Архитектура программной системы

Программная система предназначена для интеллектуального анализа наукометрических данных – некоторой совокупности значений наукометрических показателей публикационной активности для набора авторов. Пользователь программной системы представлен двумя ролями (актерами): автор (преподаватель, исследователь, имеющий публикации), использующий основной функционал программной системы для получения индивидуальных рекомендаций, и руководитель, использующий программную систему для анализа публикационной активности авторского коллектива, а также готовящий программную систему к работе с автором.

Опишем кратко функциональные возможности (варианты использования) программной системы (рис. 1).

1. Формирование публикационного рейтинга. Прецедент, предоставляющий пользователю возможность просмотра таблицы рейтинга авторов, построенной по наукометрическим данным, ранее загруженным в систему (на основе оригинальной методики оценки публикационной активности).

2. Оценка публикационного потенциала. Позволяет пользователю ознакомиться с результатами прогноза возможностей авторов по переходу к лучшему квартилю в публикационном рейтинге (на основе поиска ассоциативных правил).

3. Поиск групп публикационной активности. В рамках данного прецедента выполняется формирование групп авторов, близких по своей публикационной активности, с выделением их типичных представителей (на основе кластеризации с выделением оптимального числа кластеров или с разбиением на заданное число кластеров).

4. Получение индивидуальных рекомендаций. Позволяет пользователю получить набор рекомендаций по продвижению в рейтинге публикационной активности и реализации публикационного потенциала (на основе анализа места автора в публикационном рейтинге относительно лучшего, типичного и худшего авторов, а также подбора журналов, в которых опубликованы работы авторов).

5. Подготовка наукометрических данных. Обобщенный прецедент, доступный только руководителю, реализующий программные возможности по подготовке и предобработке наукометрических данных авторов, их загрузке в программную систему и по необходимости обновлению.

6. Настройка аналитических моделей. Выполняется руководителем. Прецедент, реализующий программные возможности по настройке рабочих методик, а также параметров моделей интеллектуального анализа наукометрических данных, в дальнейшем используемых системой для работы.

Программная система состоит из трех основных компонентов, а также использует внешнюю интеллектуальную систему, обрабатывающую данные на платформе KNIME Analytics Platform (www.knime.org/knime-analytics-platform). Основные компоненты предоставляют и используют соответствующие интерфейсы (рис. 2):

– пользовательский интерфейс программной системы и/или интерфейс программирования (компонент Gateway);

– компонент хранения данных (компонент DataSystem);

– базовая вычислительная подсистема (компонент BaseSystem), реализующая взаимодействие с хранилищем данных, некоторые вычисления и обработку результатов работы внешней интеллектуальной системы KNIME Analytics Platform.

Автор в программной системе работает с поименованным набором наукометрических данных, предоставляемых хранилищем наукометрических

