

УДК 004.3.01

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

¹Наумец А.А., ¹Соколов В.Н., ²Туманов В.Е.¹Научный центр РАН, Черноголовка;²Институт 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

¹Naumets A.A., ¹Sokolov V.N., ²Tumanov V.E.¹Federal State Institution of Science Scientific Center of the Russian Academy of Sciences, Chernogolovka;²Institute 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, structuring, 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

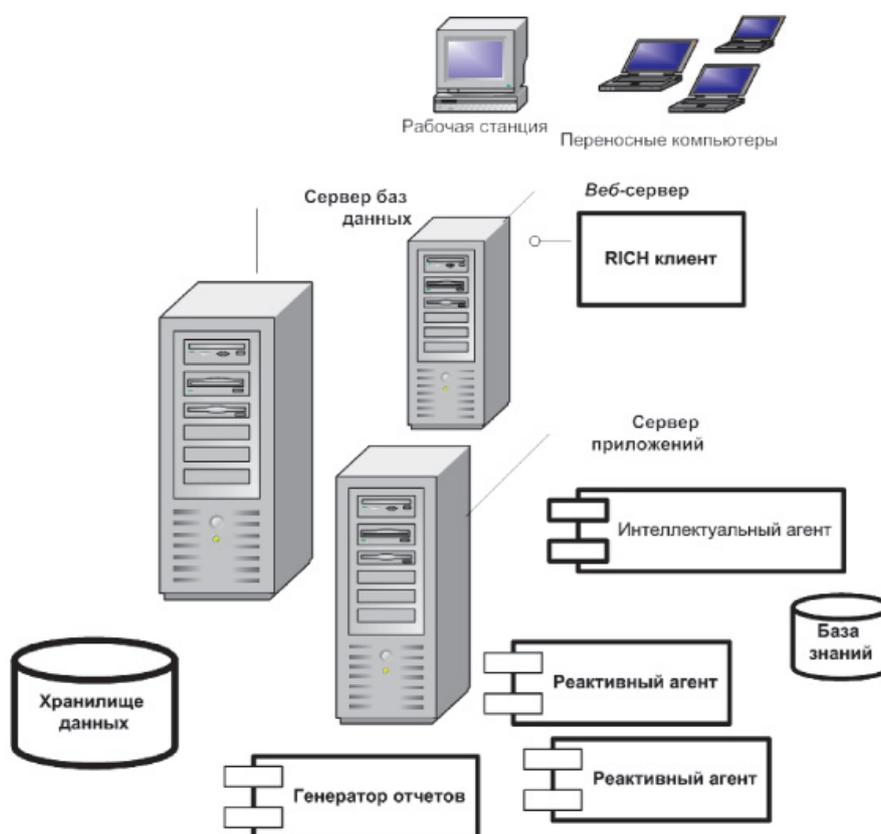


Рис. 2. Программно-технологическая архитектура системы

● 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.

