УДК 004.75: 004.031.4

АВТОМАТИЗИРОВАННАЯ ИНФОРМАЦИОННАЯ СИСТЕМА ПОИСКА И АНАЛИЗА ИНФОРМАЦИИ В СЕТИ ИНТЕРНЕТ

Карташев Е.А., Царегородцев А.Л.

АУ Ханты-Мансийского автономного округа – Югры «Югорский научно-исследовательский институт информационных технологий», Ханты-Мансийск, e-mail: tsaregorodtseval@uriit.ru

В настоящее время в нашем динамично развивающемся информационном мире особую значимость приобретает способность принимать своевременные и правильные решения, которые невозможны без сбора, обработки, хранения, анализа большого объема информации и предоставления результатов их обработки пользователю. Одной из таких задач является оперативное выявление сайтов в сети Интернет, содержащих информацию, распространение которой в Российской Федерации запрещено. Перечень данной информации представлен в ч. 2 ст. 15.1 Федерального закона от 27.07.2006 № 149-ФЗ «Об информации, информационных технологиях и о защите информации». В данной статье рассмотрено построение информационно-аналитической системы, предназначенной для оперативного поиска информации в сети Интернет, распространение которой в Российской Федерации запрещено. Предложен подход к построению информационных систем, осуществляющих поиск информации в сетях общего пользования и обработку большого объема разнородных неструктурированных данных, которые представлены в различных форматах: текст, содержащий фрагменты из нескольких документов; аудио- и видеозаписи; изображения (фотографии и рисунки).

Ключевые слова: анализ данных, информационно-поисковые системы, неструктурированные данные, загрузка данных с сайтов сети Интернет

AUTOMATED INFORMATION SYSTEM SEARCH AND ANALYSIS OF INFORMATION IN THE INTERNET

Kartashev E.A., Tsaregorodtsev A.L.

Ugra Research Institute of Information Technologies, Khanty-Mansiysk, e-mail: tsaregorodtseval@uriit.ru

Today in our agile information world a capability to accept well-timed and right decisions, which are impossible without gathering, processing, storing, and analyzing big data, as well as providing results to a user, has a special meaning. One of such points is an immediate identifying of sites in Internet, which contain forbidden information (in the territory of the Russian Federation). The whole list of information, which is forbidden to distribute, is contained in the Federal law of the Russian Federation (№ 149 «About information, information technologies and information security»). Usually forbidden information is occurred on Internet sites, which are varied in technological and functional systems. In its turn forbidden information is unstructured and can be demonstrated in formats of text, which contains parts of different documents, audio and video tapes, images (photos, pictures). The article considers an approach of building information analysis systems, for immediate search of forbidden information in Internet, scanning information in public domain networks and processing big unstructured diversified data.

Keywords: data analysis, information retrieval systems, unstructured data, downloading data from the Internet sites

В настоящее время в нашем динамично развивающемся информационном мире особую значимость приобретает способность принимать своевременные и правильные решения, которые невозможны без сбора, обработки, хранения, анализа большого объема информации и предоставления результатов их обработки пользователю.

Одной из таких задач является оперативное выявление сайтов в сети Интернет, содержащих информацию, распространение которой в Российской Федерации запрещено. Перечень данной информации представлен в ч. 2 ст. 15.1 Федерального закона от 27.07.2006 № 149-ФЗ «Об информации, информационных технологиях и о защите информации». Зачастую такая информация представлена на сайтах в сети Интернет, которые могут существенно различаться как по использующимся в них технологиям, так и по их функциональности. В свою очередь информация не структурирована и может быть представлена в различных форматах: текст, содержащий фрагменты из нескольких документов; аудио- и видеозаписи; изображения (фотографии и рисунки).

На рынке существует ряд информационных систем, осуществляющих подобную обработку данных и применяемых в других сферах, но информация об их структуре и применяемых методах обработки данных не раскрывается. Зачастую они предоставляются по технологии SaaS (англ. software as a service), что неприемлемо с учетом специфики обрабатываемых данных.

