УДК 004.9

С.В. Кулешов, С.Н. Михайлов

ВАРИАНТ АРХИТЕКТУРЫ СУБПОИСКОВОЙ СИСТЕМЫ ДЛЯ РЕАЛИЗАЦИИ ФУНКЦИИ АНАЛИТИЧЕСКОГО МОНИТОРИНГА

Кулешов С.В., Михайлов С.Н. **Вариант архитектуры субпоисковой системы для реализации функции аналитического мониторинга.**

Аннотация. В статье предложен вариант архитектуры субпоисковой системы для реализации функции аналитического мониторинга, отличающийся формированием собственной базы данных и собственного поискового индекса. При этом для ускорения процесса сбора потенциально интересующих пользователя документов на начальной стадии в системе используются внешние Интернет-поисковые системы.

Ключевые слова: поисковая система, интернет, электронный документ, мониторинг, субпоиск.

Kuleshov S.V., Mikhailov S.N. **The variant of the subsearch engine architecture for analytical monitoring.**

Abstract. The paper proposes the variant of architecture of sub-search system for analytical monitoring implementation featuring its own database formation and own search index. For accelerating the process of obtaining the documents potentially relevant for user, the external internet search systems are involved on initial stage.

Keywords: web search system, internet, electronic document, monitoring, sub-search.

1. Введение. При проведении научных исследований одним из первых этапов является подбор литературы по теме проводимого исследования. Традиционно этот этап выполняется вручную при помощи поисковых систем Интернет, чтения найденных электронных ресурсов, сохранения их копий (или ссылок), корректировки поискового запроса в соответствии со списком ключевых слов и фраз и повторения всех действий до накопления требуемого объема текстового материала, необходимого для продолжения работы или принятия решения об отсутствии материала по выбранной тематике.

Естественно, не исключается и более традиционный способ, основанный на анализе бумажных вариантов печатных изданий в библиотеках.

В случае, когда исследование длится достаточно долго или тема является основной для исследователя, интересным и крайне полезным является наблюдение динамики изменений (мониторинг) в публикациях по интересующей тематике. Подобный мониторинг по набору заданных тем необходим и при организации учебного процесса [1].

Для решения перечисленных задач предлагается автоматизированная система мониторинга научных ресурсов, способная облегчить

работу исследователей, обеспечив выполнение рутинных операций и предоставив больше функциональных возможностей по сравнению с традиционным подходом.

2. Субпоисковая система. В отличие от подхода метапоиска, при котором разрабатываемая поисковая система не имеет собственной базы данных и поискового индекса, а формирует поисковую выдачу за счет агрегации и переранжирования результатов поиска других поисковых систем [2], предлагаемая система формирует собственную базу документов и собственный поисковый индекс, но для ускорения процесса сбора потенциально интересующих документов (уменьшение мощности множества I, рис. 1) использует внешние («большие») поисковые системы (Google, Яндекс, Bing).

На рис. 1 показано: *I* — множество документов, доступных в сети интернет, *W* — множество документов, отобранных Интернет поисковой системой, *S* — множество документов, отобранных субпоисковой системой.

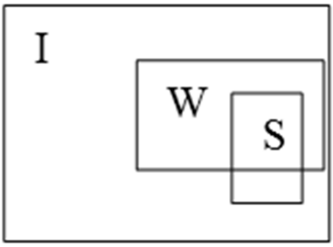

Рис. 1. Диаграмма соотношений множеств документов в субпоисковой системе.

В качестве научно-технического решения построения системы мониторинга предполагается разработка системы субпоиска, использующей внешние поисковые системы, и позволяющей производить накопление рабочего массива документов по заданной тематике с возможностью их динамического обновления. Кроме того, в рамках субпоисковой системы обеспечиваются сервисные функции, необходимые при выполнении информационно-аналитического обеспечения научной деятельности: поиск по сформированной базе, визуализация динамики обновлений, оценка близости текстов и т.п.

Необходимость использования внешних поисковых систем требует учитывать ограничения и особенности их использования.

Поисковые запросы передаются системам с использованием специализированных программных интерфейсов (API), позволяющих до-

полнительно указывать параметры поиска, настраивать критерии выдачи и получать выдачу результатов в виде XML-файлов. Использование API избавляет от необходимости производить разбор (парсинг) предназначенной для отображения пользователю через Web-интерфейс страницы поисковой выдачи (SERP), и, соответственно, не зависеть от меняющегося дизайна пользовательского интерфейса.

Среди наиболее общих ограничений «больших» поисковых систем можно отметить:

– ограничение по количеству запросов в сутки (конкретное количество разрешенных запросов зависит от конкретной системы, но в том или ином виде ограничение по бесплатному использованию API поисковых систем сторонними сервисами присутствует всегда),

– поиск ведется по неклассифицированным источникам (эклектической антологии),

– наличие механизмов поиска по предпочтениям, который будучи ориентирован на повышение эффективности показов рекламы, потенциально может искажать результаты поиска.