Рис. 1. Диаграмма вариантов использования

Fig. 1. Use case diagram

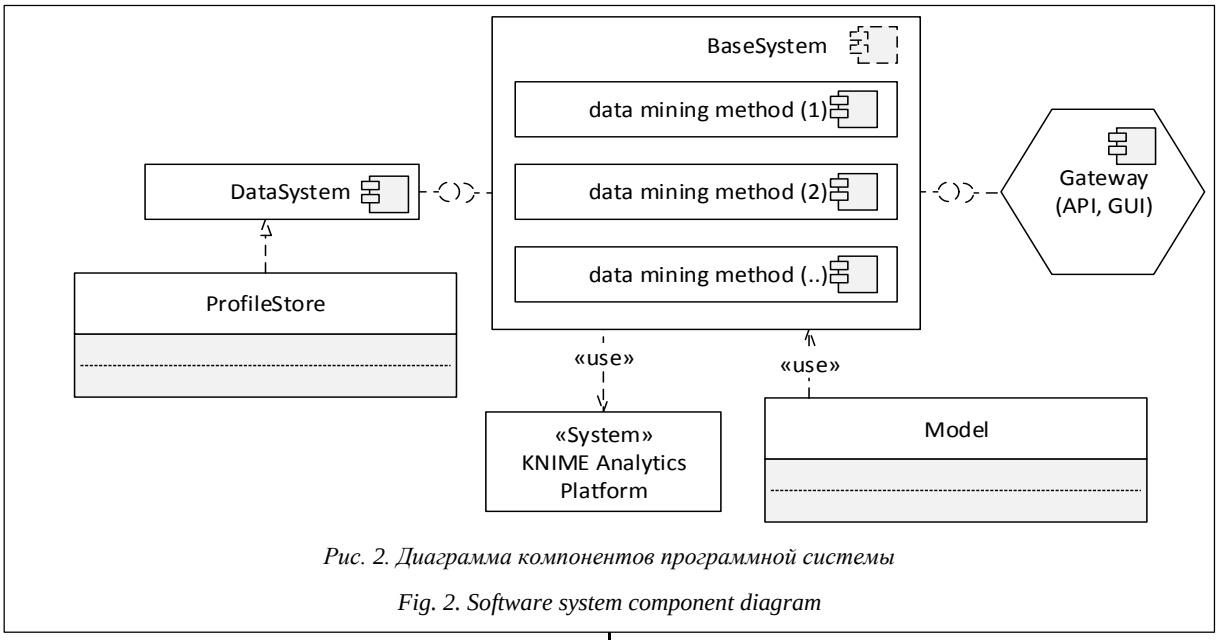

Рис. 2. Диаграмма компонентов программной системы

Fig. 2. Software system component diagram

профилей (ProfileStore) и аналитических моделей (Model). Пользователь может создавать, удалять и изменять имеющиеся в системе наборы данных и моделей. Набор создается путем отбора данных и моделей из загруженных и подготовленных руководителем (модератором) ранее в зависимости от области исследований и других критериев. На основе имеющихся наборов данных, определенных атрибутами наукометрического профиля автора в системе Scopus, программная система строит публикационный рейтинг.

Аналитическая модель (Model) представляет собой созданный и настроенный модератором интеллектуальный объект, который используется для оценки и прогнозирования принадлежности автора к той или иной публикационной группе и имплементации других методов интеллектуального анализа данных.

Базовая вычислительная подсистема (компонент BaseSystem) предусматривает модульную организацию таким образом, что каждый модуль независимо реализует один из вариантов использования. Структура основных модулей показана на рисунке 3.

Как видно из рисунка, ядром функционального модуля является вычислительный поток (KNIME