Конференция\год	2009	2010	2011	2012	итого
1.ACISP	11/13	24/26	25/17	0/0	60/56
2.ACIVSv1	0/0	0/46	0/0	0/0	0/46
3.ACIVSv2	0/0	0/36	0/0	0/0	0/36
4.ACNS	22/24	22/24	21/32	0/0	95/101
5.AFRICACRYPT	24/22	23/27	23/28	0/0	74/84
6.ASPAWITS	0/16	0/0	0/0	0/0	0/16
7.ASIACRYPT	0/0	15/17	40/42	0/0	75/79
8.CANS	24/37	23/25	20/22	0/0	77/84
9.CHES	22/34	30/32	33/35	0/0	95/101
10.CRYPTO	39/41	39/41	41/45	0/0	121/127
11.CTRSA	31/33	26/28	25/27	0/0	82/88
12.dart	0/0	0/0	0/26	0/0	0/26
13.EPEW	0/0	0/0	0/26	0/0	0/26
14.EPIA	0/0	0/0	0/52	0/0	0/52
15.ERW	0/0	0/0	0/58	0/0	0/58
16.ESDP	0/0	0/0	0/27	0/0	0/27
17.ESORICS	0/0	0/0	0/38	0/0	0/38
18.ESSoS	0/0	0/0	0/23	0/16	0/39
19.ESWCV1	0/0	0/0	0/38	0/0	0/38
20.ESWCV2	0/0	0/0	0/29	0/0	0/29
21.EUNICE	0/0	0/0	0/28	0/0	0/28
22.EUROCAST	0/0	0/0	0/62	0/0	0/62
23.EUROCRYPT	34/36	6/36	0/35	0/0	40/107
24.EuroSP	0/0	0/0	0/31	0/0	0/31
25.EuroSPI	0/0	0/0	0/42	0/0	0/42
26.EuroSPv1	0/0	0/0	0/56	0/0	0/56
27.EuroSPv2	0/0	0/0	0/45	0/0	0/45
28.EuroPKI	0/0	0/0	0/15	0/0	0/15
29.EvoBIO	0/0	0/0	0/21	0/0	0/21
30.EvoCOMNET	0/0	0/0	0/53	0/0	0/53
31.EvoNUM	0/0	0/0	0/38	0/0	0/38
32.EWSSN	0/0	0/0	0/16	0/18	0/34
33.FAC	0/0	0/0	0/27	0/0	0/27
34.FASE	0/0	0/0	0/33	0/0	0/33
35.FAWAAM	0/0	0/0	0/41	0/0	0/41
36.FCDS	0/0	0/0	0/28	0/0	0/28
37.FCW	0/0	0/0	0/18	0/0	0/18
38.FDIT	0/0	0/0	0/51	0/0	0/51
39.FIA	0/0	0/0	0/24	0/0	0/24
40.FMICS	0/0	0/0	0/20	0/0	0/20
41.FMOODS	0/0	0/0	0/24	0/0	0/24
42.FOSAD	0/0	0/0	0/11	0/0	0/11
43.FOSSACS	0/0	0/0	0/33	0/0	0/33
44.FPS	0/0	0/0	0/22	0/0	0/22
45.FQAS	0/0	0/0	0/48	0/0	0/48
46.FROCUS	0/0	0/0	0/20	0/0	0/20

Рис. 3. Экранная форма «Статистика ввода»

Criterion of Noisy Images Quality

Sergey V. Sai, Ilya S. Sai, and Nikolay Yu. Sorokin

Pacific National University, Tikhookevskaya str. 136,
Khabarovsk, Russia, 680035
sai@vsn.khpu.ru

Abstract: This work describes an objective criterion of quality estimation of fine details in the noisy images in the normalized equal color. Comparison with the standard PSNR criterion is presented for images.

Index Terms: Image analysis, fine details, filtering.

Introduction: Peak-signal-to-noise ratio (PSNR) is considered nowadays the most popular noisy images [1]. According to this criterion the normalized root-mean-square deviation of color coordinates is calculated and the average deviation of color coordinates is calculated and the average deviation of all pixels of the image. The ratio of the maximal amplitude of the root-mean-square deviation in logarithmic scale defines PSNR:

$$PSNR = 20 \lg \frac{A}{\sqrt{\frac{1}{N_x N_y} \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} \Delta C_{i,j}^2}}$$

where $\Delta C_{i,j} = (R_{i,j} - \hat{R}_{i,j})^2 + (G_{i,j} - \hat{G}_{i,j})^2 + (B_{i,j} - \hat{B}_{i,j})^2$, R, G, B – are without noise, $\hat{R}, \hat{G}, \hat{B}$ – are the color signals with noise and N_x, N_y – are the number of pixels in the image.

Close the noisy image to the original, the bigger the PSNR value, the better its quality we have. However this and other similar metrics that use only root-mean-square difference between images, differ from the metric point of view are not always correspond to perception. For instance, the noisy image containing fine details can have high PSNR value even when the details are not visible and noise is high.

Algorithms of noisy images are well investigated and described in [2]. They are usually specialize on suppression of a particular kind of noise. There are no universal filters, that could detect and suppress noise. However many kinds of noise can be rather well approximated by Gaussian noise. And therefore the majority of algorithms of suppression of this kind of noise. The basic problem at noise filtering is the sharpness of details borders of the image, and also not to lose them is comparable complexity with noise [3].