Общая структура взаимодействия в рамках системы мониторинга и место субпоисковой системы в ней в рамках предлагаемого подхода приведены на рис. 2.

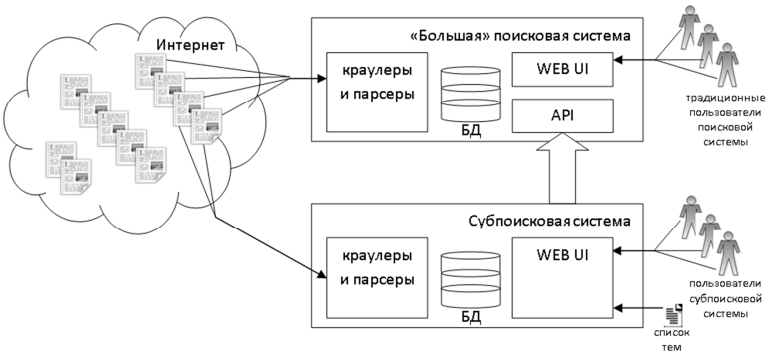

Рис. 2. Окружение субпоисковой системы и общая структура взаимодействия.

3. Принцип работы системы мониторинга. Работа системы начинается с задания списка тем для проведения мониторинга и автоматического (или полуавтоматического) формирования глоссария.

Субпоисковая система формирует набор запросов к поисковой системе, стараясь минимизировать их количество.

По выдаче результатов поисковой системы (SERP) формируется множество документов *W* (рис. 1), заносящееся в базу данных субпоисковой системы. Далее каждый из документов множества *W* загружается и анализируется уже собственными средствами субпоисковой системы и анализируется на предмет соответствия интересующей предметной области, заданной фрагментом семантического окружения [3].

В процессе анализа документов определяется дополнительный набор параметров для каждого документа (параметры качества и связности текста [4], степень принадлежности образца семантического окружения к семантическому окружению, построенному по документу).

Отобранные документы, прошедшие по заданным критериям качества и принадлежащие заданной тематической области, составляют множество *S* и сохраняются для возможности дальнейшей работы с ними в режиме «оффлайн».

В случае наличия внешних ссылок (например, в списках литературы), субпоисковая система производит загрузку и анализ документов, расположенных по этим ссылкам, расширяя тем самым множество *S*.

Возможна одновременная работа системы по нескольким интересующим темам. Ограничениями в этом случае являются производительность аппаратного обеспечения хостинга, на котором размещается система, и суточный лимит запросов к внешним поисковым системам.

С заданной периодичностью (несколько дней, недель) субпоисковая система повторяет операции для каждой заданной темы, добавляя новые документы и формируя отчеты по динамике появления и исчезновения интересующих публикаций.

Зависимость начальной выборки (множества *S*) от способа построения запросов, что потенциально может приводить к пропускам некоторых источников, можно назвать недостатком системы, который частично компенсируется расширением множества *W* собственными средствами субпоисковой системы.

4. Заключение. Предложен вариант реализации системы аналитического мониторинга для осуществления информационно-аналитического обеспечения научной деятельности на основе субпоисковой системы.

Предложенное техническое решение позволяет в автоматическом режиме осуществлять загрузку документов, соответствующих заданной тематике и обладает следующими достоинствами:

– значительное снижение требований к аппаратному обеспечению системы за счет использования ресурсов Интернет поисковых систем, по сравнению с использованием специализированной версии полнофункциональной Интернет поисковой системы;

– обеспечение фильтрации контента, предлагаемого выдачей поисковой системы, по критериям качества текста и соответствия глоссарию предметной области [5], а также возможность автоматического расширения областей поиска по сравнению с вариантом применения метапоиска.

Подобная субпоисковая система может решать задачу аналитического мониторинга по глоссариям предметной области исследований и войти в состав системы обеспечения научных исследований.

Литература

1. *Михайлов С.Н.* Способ тематической кластеризации текстовых документов на основе их инфологической обработки // Наукоёмкие технологии. 2012. № 9.
2. Метапоисковая система. — [Электронный ресурс]. — Доступ: http://ru.wikipedia.org/wiki/Метапоисковая_система.
3. *Александров В.В., Андреева Н.А., Кулешов С.В.* Методы построения информационно-логистических систем. СПб.: Изд-во Политехнического университета. 2006. 93 с.
4. *Александров В.В., Кулешов С.В., Цветков О.В.* Цифровая технология инфокоммуникации. Передача, хранение и семантический анализ текста, звука, видео. СПб.: Наука, 2008. 244 с.
5. *Александров В.В., Кулешов С.В.* Аналитический мониторинг Internet контента. Инфологический подход // Качество. Инновации. Образование. 2008. № 3. С. 68–70.

Поддержка исследований. В публикации представлены результаты исследований, поддержанные грантом РФФИ 13-07-00137.