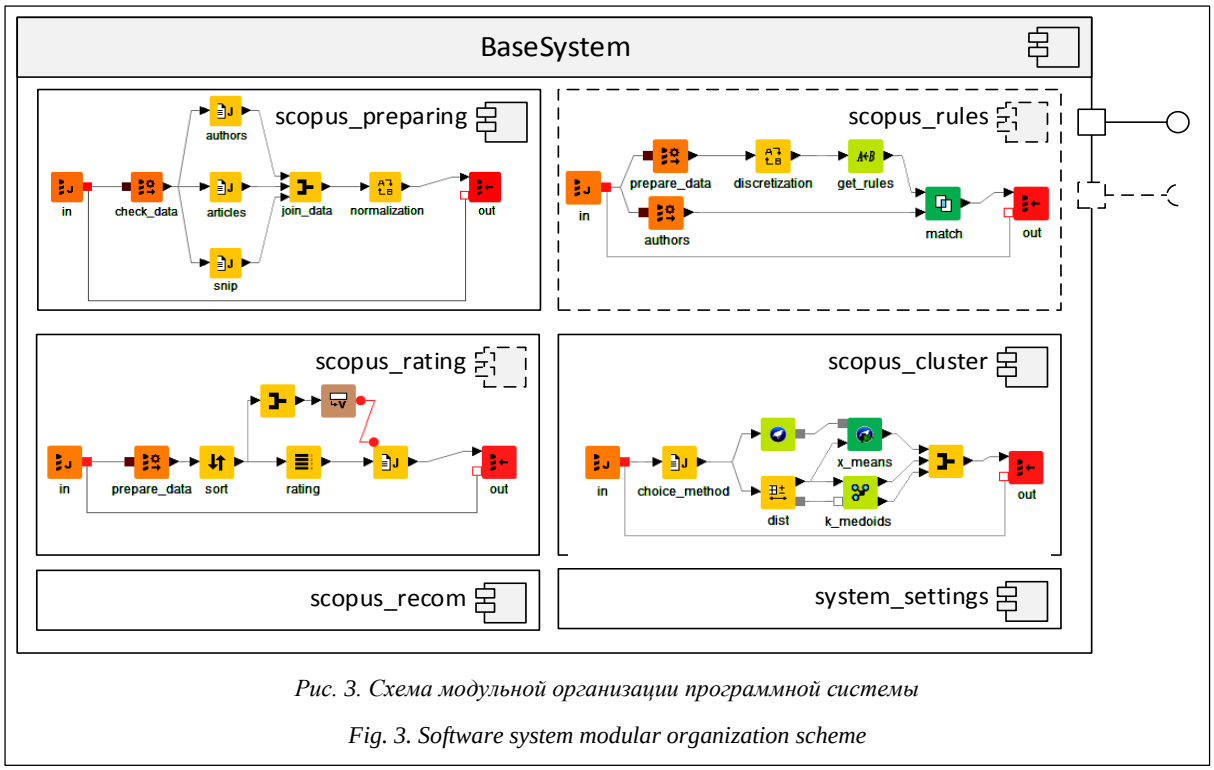

Рис. 3. Схема модульной организации программной системы

Fig. 3. Software system modular organization scheme

workflow), отдельный экземпляр которого обрабатывает данные на платформе KNIME. Из наименования модулей ясно, что *scopus_rating* реализует формирование публикационного рейтинга на основе данных системы Scopus, *scopus_rules* – оценку публикационного потенциала, *scopus_cluster* – поиск групп публикационной активности, а *scopus_preparing* – подготовку и предобработку наукометрических данных. Модульная структура, где каждый модуль автономен и реализует один из методов интеллектуального анализа, позволяет масштабировать систему в условиях распределенных вычислений и применения микросервисов. При этом компонент «шлюз» (gateway) позволяет каждому актору независимо получать результаты интеллектуального анализа и по-своему интерпретировать их для разработки рекомендаций. Полагаем, что такая архитектура позволит эффективно реализовать широкий инструментарий методов интеллектуального анализа данных, что выгодно отличает предложенную архитектуру от известных решений, например [8].

Выскажем некоторые соображения по организации пользовательского интерфейса (схема организации рабочего окна представлена на рисунке 4).

Исходя из привычного пользовательского дизайна распространенных операционных систем линейки Windows, рабочее окно должно содержать главное меню (1), предоставляющее доступ к основным параметрам программной системы, справке и помощи, а также область статуса (5), в которой будут отображены лог основных событий и прочая служебная информация.

Основная рабочая область (3) окна предоставляет пользователю различные способы отображения результатов наукометрического анализа, в частности, таблицы, текстовые сведения, графическую информацию. Информация, представленная в рабочей области, должна предоставляться в зависимости от выбранного варианта использования и примененных инструментов, при этом кнопки инструментария размещаются на инструментальной панели (4).

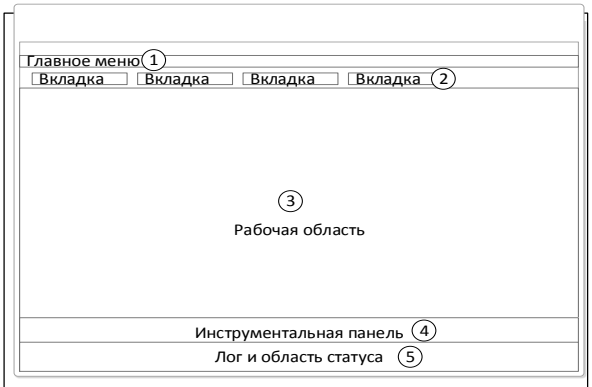

Рис. 4. Схема рабочего окна пользовательского интерфейса

Fig. 4. GUI scheme

Выбор доступных вариантов использования системы осуществляется путем переключения вкладок (2), каждая из которых отвечает за свой вариант использования и предоставляет свой инструментарий на панели (4).

Разумеется, данная схема интерфейса – только достаточный вариант представления функциональных возможностей системы. В силу модульной организации системы и в случае ее реализации в виде, например, микросервисов пользовательский интерфейс может быть выполнен на основе независимого web-представления.

Некоторые результаты использования программной системы

Прототип описанной системы был реализован в виде десктопного приложения со встроенным ADO-хранилищем данных, развернутого на операционной системе Windows 10 для непосредственного взаимодействия с развернутой здесь же системой KNIME Analytics Platform. В рамках апробации программной системы в части реализованных методик построения публикационного рейтинга и оценки публикационного потенциала рассмотрим результаты работы на примере Южно-Уральского государственного университета (см. табл. 1). Для этой цели в хранилище программной системы импортированы данные наукометрических профилей авторов, отнесенных Scopus к области исследования Computer Science и аффилированных с университетом на 1 июля 2017 г.

