УДК 004.3.01

ПРЕДМЕТНО-ОРИЕНТИРОВАННАЯ ИНФОРМАЦИОННО-АНАЛИТИЧЕСКАЯ СИСТЕМА МОНИТОРИНГА НАУЧНЫХ ИССЛЕДОВАНИЙ ПО ПУБЛИКАЦИЯМ КОНФЕРЕНЦИЙ

[1]Наумец А.А., [1]Соколов В.Н., [2]Туманов В.Е.

[1]Научный центр РАН, Черноголовка;

[2]Институт проблем химической физики РАН, Черноголовка, e-mail: tve90@yandex.ru

В статье рассмотрена программно-технологическая архитектура информационно-аналитической системы мониторинга материалов научных и научно-технических конференций. Система предназначена для сбора, структуризации, хранения, поиска документов и производства оперативных справок, аналитических справок и отчетов. Система является компонентом автоматизации работы аналитиков в области научно-технической сферы: получение оперативных и аналитических справок, аналитических отчетов по различным поисковым профилям в рамках направлений научных исследований в определенной предметной области знаний. В системе реализована авторская методика анализа публикаций конференций по ряду наукометрических показателей. Также в системе предусмотрен рубрикатор, который относит входной документ к наиболее вероятному направлению научных и научно-практических исследований. Рубрикатор основан на использовании алгоритма классификации Байеса и использует извлеченные из документов рубрики.

Ключевые слова: информационно-аналитическая система, мониторинг, научно-технические конференции, программно-технологическая архитектура

INFORMATION-ANALYTICAL SYSTEM FOR MONITORING OF MATERIALS OF SCIENTIFIC AND TECHNICAL CONFERENCES

[1]Naumets A.A., [1]Sokolov V.N., [2]Tumanov V.E.

[1]Federal State Institution of Science Scientific Center of the Russian Academy of Sciences, Chernogolovka;

[2]Institute of Problems of Chemical Physics RAS, Chernogolovka, e-mail: tve90@yandex.ru

In article the program and technological architecture of information and analytical system of monitoring of materials of scientific and scientific and technical conferences is considered. The system is intended for collecting, structurization, storage, search of documents and production of operational references, analytical references and reports. The system is a component of automation of work of analysts in the field of the scientific and technical sphere: obtaining operational and research opinions, analytical reports on various search profiles within the directions of scientific researches in a certain subject domain of knowledge. In system the author's technique of the analysis of publications of conferences on a number of scientometric indicators is realized. Also the rubricator which carries the entrance document to the most probable direction of scientific and scientific and practical researches is provided in system. The rubricator is based on use of algorithm of classification of Bayes and uses the headings taken from documents.

Keywords: information-analytical system, monitoring, scientific and technical conference, software platform

Согласно заключениям экспертов Gartner Groups в течение нескольких лет будет наблюдаться *рост рынка инструментов и систем бизнес-аналитики* (Business Intelligence, BI) и *интеллектуального анализа данных* (Data mining). Этот вывод подтверждает то обстоятельство, что в условиях построения ведущими государствами экономики, основанной на знаниях, роль информации (и извлекаемых из нее знаний) как промышленного ресурса во многом определяет конкурентные способности как государств, так и отдельных компаний. С этой точки зрения информационно-аналитические системы конкурентной разведки можно рассматривать как особый класс BI систем.

В настоящее время на BI сегменте ИТ рынка, как российского, так и зарубежного, предлагается достаточно много готовых систем и программно-аппаратных решений для обеспечения работы аналитиков, в том числе в области конкурентной разведки. На эту тему существует много доступных обзоров, рекламных материалов и т.д., ориентированных на различные читательские аудитории. Можно утверждать, что функциональные требования к таким системам в целом на сегодняшний момент определены достаточно полно. Есть ряд интересных предложений по разработке высокопроизводительных программно-аппаратных платформ для эффективной бизнес-аналитики.

Тенденция разработки современных информационно-аналитических систем для конкурентной разведки (факт) – *их универсальность с точки зрения поддержки всего комплекса работ аналитика*. Эта тенденция будет сохраняться в течение ближайших 5–10 лет. Исключением являются программные продукты класса «условно бесплатные»,

которые используются в организациях, не имеющих отдела конкурентной разведки (или не имеющих возможности содержать такой отдел). Поэтому разработка специализированной предметно-ориентированной информационной системы для мониторинга научных публикаций является актуальной научной и технической задачей.