Кулешов Сергей Викторович — д.т.н.; ведущий научный сотрудник лаборатории автоматизации научных исследований СПИИРАН. Область научных интересов: инфологические информационные системы, инфокоммуникационные системы, гибридные кодеки, обработка потоков видеоданных. Число научных публикаций — 70. kuleshov@iias.spb.su, sial.iias.spb.su; СПИИРАН, 14-я линия В.О., 39, Санкт-Петербург, 199178, РФ; р.т. +7(812)323-5139, факс +7(812)328-4450.

Kuleshov Sergey Victorovich — Ph.D., Dc.Sci.; Leading researcher, Laboratory of Research Automation, SPIIRAS. Research interests: infology information systems, infocommunication systems, hybrid codecs, video data streams processing. The number of publication — 70. kuleshov@iias.spb.su, sial.iias.spb.su; SPIIRAS, 14-th Line V.O., 39, St. Petersburg, 199178, Russia; office phone +7(812)323-5139, fax +7(812)328-4450.

Михайлов Сергей Николаевич — к.т.н.; доцент кафедры Защиты информации и систем связи Юго-Западного Государственного университета, г. Курск. Область научных интересов: средства диагностирования систем управления, моделирование процессов

функционирования средств диагностирования. Число научных публикаций — более 170. tk_kursk@mail.ru, ЮЗГУ, 305040, г.Курск, ул. 50 лет Октября, 94. т. +7 (4712) 57 55 53.

Mikhailov Sergey Nikolaevich — Ph.D.; Associate Professor, Department of Protection of information and communication systems of the Southwestern State University, Kursk. Research interest: means of diagnosing control systems, modeling of the functioning of diagnostics tools. The number of publication — 170. tk_kursk@mail.ru; 94, 50 let Oktyabrya, Kursk, 305040, Russia, phone +7 (4712) 57 55 53

Рекомендовано лабораторией автоматизации научных исследований СПИИРАН, заведующий лабораторией Александров В.В., д.т.н., проф.
Статья поступила в редакцию 16.09.2013.

РЕФЕРАТ

Кулешов С.В., Михайлов С.Н. **Вариант архитектуры субпоисковой системы для реализации функции аналитического мониторинга.**

При проведении научных исследований одним из первых этапов является подбор литературы по теме проводимого исследования. Традиционно этот этап выполняется вручную с использованием поисковых систем Интернет, чтения найденных электронных ресурсов, сохранения их копий (или ссылок), корректировки поискового запроса в соответствии со списком ключевых слов и фраз и повторения всех действий до накопления требуемого объема текстового материала, необходимого для продолжения работы. В случае, когда исследование длится достаточно долго или тема является основной для исследователя, интересным и крайне полезным является наблюдение динамики изменений (мониторинг) в публикациях по интересующей тематике.

Для решения перечисленных задач предлагается автоматизированная система мониторинга научных ресурсов, способная облегчить работу исследователей, обеспечив выполнение рутинных операций и предоставив больше функциональных возможностей по сравнению с традиционным подходом.

В качестве научно-технического решения построения системы мониторинга предполагается разработка системы субпоиска, использующей внешние поисковые системы, и позволяющей производить накопление рабочего массива документов по заданной тематике с возможностью их динамического обновления. В рамках субпоисковой системы обеспечиваются сервисные функции, необходимые при выполнении информационно-аналитического обеспечения научной деятельности: поиск по сформированной базе, визуализация динамики обновлений, оценка близости текстов и т. п.

Предложенное техническое решение позволяет в автоматическом режиме осуществлять загрузку документов, соответствующих заданной тематике, и обладает следующими достоинствами: значительное снижение требований к аппаратному обеспечению системы по сравнению с использованием специализированной версии полнофункциональной Интернет поисковой системы, обеспечение фильтрации контента, возможность автоматического расширения областей поиска по сравнению с вариантом применения метапоиска.

SUMMARY

Kuleshov S.V., Mikhailov S.N. **The variant of the subsearch engine architecture for analytical monitoring.**

When conducting research one of the first steps is the selection of literature on the research topic. Traditionally, this step is performed manually using Internet search engines, reading electronic resources, preservation copies (or links), adjusting the search query based on a list of keywords and phrases, and repeating all the actions to accumulate the required amount of textual material. When the research lasts long enough or is the main topic for researchers, interesting and extremely useful is the observation of the dynamics of change (monitoring) in publications on topics of interest.

To solve these problems automatic monitoring system of scientific resources is proposed that can facilitate the work of researchers, providing routine operations and providing more functionality than the traditional approach.

As scientific and technological solution for developing of monitoring system it is proposed to develop a system of sub-search which uses an external search engines, and allowing the accumulation of the work array of documents on a particular subject with the ability to dynamically update them. As part of the service provided by sub-search system are the functions required for the performance of information-analytical support of scientific activity: the search on the generated database, visualization of dynamic updates, estimation of the texts similarity, etc.

The proposed solution allows you to automatically upload documents relevant to the theme and has the following advantages: a significant reduction in hardware requirements compared to using a specialized version of a full-featured Internet search engine, providing content filtering, and the ability to automatically expand search regions compared to metasearch application.