Таблица 1

Общая характеристика объекта исследования

Table 1

General characteristic of the research object

Международное наименование университета	Общий QS-рейтинг за 2016	QS-research in 2016	Число авторов в области Computer Science
South Ural State University (SUSU)	151–200	Medium	228

Подготовительные этапы интеллектуального анализа в соответствии с применяемой методикой формируют наборы ассоциативных правил, имеющих вид, представленный на рисунке 5.

Для поиска правил анализируются наборы с не менее чем 5 элементами, что, с одной стороны, гарантирует гибкость правил (поскольку соответствие 7 из 7 элементов обеспечивает сам публикационный рейтинг), а с другой – отсекает огромное множество заведомо недостоверных правил с малым числом элементов.

В целом программная система на имеющихся данных строит 4 163 правила при минимальной поддержке и достоверности в 5,0 %. Выборка с надежностью не менее 50,0 % обнаруживает 78 правил. Однако для перехода авторов в публикационные квартили рейтингов Q1 и Q2, а именно

Поддержка набора (support), %	**Достоверность правила (confidence), %**	**Значимость правила (lift), %**	**Результирующая группа (consequent)**		**Исходный набор уровней показателей (antecedent)**		
5,12	93,33	392,47	rank_Q1	←	h_index _Q3	snip_mean _Q1	…
5,61	88,50	371,99	rank_Q1	←	cit_count_Q1	snip_mean _Q1	…
…	…	…	…		…	…	…

Рис. 5. Схема набора полученных ассоциативных правил

Fig. 5. A scheme of a set of obtained associative rules

они в большей степени интересуют с точки зрения анализа публикационного потенциала, выделено 20 лучших правил, отвечающих заданным критериям достоверности.

В соответствии с выделенными правилами 28 авторов могут быть потенциально (с разной степенью достоверности) переквалифицированы в лучший квартиль. Из этических соображений ID авторов в системе Scopus не публикуются, поэтому в таблице 2 представлены групповые оценки.

Таблица 2

Публикационный потенциал авторов в области Computer Science

Table 2

The publication potential of authors in Computer Science

Исходный квартиль	**Число авторов**	**Средняя поддержка правил по квартилю, %**	**Средняя достоверность правил по квартилю, %**	**Целевой квартиль (consequent)**
Q2	13	6,0	90,2	Q1
Q3	48	9,0	65,8	Q2
Q4	4	7,6	60,2	Q2

Детализация оценок публикационного потенциала в разрезе отдельных авторов может стать основой для принятия управленческих решений в сфере менеджмента научных исследований, в частности, о целевом стимулировании авторов в целях прогнозируемого улучшения наукометрических показателей. На рисунках (см. http://www.swsys.ru/uploaded/image/2018_2/2018-2-dop/3.jpg, http://www.swsys.ru/uploaded/image/2018_2/2018-2-dop/4.jpg) показаны примеры индивидуальных рекомендаций, формируемых программной системой.

Заключение

В ходе исследования в целом показано, что в настоящее время целесообразна разработка программной системы, призванной помочь руководителям исследовательских групп, а также индивидуальным исследователям принимать решения по поводу улучшения публикационной активности, наукометрической результативности, определять слабые и сильные стороны, направления и конкретные решения в виде автоматизированных рекомендаций по повышению величины и репрезентативности наукометрических показателей.

В ходе проектирования сформулированы общие и технические требования к программной системе; с использованием среды моделирования Microsoft Visio построены необходимые диаграммы в нотации UML 2.0, описывающие систему.

Разработан методический инструментарий анализа публикационной активности с применением интеллектуального анализа наукометрических данных.

Разработаны модульная архитектура, алгоритмы и реализована программная система, позволяющая на основе интеллектуального анализа наукометрических данных формировать публикационный рейтинг и индивидуальные рекомендации по улучшению публикационной активности автора.

Проведены эксперименты по оценке публикационного потенциала исследователей, аффилированных с Южно-Уральским государственным университетом.

Работа выполнена при финансовой поддержке Правительства РФ (Постановление № 211 от 16.03.2013 г.), соглашение № 02.A03.21.0011.