Цель данной работы – предложить структуру информационной системы, обеспечивающей возможность оперативного получения неструктурированной информации с большого количества различных

сайтов в сети Интернет и ее хранения для последующей обработки, при этом должна предусматриваться возможность увеличения объема обрабатываемых данных за счет увеличения количества применяемого оборудования (горизонтальное масштабирование) и использование невысокопроизводительного серверного оборудования.

Разработка автоматизированной информационной системы поиска и анализа информации в сети Интернет (далее АИС Поиск) осуществлялась в Югорском научно-исследовательском институте информационных технологий и предназначена: для взаимодействия с сайтами в сети Интернет; хранения и анализа собранной информации; предоставления результатов обработки информации в виде отчетов пользователю.

Взаимодействие с сайтами в сети Интернет направлено на сбор с них исходной информации, предусматривает работу в режиме запрос – ответ по следующим направлениям: поиск требуемой информации на сайте сети Интернет; загрузка найденной информации в АИС Поиск; актуализация информации, хранящейся в АИС Поиск, за счет сравнения с версией [3], расположенной на сайте сети Интернет (выполняется через определенный интервал времени, определяемый с учетом обновления информации).

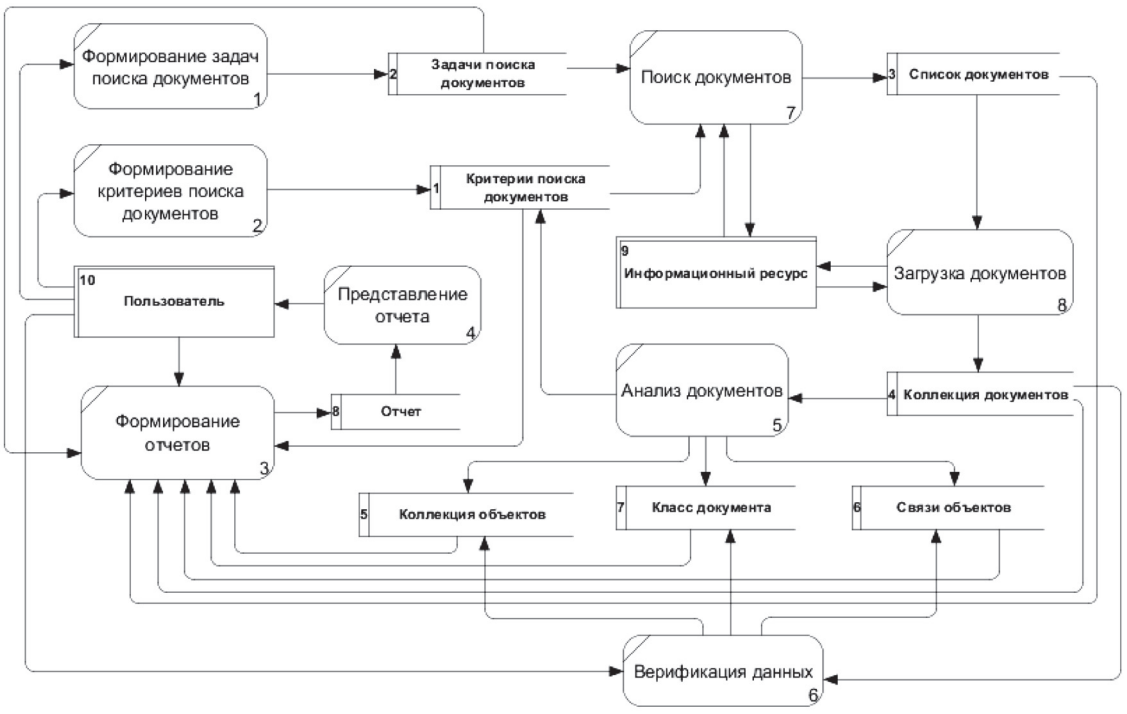
Хранение собранной информации с сайтов в сети Интернет предусматривает множество точек входа для сбора и обработки информации, при этом каждая из них может собирать и обрабатывать данные по своим уникальным правилам.

