Л. Н. Пирумова ; Моск. гос. ун-т культуры и искусств. – М., 2003. – 19 с.

ИНФОРМАЦИОННАЯ СИСТЕМА ИНТЕРНЕТ-МОНИТОРИНГА ПУБЛИКАЦИЙ: ФУНКЦИИ И СТРУКТУРА

С. Ф. Липницкий, Л. В. Степура, А. Г. Буравкин
Объединенный институт проблем информатики НАН Беларуси,
г. Минск, Беларусь

Введение

В настоящее время в связи с интенсивным увеличением объемов текстовой информации, представленной в электронном виде, все более необходимой становится разработка систем, обеспечивающих решение широкого круга задач мониторинга и аналитической обработки таких данных. При этом постоянно повышаются требования к эффективности процессов поиска текстов и формирования отчетных документов по результатам мониторинга. Все это существенно усложняет получение высоких характеристик работы информационных систем. Другими словами, результаты мониторинга должны отвечать приемлемым требованиям полноты и точности.

В существующих системах автоматической обработки текстовой информации используются подходы по увеличению эффективности поисковых механизмов, которые основаны главным образом на различных лингвистических и статистических методах, что не позволяет достичь требуемого качества работы данных систем. Значительными проблемами для современных информационных служб являются адаптация к информационным запросам конкретных пользователей, а также необходимость проведения углубленного мониторинга по интересующим их тематикам.

Целью данной работы является определение функций и разработка структуры информационной системы Интернет-мониторинга публикаций, в которой настройка на различные предметные области и создание соответствующих баз знаний будет

осуществляться полностью в автоматическом режиме без привлечения дополнительных специалистов [1].

Предлагаемая технология будет реализована с использованием подхода, основанного на применении в качестве знаний о предметной области тематических и динамических корпусов текстов и сформированных на их основе специализированных словарей базы знаний [2]. Данные корпусы текстов могут создаваться предварительно под прогнозируемые задачи, а также в оперативном режиме непосредственно при поиске информации путем объединения наборов документов, релевантных каждому конкретному тексту или запросу пользователя (динамические корпусы). Это предоставляет возможность осуществлять адаптацию к информационным потребностям пользователей, эффективно индексировать и искать не только полнотекстовые документы, но и краткие сообщения, объем которых мал и не позволяет выявить их статистические характеристики.

Функции системы интернет-мониторинга публикаций

К основным функциям информационной системы интернет-мониторинга публикаций относятся:

– индексирование текстовых документов из интернет-источников;
– поиск документов по результатам индексирования;
– реферирование текстовых документов;
– создание и актуализация словарей базы знаний;
– создание и актуализация корпусов текстов по различным предметным областям;
– хранение результатов интернет-мониторинга.

Мониторинг Интернета осуществляется тремя основными процедурами:

– автоматический поиск веб-страниц (по запросу пользователя);
– автоматический поиск текстовых сообщений на найденных веб-страницах;
– автоматизированная или автоматическая генерация отчетов по результатам мониторинга информации в Интернете (в зависимости от типа мониторинга).

Структура системы

Система интернет-мониторинга публикаций по космической тематике включает следующие четыре подсистемы:

– автоматизированное рабочее место (АРМ) эксперта-лингвиста;

– АРМ аналитика;

– подсистема поиска и аналитической обработки информации;

– подсистема накопления и поиска данных о научно-технических достижениях в требуемой предметной области.

АРМ эксперта-лингвиста предназначен для автоматизированного создания и актуализации баз данных и знаний. В базе данных эксперт-лингвист накапливает различные электронные документы, на основе которых формируются тематические корпусы текстов. В базе знаний представлены лингвистические словари.

АРМ аналитика используется при решении следующих основных задач:

– формирование набора интернет-сайтов для поиска и аналитической обработки информации в рамках требуемой предметной области;

– создание аналитических отчетов по результатам информационного мониторинга публикаций;

– построение и актуализация шаблонов отчетов по результатам интернет-мониторинга.