В настоящее время такие информационно-аналитические системы разрабатываются в рамках крупных проектов по созданию систем по актуальным научным исследованиям и систем различных фондов, финансирующих научные и научно-практические исследования. В качестве примеров можно привести европейский проект разработки системы euroCRIS [4], российский проект разработки системы SCIENCE INDEX [1], аналитическую подсистему РФФИ. Однако разработке и созданию информационно-аналитических систем для исследования и наукометрического анализа публикаций, ориентированных непосредственно для решения задач аналитиков в научной сфере, уделяется недостаточно внимания [2, 3].

Цель настоящей работы – описать архитектуру и функциональные возможности предметно-ориентированной информационно-аналитической системы мониторинга результатов научных исследований по материалам конференций (далее система).

Постановка задачи и метод решения

Предметная область – материалы конференций по определенным областям знаний по сравнению с другими источниками научной информации являются актуальными по времени – срок их публикации составляет в среднем от трех месяцев до года. Для описываемой системы выбор предметной области не имеет особого значения: тематика исследований может быть изменена, при этом система сохраняет свой основной функционал. Это обстоятельство обусловлено тем фактом, что в основу функционала положена специально разработанная система анализа научных публикаций и методика ее применения [5].

Основные требования к функционалу системы были следующие:

- Структуризация документа (статьи в сборники трудов) на вводе (парсинг документа);
- автоматическое формирование рубрикатора и понятийного тезауруса при вводе документов;
- семантическая разметка документа при вводе на двух уровнях: степень завершенности результатов и принадлежность к определенной рубрике (направлению исследований);
- если направление исследований не следует из обрабатываемого пакета документов, предусмотрено автоматическое отнесение документов к уже известным в системе рубрикам на основе алгоритма машинного обучения по Байесу;
- динамически настраиваемый интерфейс (автоматическое формирование имен полей экранных форм и их привязка к полям базы данных системы);
- формирование комплекса встроенных справок и аналитических отчетов.

В основу программно-технологических решений была положена специально разработанная методика анализа данных. В качестве программного решения была выбрана клиент-серверная архитектура с RICH-клиентом и использованием программных реактивных агентов и одного интеллектуального агента со специально настраиваемой базой знаний. Реализация системы была выполнена на программно-технологической платформе MS Windows Server 2008, IIS для веб-сервера и MS SQLServer 2008 для реализации БД.

Описание программно-технологической архитектуры системы

На рис. 1 показан бизнес-процесс обработки документов в системе. При вводе отдельно обрабатывается оглавление документа, выделяются тематические направления (секции) работы конференции, а затем последовательно структурируются статьи конференций (название, авторы, организация, страны, гранты, ключевые слова, аннотации, текст). Структурированная информация заносится в базу данных (БД) системы.

При проектировании БД была использована методика многомерного проектирования (для схемы снежинка), которая затем была оптимизирована под типовые запросы.

На рис. 2 показана программно-технологическая архитектура системы.

В процессе разработки и создания были решены, помимо общих для систем конкурентной разведки, следующие задачи:

- Задача автоматической классификации (кластеризации) входного потока, его фильтрации, сжатию данных ранжирования и доставки потребителю (аналитику) в виде пригодном для обозрения и анализа.
- Задача построения ассоциативных тезаурусов, достаточно интеллектуальных, чтобы связывать модели предметных областей (интересов аналитиков) с расклассифицированным предварительно потоком данных с учетом смежных областей.