Литература

1. Питерс Д., Марш Р. Rate my research dot com: измеряем то, что ценим, ценим, что измеряем // Научная периодика: проблемы и решения. 2011. № 1. С. 40–45.

2. Koenigstein N., Dror G., Yehuda Koren Y. Yahoo! music recommendations: modeling music ratings with temporal dynamics and item taxonomy. Proc. 5th ACM Conf. on Recommender systems. ACM, 2011, pp. 165–172.

3. Butler D. Computing giants launch free science metrics. Nature, 2011, vol. 476, p. 18.

4. He Y., Hui S.C. Mining a Web Citation Database for author co-citation analysis. Information Processing and Management, 2002, vol. 38, pp. 491–508.

5. Padrós-Cuxart R., Riera-Quintero C., March-Mir F. Bibliometrics: a Publication Analysis Tool. Proc. 3rd Workshop on Bibliometric-enhanced Information Retrieval (BIR 2016), 2016, pp. 44–53.

6. Harzing AW., Alakangas S. Microsoft Academic: is the Phoenix getting wings? 2016, vol. 106. DOI: 10.1007/s11192-015-1798-9.

7. Луценко Е.В., Орлов А.И., Глухов В.А. Наукометрическая интеллектуальная измерительная система по данным РИНЦ на основе АСК-анализа и системы «Эйдос» // Науч. журн. КубГАУ. 2016. № 122. С. 1–56.

8. Синицын А.А., Никифоров О.Ю., Андреев М.А. Концепция и структура информационно-аналитической системы анализа публикационной активности сотрудников научно-образовательной организации // Фундаментальные исследования. 2014. № 11. С. 1276–1280.

9. Синицын А.А., Никифоров О.Ю., Андреев М.А. Особенности применения информационно-аналитической системы для оценки направления поддержки по созданию результатов интеллектуальной деятельности научно-образовательной организации // Фундаментальные исследования. 2014. № 11-6. С. 1271–1275.

10. Мбого И.А., Прокудин Д.Е. Подходы к развитию инструментов автоматизации и интеграции ресурсов информационного пространства поддержки междисциплинарного научного направления // Интернет и современное общество IMS-2015: сб. науч. стат. XVIII объединен. конф. СПб, 2015. С. 290–302.

11. Мбого И.А., Прокудин Д.Е., Чугунов А.В. Разработка инструментов интеграции научной информации в пространстве разнородных информационных систем // Научный сервис в сети Интернет: тр. XVIII Всерос. науч. конф. М.: Изд-во ИПМ им. М.В. Келдыша, 2016. С. 249–258. DOI: 10.20948/abrau-2016-44.

12. Agrawal R., Srikant R. Fast algorithms for mining association rules. Proc. 20th Int. Conf. Very Large Databases, Santiago, Chile, 1994, pp. 487–499.

13. Borgelt C. Efficient Implementations of Apriori and Eclat. Proc. Workshop of Frequent Item Set Mining Implementations (FIMI 2003, Melbourne, FL, USA). URL: http://www.borgelt.net/papers/fimi_03.pdf (дата обращения: 02.10.2017).

14. Валько Д.В., Колташев А.С. Модуль агрегации наукометрических данных открытых сервисов в сети Интернет: пат. 2016619028. Рос. Федерация. № 2016616489; заявл. 21.06.16; опубл. 11.08.16, Бюл. № 9. 1 с.

Software & Systems
DOI: 10.15827/0236-235X.122.275-283

Received 12.10.17
2018, vol. 31, no. 2, pp. 275–283

A RECOMMENDATION SYSTEM BASED ON DATA MINING OF A SCIENTOMETRIC RESEARCH PROFILE

D.V. Valko [1], *Graduate Student, ell.science@mail.ru*

[1] *South Ural State University (National Research University), Lenin Ave. 76, Chelyabinsk, 454080, Russian Federation*

Abstract. Nowadays scientific results can be represented in various scientometric bases and systems. They are often popular not because of their relevance, but due to global availability. In fact, scientific results may be out of the scope of a scientific community simply because they are not placed in a popular scientometric system. From scientific point of view, such situation devalues a researcher regardless of quality and relevance of his scientific results.

According to the author, the development of recommendations for individual researchers, research teams and their managers at all levels of management would make it possible to pay attention to promising scientific results and reasonably accumulate necessary resources to include such results in popular scientometric systems.