На этапе проектирования были разработаны диаграммы потоков данных, описывающие основные процессы АИС Поиск и потоки данных, циркулирующих в системе. На рисунке представлена контекстная диаграмма потоков данных АИС Поиск.

Рассмотрим процессы контекстной диаграммы подробнее.

1. Формирование критериев поиска документов (ключевые слова, тематические фразы, поисковые запросы, образцы изображений, фрагменты аудио- и видеозаписей) – определяются требования к содержанию документов, которые должны быть найдены на информационных ресурсах, расположенных в сети Интернет, и загружены в базу данных. Первоначальное наполнение осуществляется оператором, в последующем уточняется по результатам анализа документов.

2. Формирование задач поиска документов – определяется режим поиска документов на информационных ресурсах с учетом имеющихся возможностей, периодичности обновления информации и приоритетов пользователя. Формируется в виде задачи,



Контекстная диаграмма потоков данных АИС Поиск

для которой определяется: время запуска, информационные ресурсы, критерии поиска документов.

3. Поиск документов – обеспечивает выполнение задач по поиску документов: периодическая проверка наличия требующих запуска задач поиска документов, выполнение задачи поиска документов в рамках которой по количеству установленных критериев поиска документов и информационных ресурсов выполняется набор действий:

а) формирование запроса на получение данных к информационному ресурсу на основе определенных критериев поиска документов и его синтаксиса;

б) направление запроса на получение данных в информационный ресурс и ожидание ответа;

в) обработка ответа информационного ресурса (запись ссылок на найденные документы в базу данных).

4. Загрузка документов – обеспечивает загрузку документа по найденной ссылке: проверка доступности документа по найденной ссылке; сравнение загруженного документа с предыдущей версией, при ее наличии (проверка на наличие изменений) в базе данных; запись загруженного документа в базу данных.

5. Анализ документов – обеспечивает автоматическую обработку загруженных документов: извлечение объектов из документа (структурированные данные: ФИО, должности, название территорий и веществ, контактная информация, события и т.д.); определение характера связи для выявленных объектов: объект – субъект, негатив – позитив и т.д.; расчет вероятности отнесения документа к различным группам документов, ранее определенных пользователем (классификация документа); выявление похожих документов (с использованием методов классификации объектов по группам за счет выявления наперед неизвестных общих признаков (введен в 1939 году Robert Tryon) [1]); уточнение критериев поиска документов на основе ранее классифицированных и кластеризованных документов.

6. Формирование отчетов – подготовка данных для отображения пользователю (выполнение операций, которые не могут быть выполнены за время ожидания пользователем отклика АИС Поиск).

7. Представление отчетов – представление данных в виде отчетов на основе определенных шаблонов с учетом предпочтений пользователя, при этом ему предоставляется возможность установки фильтра для отбора данных в него включаемых.

8. Верификация данных – подтверждаются пользователем результаты анализа документов: классификация, извлеченные объекты, установленные связи.

По результатам изучения опыта построения подобных систем, в том числе представленных в [2, 4], была выбрана модульная архитектура системы. Использование модульного подхода в качестве основы для такого инструментария позволяет не только просто строить сложные приложения, собирая их из «кирпичиков», но и обеспечивать их взаимозаменяемость для доработки программного обеспечения и расширения возможностей информационных систем. Основные преимущества модульной архитектуры этим не ограничиваются. Также к ключевым особенностям выбранного подхода к построению АИС Поиск можно отнести возможность выборочной ее компоновки, многократное использование однажды написанного кода и разработанных классов [5].

В общем виде структура АИС Поиск состоит из следующих модулей:

– База данных (совокупность средств для обеспечения хранения и доступа к найденным данным).