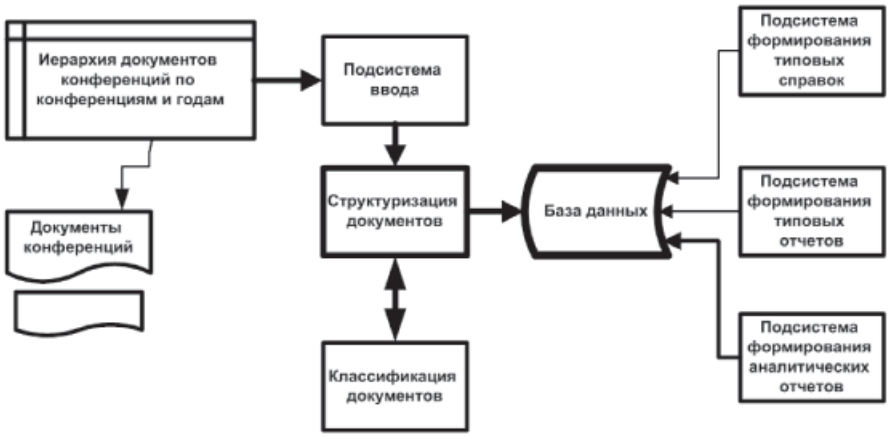

Рис. 1. Бизнес-процесс обработки документов в системе

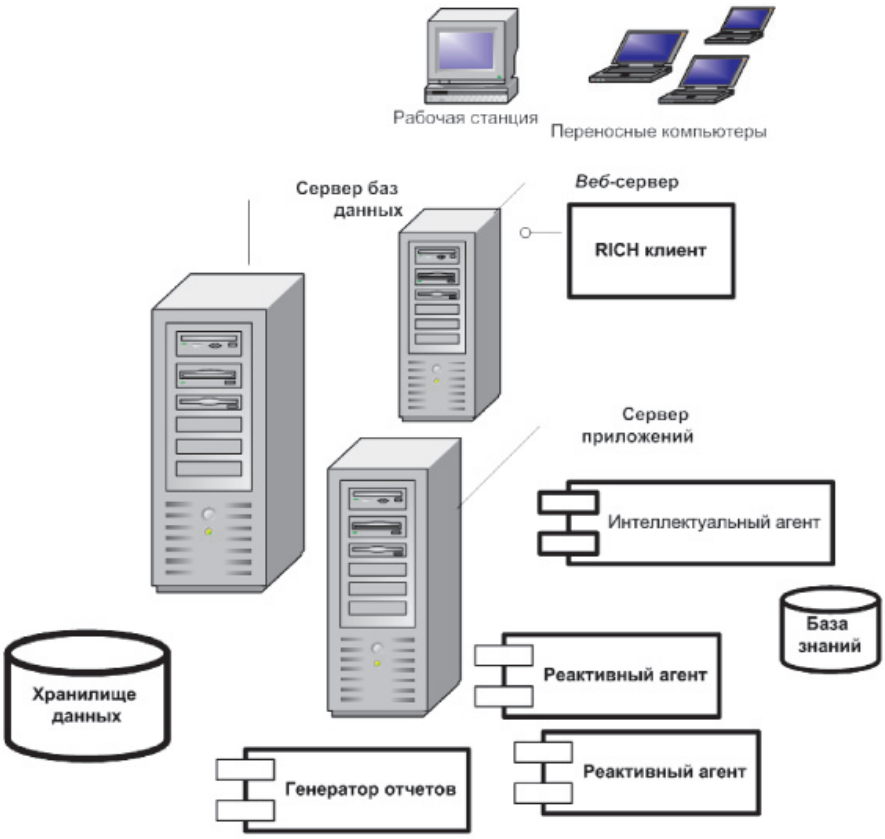

Рис. 2. Программно-технологическая архитектура системы

● Задача создания интеллектуального человеко-машинного интерфейса, с помощью которого аналитик вернет себе контроль над своими данными, т.е. создание своеобразной интеллектуальной рабочей доски (тетради), которая связывает данные, раскрывает данные, обеспечивает их фиксацию, организацию, анализ, визуализацию и публикацию.

Полуавтоматическая система ввода данных

Система включает в себя подсистему автоматического ввода данных, основанную на парсинге документов, поступающих на вход системы. При начале ввода документов в систему пользователям предоставляется статистика ввода, как показано на рис. 3, 4, приведен пример такого парсинга документов.

