно-исследовательской работы. Предполагается, что результаты данной работы (которая находится на начальной фазе) 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 по-

высит качество обработки информации из глобальных сетей, и, соответственно, эффективность информационно-аналитической поддержки научно-аналитической деятельности отечественных ученых и специалистов.

1. *Ландэ Д.В.* Основы интеграции информационных потоков / Д.В. Ландэ. — К.: Інжиніринг, 2006. — 240 с.

2. *Lande D.* Informationsfluesse im Internet / D. Lande, S. Braichevski, D. Busch // IWP — Information Wissenschaft & Praxis, 59(2007). — Vol. 5. — P. 277–284.

3. *Додонов О.Г.* Інформаційно-аналітична підтримка прийняття управлінських рішень / О.Г. Додонов, В.Г. Путятін, В.О. Валетчик // Реєстрація, зберігання і оброб. даних. — 2005. — Т. 7, № 2. — С. 77–93.

4. *Ланде Д.В.* Новітні підходи й технології інформаційно-аналітичної підтримки прийняття рішень // Національна безпека: український вимір: щокв. наук. зб. / Рада нац. безпеки і оборони України, Ін-т пробл. нац. безпеки. — К., 2008. — Вип. 1–2 (20–21). — С. 87–105.

5. *Додонов А.Г.* Самоподобие массивов сетевых публикаций по компьютерной вирусологии / А.Г. Додонов, Д.В. Ландэ // Реєстрація, зберігання і оброб. даних. — 2007. — Т. 9, № 2. — С. 53–60.

6. *Додонов А.Г.* Организация сети информационных прокси-серверов / А.Г. Додонов, Д.В. Ландэ // Реєстрація, зберігання і оброб. даних. — 2006. — Т. 8, № 3. — С. 24–31.

7. *Document* Management — Portable Document Format. — Part 1: PDF 1.7 [Елетронний ресурс] // Adobe Systems Inc. — 2008. — 756 p. — Режим доступу: http://www.adobe.com/devnet/acrobat/PDFs/PDF32000_2008.PDF

8. *Ландэ Д.В.* Дорожная карта сетевого поискового бизнеса / Д.В. Ландэ // Сети и бизнес. — 2009. — № 3. — С. 102–106.

9. *Meng W.* Building Efficient and Effective Metasearch Engines / W. Meng, C. Yu, K.L. Liu // ACM Comput. Surv. — 2002, Mar. — 34, 1 — P. 48–89.

Поступила в редакцию 06.11.2009