– Интерфейс пользователя (предоставляет инструменты пользователю для просмотра имеющихся данных и результатов их обработки, а также по управлению работой каждого из модулей).

– Подсистема анализа (осуществляет обработку (классификация, определение объектов и связей) найденных данных).

– Подсистема сбора данных (реализует заданный пользователем алгоритм работы Модулей взаимодействия (запуск, формирование параметров) и обеспечивает загрузку получаемых от них данных в Базу данных).

– Модуль взаимодействия (обеспечивает получение данных с определенного информационного ресурса в соответствии с установленными параметрами).

Все эти собранные неструктурированные данные требуется быстро анализировать, что в свою очередь невозможно без соответствующей организации хранения этих данных. Тенденции последних лет показывают, что для хранения неструктурированных данных используются современные СУБД, сочетающие в себе гибкость модели хранилища документов и строгость и простоту реляционной модели.

Например, в СУБД PostreSQL 9.2 появилась поддержка типа данных JSON (JavaScript Object Notation), а в 9.3 добавились функции обработки значений в нём. Этот же тип данных теперь поддерживается и в MySQL начиная с версии 5.7.8. Аналогичный функционал есть и в СУБД Oracle, MSSQL.

Существует несколько подходов к хранению неструктурированных данных в информационных системах:

– непосредственно в базе данных, при этом большинство современных СУБД предусматривают для этого специализированный тип данных: JSONB в PostgreSQL, CLOB в Oracle и т.д.;

– вне базы данных (в виде файлов в соответствующих хранилищах), при этом в базе данных хранятся только ссылки на них. Основными недостатками данного варианта являются сложности с администрированием, обеспечением доступности и целостности данных. В свою очередь преимуществом данного подхода является возможность использования стандартных приложений по их обработке (просмотр), сокращение общего объема базы данных (не требуется выделять большой объем дискового пространства в одном месте), данные могут храниться на большом количестве различных серверов с небольшим объемом дискового пространства. На сегодняшний день данное направление активно поддерживается разработчиками СУБД и ведутся работы по устранению указанных недостатков, в частности в MS SQL Server 2012 появились таблицы FileTable для работы с файлами, а в Oracle – параметр SecureFiles для типа данных LOB.

Принимая во внимание, что наибольшую часть (объем) будут занимать неструктурированные данные, доступ к которым нужен будет эпизодически (на этапе загрузки для извлечения метаданных и несколько раз для демонстрации результатов пользователю), была предложена следующая структура: Файловый сервер – Драйвер доступа – СУБД.

В качестве файловых серверов было принято решение использовать сервера под управлением свободно распространяемой операционной системы Linux (Debian, или Astra Linux), а в качестве СУБД Postgres, так как она: свободно распространяемая, имеет развитые инструменты для полнотекстового поиска и может быть сертифицирована по требованиям безопасности информации например в составе операционной системы Astra Linux.

В соответствии с предложенным подходом нами в Югорском НИИ информационных технологий была осуществлена реализация АИС Поиск, которая используется компетентными ведомствами Ханты-Мансийского автономного округа – Югры для поиска доменных имен, указателей страниц сайтов в информационно-телекоммуникационной сети Интернет и сетевых адресов, позволяющих идентифицировать сайты в информационно-телекоммуникационной сети Интернет, содержащие информацию, распространение которой в Российской Федерации запрещено.

В настоящее время было обработано более 75 тыс. ссылок, загружено в базу данных более 21 тыс. уникальных документов. Для 922 документов было определено с высокой долей вероятности, что они содержат информацию, распространение которой в Российской Федерации запрещено, более 75 % из них были включены в соответствующий реестр, который ведется Роскомнадзором в соответствии с ч. 3 ст. 15.1 Федерального закона от 27.07.2006 № 149-ФЗ «Об информации, информационных технологиях и о защите информации».