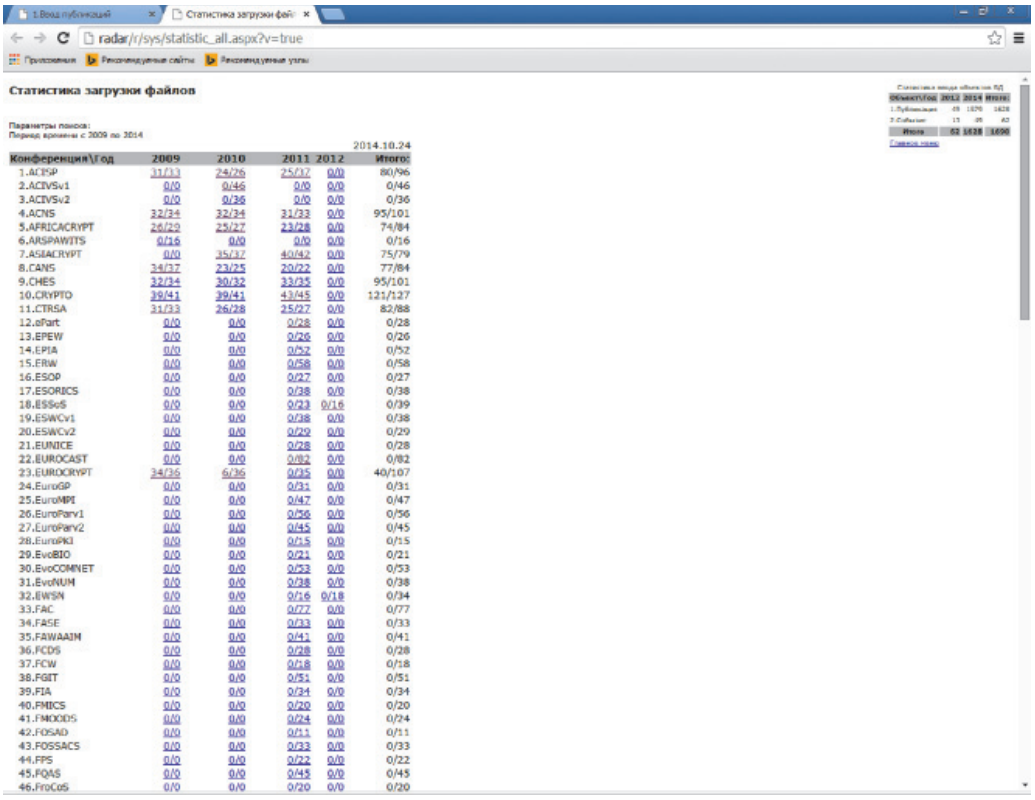

Рис. 3. Экранная форма «Статистика ввода»