В состав подсистемы поиска и аналитической обработки информации в процессе мониторинга входят следующие информационно-программные средства:

– программы индексирования веб-страниц;

– программы документального поиска веб-страниц;

– программы фактографического поиска на найденных веб-страницах;

– шаблоны отчетов;

– тематические и полные корпусы текстов, а также служебные базы данных.

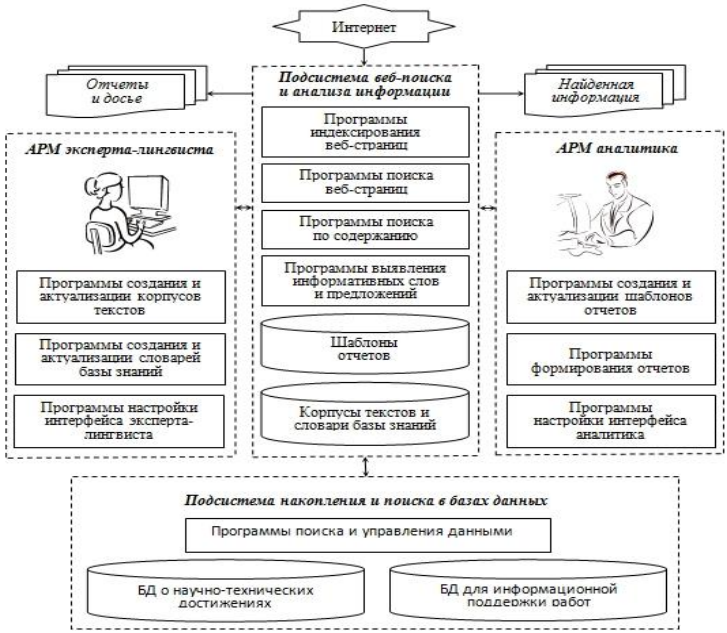

Рис. 1. Структурная схема взаимодействия подсистем

В состав подсистемы поиска и накопления данных входят следующие служебные базы данных:

– база данных о научно-технических достижениях в определенной тематической области;

– база данных для информационной поддержки работ в определенной тематической области.

Состав функций и задач базы данных и знаний

В базе данных системы интернет-мониторинга представлены тематические и полные корпусы текстов. Тематический корпус — это совокупность текстовых документов по конкретной тематике, характеризующей соответствующую предметную область. Полный корпус представляет собой объединение всех тематических корпусов. Кроме корпусов текстов в базе данных могут храниться различные полнотекстовые документы, их рефераты и прочая актуальная информация.