В ходе опытной эксплуатации АИС Поиск получены положительные оценки от конечных пользователей, также ими отмечается предсказуемость появления документов в базе данных в зависимости от сформированных критериев поиска документов (результаты аналогичны полученным при ручном поиске) и снижение трудоемкости. По результатам также было рекомендовано ввести АИС Поиск в промышленную эксплуатацию.

В дальнейшем планируется проведение работ по повышению эффективности работы пользователей с АИС Поиск, в частности за счет внесения изменений в интерфейс пользователя, сокращению время отклика системы на действия пользователя за счет предварительной подготовки данных и повышению скорости работы алгоритмов обработки данных. Планируется также проведение работ по сравнению результатов классификации документов с использованием различных алгоритмов и методов.

Список литературы

1. Бериков В.С., Лбов Г.С. Современные тенденции в кластерном анализе // Всероссийский конкурсный отбор обзорно-аналитических статей по приоритетному направлению «Информационно-телекоммуникационные системы». – 2008. – 26 с.

2. Ерохин Г.Н., Дружинин В.А., Царегородцев А.Л., Махнева Т.В., Огородников И.Н., Карташев Е.А. Телемедицина отложенных консультаций на примере северных регионов // Информационно-измерительные и управляющие системы. – 2009. – Т. 7. – № 12. – С. 49–53.

3. Зеленков Ю.Г., Сегалович И.В. Сравнительный анализ методов определения нечетких дубликатов для WEB-документов // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: труды 9-й Всероссийской научной конференции RCDL'2007: Сб. работ участников конкурса. – Т. 1. – Переславль- Залесский: «Университет города Переславля», 2007. – С. 166–174.

4. Карташев Е.А., Самков Л.М. Онлайновая информационно-аналитическая система мониторинга индикаторов жизнеобеспечения территориальных объектов Управление большими системами: сборник трудов. – 2009. – № 24. – С. 112–129.

5. Макунин, Алексей Анатольевич. Технология построения модульных автоматизированных информационных систем для сложных предметных областей и ее применение на примере информационной поддержки системы муниципального заказа органов местного самоуправления: дис. ... канд. техн. наук: 05.13.11. – Томск, 2005. – 228 с.

References

1. Berikov V.S., Lbov G.S. Sovremennye tendencii v klasternom analize // Vserossijskij konkursnyj otbor obzorno-analiticheskih statej po prioritetnomu napravleniyu «Informacionno-telekommunikacionnye sistemy», 2008. 26 p.

2. Erohin G.N., Druzhinin V.A., Tsaregorodtsev A.L., Mahneva T.V., Ogorodnikov I.N., Kartashev E.A. Telemedicina otlozhennyh konsultacij na primere severnyh regionov // Informacionno-izmeritelnye i upravlyayushchie sistemy. 2009. T. 7. no. 12. pp. 49–53.

3. Zelenkov YU.G., Segalovich I.V. Sravnitelnyj analiz metodov opredeleniya nechetkih dublikatov dlya WEB-dokumentov // Trudy 9-j Vserossijskoj nauchnoj konferencii «EHlektronnye biblioteki: perspektivnye metody i tekhnologii, ehlektronnye kollekcii» RCDL2007: Sb. rabot uchastnikov konkursa. T. 1. Pereslavl- Zalesskij: «Universitet goroda Pereslavlya», 2007. pp. 166–174.

4. Kartashev E.A., Samkov L.M. Onlajnovaya informacionno-analiticheskaya sistema monitoringa indikatorov zhizneobespecheniya territorialnyh obektov Upravlenie bolshimi sistemami: sbornik trudov. 2009. no. 24. pp. 112–129.

5. Makunin, Aleksej Anatolevich. Tekhnologiya postroeniya modulnyh avtomatizirovannyh informacionnyh sistem dlya slozhnyh predmetnyh oblastej i ee primenenie na primere informacionnoj podderzhki sistemy municipalnogo zakaza organov mestnogo samoupravleniya: dissertaciya kandidata tekhnicheskih nauk: 05.13.11. Tomsk, 2005. 228 p.