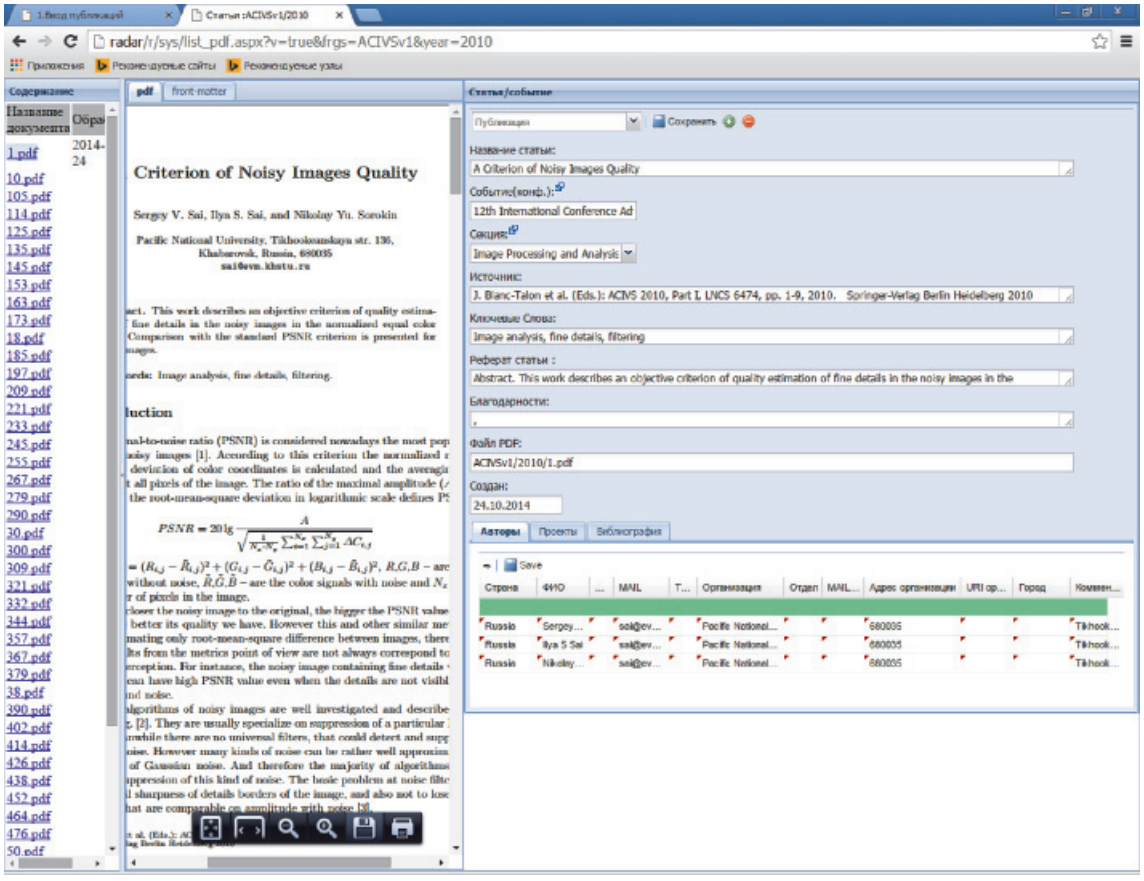

Рис. 4. Экранная форма «Структуризация документа»

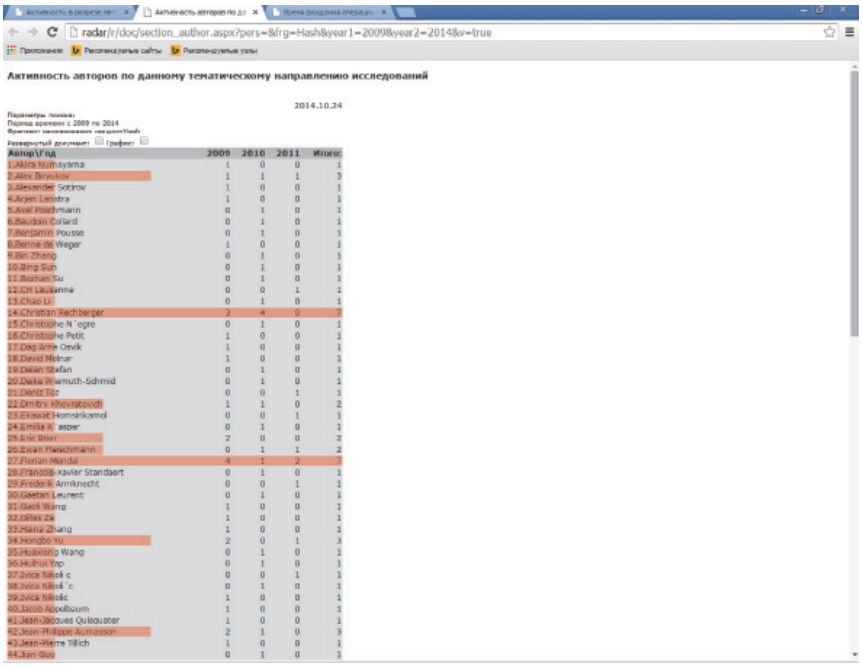

Рис. 5. Отчет «Активность авторов по выделенному тематическому направлению исследований»