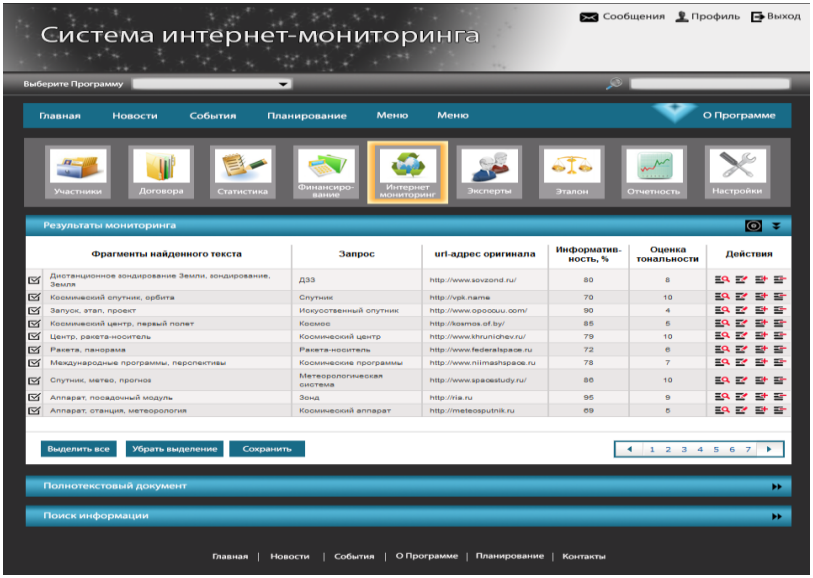

Рис. 2. Веб-интерфейс системы интернет-мониторинга

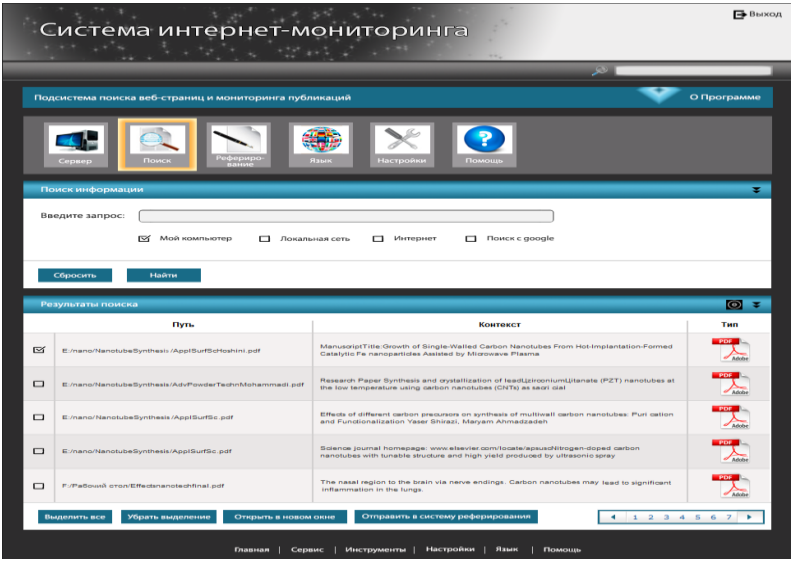

Рис. 3. Поиск веб-страниц и текстовых документов

Рис. 4 Реферирование текстовых документов

База знаний системы интернет-мониторинга включает следующие лингвистические словари, которые используются при индексировании веб-страниц и запросов пользователей:

– частотный словарь словоформ;
– частотный словарь вербально-ассоциативных пар слов;
– словарь словоизменительных парадигм;
– словарь синонимичных словоформ.

В настоящее время ведутся работы по созданию базовых модулей программного обеспечения системы Интернет-

мониторинга публикаций. Далее представлен веб-интерфейс различных модулей разрабатываемой информационной системы.

Заключение

В докладе сформулированы функции и представлена структура информационной системы интернет-мониторинга публикаций.

К числу наиболее актуальных можно отнести следующие задачи информационного мониторинга:

– подборка веб-страниц по запрашиваемой тематике (информирование пользователей о новых публикациях в их предметных областях, информация для принятия решений, деловая и экономическая разведка, тенденции развития и состояния рынков товаров и услуг и т. п.);

– информационный мониторинг текстовых документов по запрашиваемой тематике (подборка информативных выдержек из веб-страниц, дайджест новостей, выявление тонально окрашенной информации);

– классификация и рубрикация найденной информации;

– поиск документов по рубрикам;

– контроль обновляемости сайтов Интернета;

– формирование отчетов по результатам мониторинга.

Список использованных источников:

1. Липницкий, С. Ф. Моделирование информационного мониторинга Интернета на основе тематических корпусов текстов / С. Ф. Липницкий // Вес. Нац. акад. навук Беларусі. Сер. фіз.-тэхн. навук. – 2011. – № 3. – С. 92–99.

2. Липницкий, С. Ф. Поиск и реферирование текстовой информации в многоязычной среде / С. Ф. Липницкий, А. А. Мамчич, Л. В. Степура // Открытые семантические технологии проектирования интеллектуальных систем (OSTIS-2013) : материалы III Междунар. научн.-техн. конф. (Минск, 21–23 февр. 2013 г.) / Белорус. гос. ун-т информатики и радиоэлектроники ; [редкол.: В. В. Голенков и др.]. – Минск, 2013. – С. 229–232.