Authors:

Страна	Имя	Фамилия	Адрес организации	ORCID	Комментарий
Russia	Sergey V. Sai	sai@vsn.khpu.ru	Pacific National University	680035	Tikhookevskaya str. 136, Khabarovsk, Russia
Russia	Ilya S. Sai	sai@vsn.khpu.ru	Pacific National University	680035	Tikhookevskaya str. 136, Khabarovsk, Russia
Russia	Nikolay Yu. Sorokin	sorokin@vsn.khpu.ru	Pacific National University	680035	Tikhookevskaya str. 136, Khabarovsk, Russia

Рис. 4. Экранная форма «Структуризация документа»

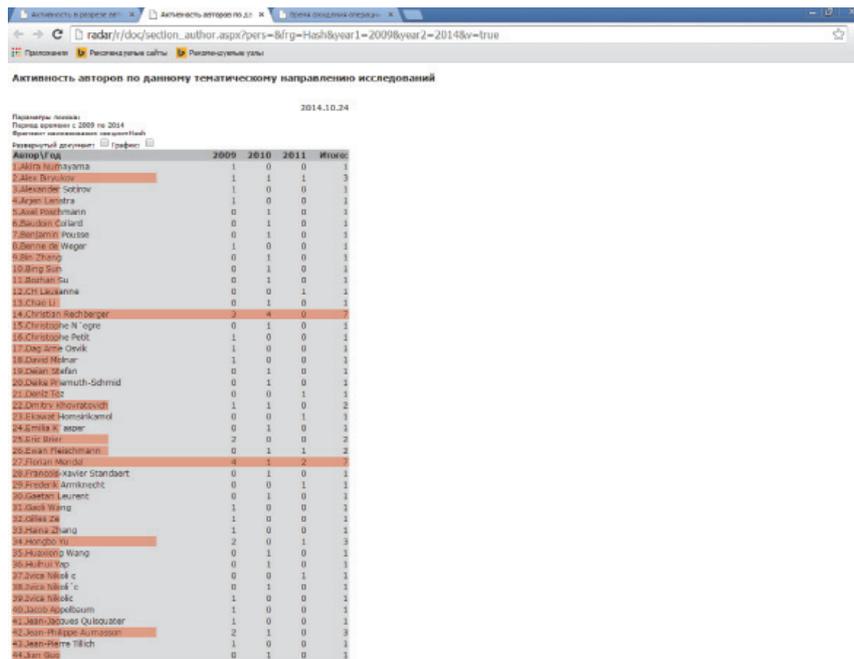


Рис. 5. Отчет «Активность авторов по выделенному тематическому направлению исследований»