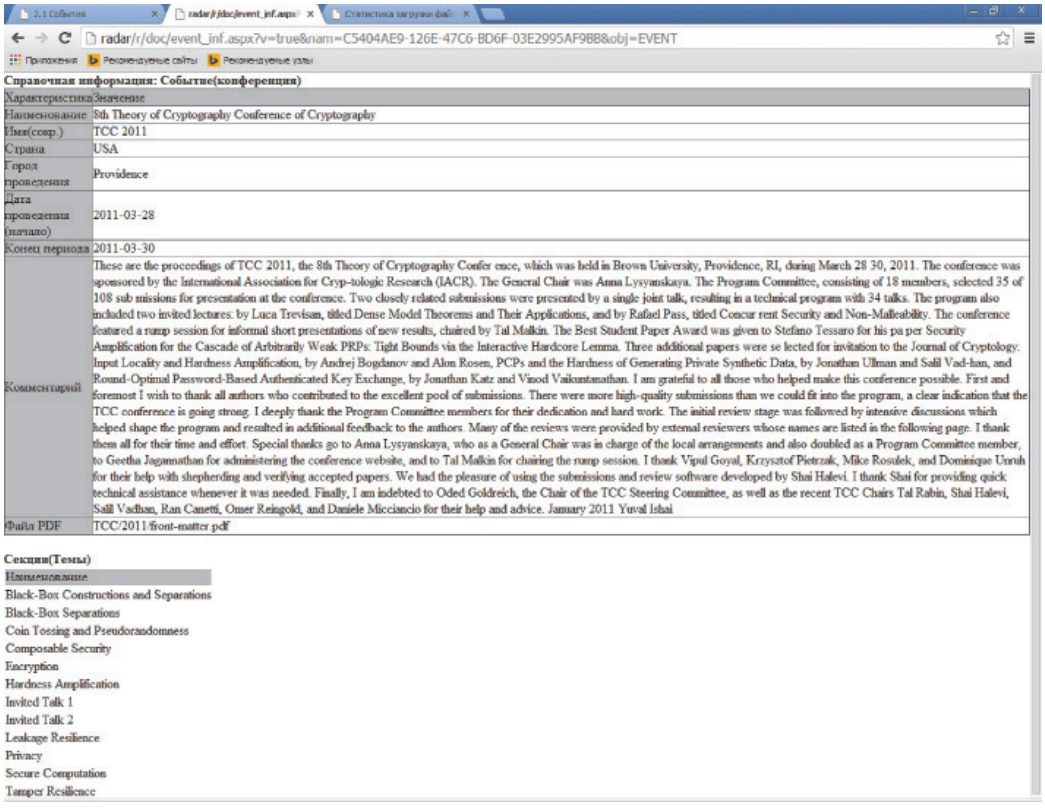

Рис. 6. Справка о конференции

Подсистема получения динамически формируемых справок и аналитических отчетов

На рис. 5–7 приведены типовые отчеты, генерируемые системой по запросу:

- отчет «Активность авторов по выделенному тематическому направлению исследований»;
- справка о конференции;
- отчет «Активность авторов по выделенному тематическому направлению исследований» в развернутом виде.

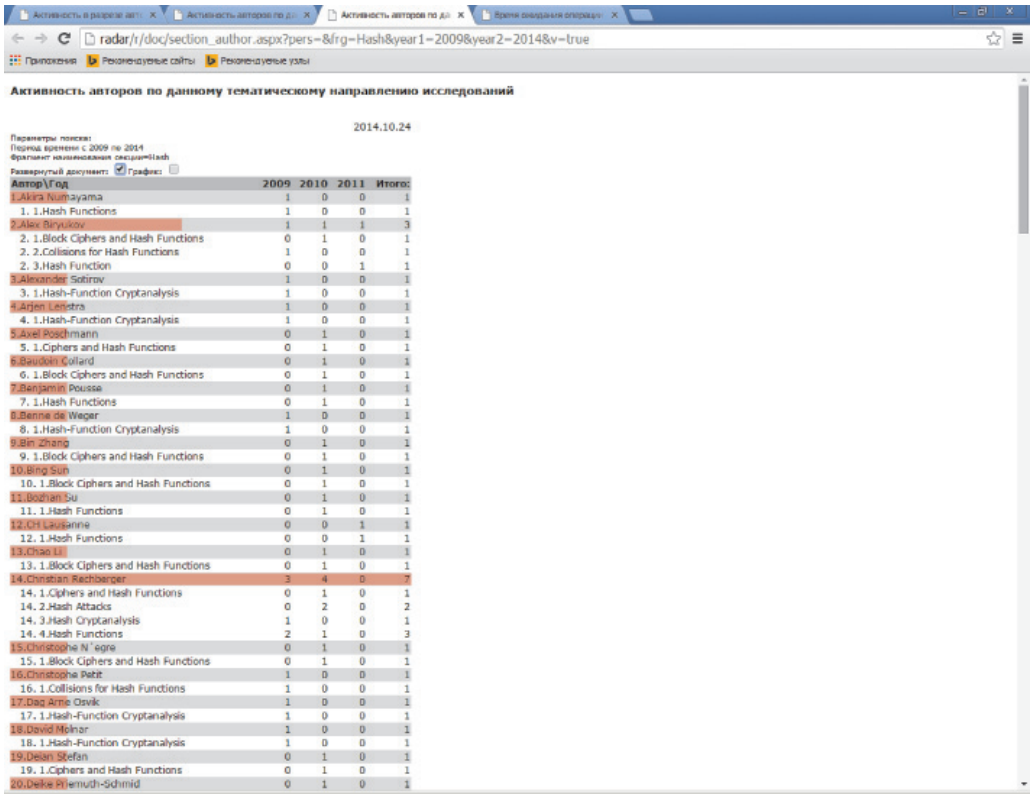

Рис. 7. Отчет «Активность авторов по выделенному тематическому направлению исследований» в развернутом виде

Заключение

Разработана информационно-аналитическая система мониторинга материалов научных и научно-практических конференций для информационного обеспечения работы аналитиков в области научно-технологической сферы.

Функционал системы разработан на основе оригинальной авторской методики наукометрического анализа тематических массивов публикаций. Разработан взаимосвязанный комплекс аналитических справок и отчетов для информационного обеспечения работы аналитиков в научно-технической сфере.

Разработанная информационно-аналитическая система может быть дополнена средствами семантического поиска и когнитивной визуализации на основе создания предметной онтологии.

Список литературы

1. Еременко Г.О. Российский индекс научного цитирования и информационно-аналитическая система SCIENCE INDEX [Электронный ресурс] // Science index: аналитические инструменты и сервисы для оценки научной деятельности: материалы научно-практической конференции: сайт. – URL: http://science.usue.ru/index/news/745-1.html (дата обращения 18.02.2016 г.).

2. Маркусова В.А. Информационные ресурсы для мониторинга российской науки // Вестник РАН. – 2005. – Т. 75, № 7. – С. 607–612.

3. Хайтун С.Д. Проблемы количественного анализа науки. – М.: Наука, 1989. – 280 с.

4. European current research information systems (CRIS) community. 2016. – URL: // http:www.eurocris.org (дата обращения 18.02.2016 г.).

5. Naumets A.A. Approccio allo sviluppo di metodi analisi scientometricisullesempio la pubblicazione di conferenze scientifiche / A.A. Naumets, V.N. Sokolov, V.E. Tumanov // Italian Science Review. – 2015. Vol. 8, № 29. – P. 30–39. – URL: http://www.ias-journal.org/archive/2015/august/Naumets.pdf (дата обращения 18.02.2016 г.).

References

1. Eremenko G.O. Rossijskij indeks nauchnogo citirovanija i informacionno-analiticheskaja sistema SCIENCE INDEX [Jelektronnyj resurs] // Materialy nauchno-prakticheskoj konferencii Science index: analiticheskie instrumenty i servisy dlja ocenki nauchnoj dejatelnosti: sajt. URL: http://science.usue.ru/index/news/745-1.html (дата обращения 18.02.2016).

2. Markusova V.A. Informacionnye resursy dlja monitoringa rossijskoj nauki // Vestnik RAN. 2005. T. 75, no. 7. pp. 607–612.

3. Hajtun S.D. Problemy kolichestvennogo analiza nauki. M.: Nauka, 1989. 280 p.

4. European current research information systems (**CRIS**) community. 2016. URL: // http:www.eurocris.org (дата обращения 18.02.2016).

5. Naumets A.A. Approccio allo sviluppo di metodi analisi scientometricisullesempio la pubblicazione di conferenze scientifiche / A.A. Naumets, V.N. Sokolov, V.E. Tumanov // *Italian Science Review*. 2015. Vol. 8, no. 29. pp. 30–39. URL: http://www.ias-journal.org/archive/2015/august/Naumets.pdf (дата обращения 18.02.2016).