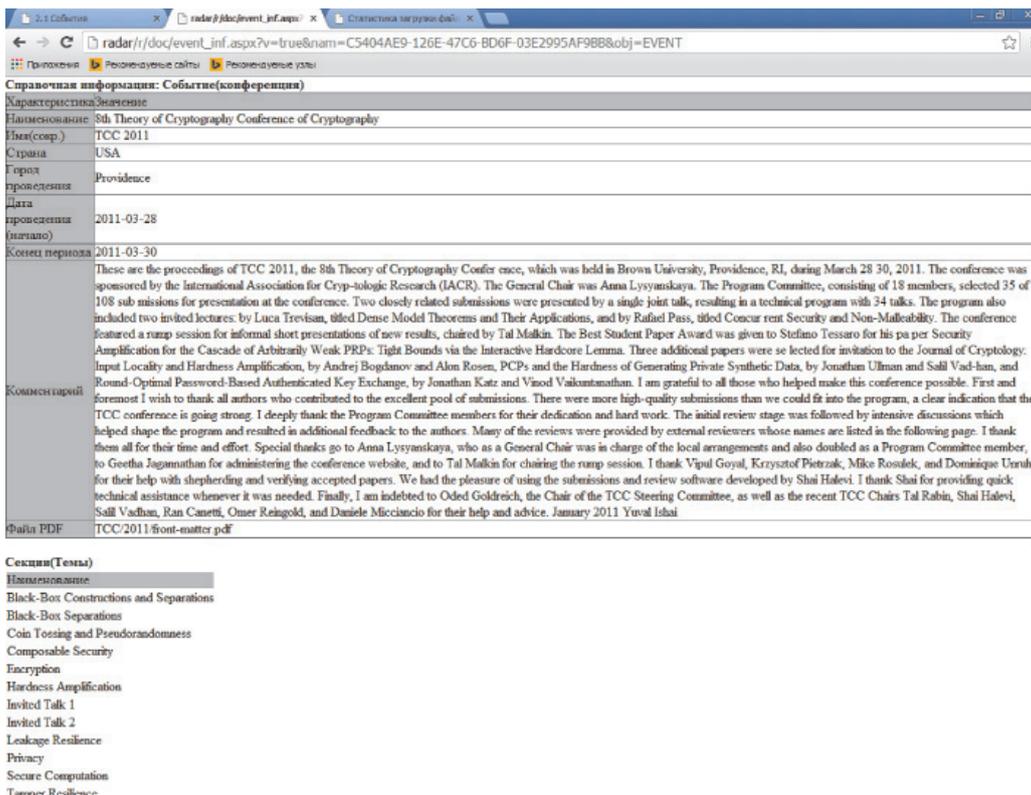


Рис. 6. Справка о конференции

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.

Активность авторов по данному тематическому направлению исследований

Параметры поиска: 2014.10.24
Период: все время с 2009 по 2014
Фрагмент: полевые значения секции hash
Развернутый документ: Графики:

Автор/Год	2009	2010	2011	Итого:
1. Alex Nishiyama	1	0	0	1
1. 1.Hash Functions	1	0	0	1
2. Alex Biryukov	1	1	1	3
2. 1.Block Ciphers and Hash Functions	0	1	0	1
2. 2.Collisions for Hash Functions	1	0	0	1
2. 3.Hash Function	0	0	1	1
3. Alexander Sotirov	1	0	0	1
3. 1.Hash-Function Cryptanalysis	1	0	0	1
4. Gilles Leurent	1	0	0	1
4. 1.Hash-Function Cryptanalysis	1	0	0	1
5. Axel Poschmann	0	1	0	1
5. 1.Ciphers and Hash Functions	0	1	0	1
6. Aurélien C. Allard	0	1	0	1
6. 1.Block Ciphers and Hash Functions	0	1	0	1
7. Benjamin Housheer	0	1	0	1
7. 1.Hash Functions	0	1	0	1
8. Sennae de Weeger	1	0	0	1
8. 1.Hash-Function Cryptanalysis	1	0	0	1
9. Bin Zhang	0	1	0	1
9. 1.Block Ciphers and Hash Functions	0	1	0	1
10. Bin Wang	0	1	0	1
10. 1.Block Ciphers and Hash Functions	0	1	0	1
11. Bojhan Li	0	1	0	1
11. 1.Hash Functions	0	1	0	1
12. Oh-Laurin	0	0	1	1
12. 1.Hash Functions	0	0	1	1
13. Chao Li	0	1	0	1
13. 1.Block Ciphers and Hash Functions	0	1	0	1
14. Christian Bachmeier	2	4	0	6
14. 1.Ciphers and Hash Functions	0	1	0	1
14. 2.Hash Attacks	0	2	0	2
14. 3.Hash Cryptanalysis	1	0	0	1
14. 4.Hash Functions	2	1	0	3
15. Christophe N. egré	0	1	0	1
15. 1.Block Ciphers and Hash Functions	0	1	0	1
16. Christophe Pezz	1	0	0	1
16. 1.Collisions for Hash Functions	1	0	0	1
17. Daq Arno Osvik	1	0	0	1
17. 1.Hash-Function Cryptanalysis	1	0	0	1
18. David Meinar	1	0	0	1
18. 1.Hash-Function Cryptanalysis	1	0	0	1
19. Debra Stefan	0	1	0	1
19. 1.Ciphers and Hash Functions	0	1	0	1
20. Deke P. emuth-Schimid	0	1	0	1

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 citirovani-ja i informacionno-analiticheskaja sistema SCIENCE INDEX [Jelektronnyj resurs] // Materialy nauchno-prakticheskoy konferencii Science index: analiticheskie instrumenty i servisy dlja ocenki nauchnoj dejatel'nosti: 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).