Development of tools that operate the big scientometric open data cannot be without data mining methods. The paper shows that based on the algorithm of intellectual analysis of interrelations (like apriori), which is adapted to scientometric data in the Scopus, it is possible to formulate certain sets of associative rules suitable for forecasting probable future scientific results. It is also possible to develop automated recommendations for improving publication activity. The paper proposes the developed methodical tools for analyzing publication and scientometric data using data mining methods. In addition, it describes a modular architecture and a prototype of a software system that allows forming a publication rating and individual recommendations for improving author's publication activity based on scientometric data mining.

The paper shows some experimental results on assessing publication potential of researchers affiliated with the South Ural State University.

The authors built necessary diagrams in the UML 2.0 notation describing that software system using Microsoft Visio modeling environment.

Keywords: scientometrics, recommendation system, data mining, publication activity.

Acknowledgements. *The work has been financially supported by the Russian government (Resolution no. 211 dated March 16, 2013), Agreement no. 02.A03.21.0011.*

References

1. Piters D., Marsh R. Rate my research dot com: we measure what we value, we value what we measure. *Nauchnaya periodika: problemy i resheniya* [Scholarly Communication Review]. 2011, no. 1, pp. 40–45 (in Russ.).

2. Koenigstein N., Dror G., Yehuda Koren Y. Yahoo! music recommendations: modeling music ratings with temporal dynamics and item taxonomy. *Proc. 5th ACM Conf. on Recommender Systems*. ACM Publ., 2011, pp. 165–172.

3. Butler D. Computing giants launch free science metrics. *Nature*. 2011, vol. 476, p. 18.

4. He Y., Hui S.C. Mining a Web Citation Database for author co-citation analysis. *Information Processing and Management*. 2002, vol. 38, pp. 491–508.

5. Padrós-Cuxart R., Riera-Quintero C., March-Mir F. Bibliometrics: a Publication Analysis Tool. *Proc. 3rd Workshop on Bibliometric-enhanced Information Retrieval (BIR 2016)*. 2016, pp. 44–53.

6. Harzing AW., Alakangas S. Microsoft Academic: is the Phoenix getting wings? *Harzing.com*. 2016, vol. 106, DOI: 10.1007/s11192-015-1798-9.

7. Lutsenko E.V., Orlov A.I., Glukhov V.A. A scientometric intelligent measuring system of RSCI data based upon the ask analysis and Eidos system. *Nauchny zhurnal KubGAU* [Scientific J. of KubSAU]. 2016, no. 122 (08), pp. 1–56 (in Russ.).

8. Sinitsyn A.A., Nikiforov O.Yu., Andreev M.A. Concept and structure of the information-analytical system for to analyze scientific activity efficiency of the research and educational institution. *Fundamentalnye issledovaniya* [Fundamental research]. 2014, no. 11, pp. 1276–1280 (in Russ.).

9. Sinitsyn A.A., Nikiforov O.Yu., Andreev M.A. Features of the information-analytical system application for estimation the support areas for creation of the results of the intellectual activity of the research and educational institutions. *Fundamentalnye issledovaniya* [Fundamental research]. 2014, no. 11-6, pp. 1271–1275 (in Russ.).

10. Mbogo I.A., Prokudin D.E. Approaches to development of tools for automation and integration of information space resources to support interdisciplinary research. *Sb. nauch. st. XVIII Obyedinen. konf. "Internet i sovremennoe obshchestvo" IMS-2015* [Proc. 18th Joint Conf. "Internet and Contemporary Society" IMS-2015]. St. Petersburg, 2015, pp. 290–302 (in Russ.).

11. Mbogo I.A., Prokudin D.E., Chugunov A.V. Development of tools for integration of scientific information among heterogeneous information systems. *Nauchny servis v seti Internet: tr. XVIII Vseros. nauch. konf.* [Scientific Service on the Internet: Proc. 18th All-Russ. Sci. Conf.]. Moscow, IPM im. M.V. Keldysha Publ., 2016, pp. 249–258 (in Russ.).

12. Agrawal R., Srikant R. Fast algorithms for mining association rules. *Proc. 20th Int. Conf. on Very Large Databases*. Santiago, Chile, 1994, pp. 487–499.

13. Borgelt C. Efficient Implementations of Apriori and Eclat. *Proc. Workshop of Frequent Item Set Mining Implementations (FIMI 2003)*. Available at: http://www.borgelt.net/papers/fimi_03.pdf (accessed October 2, 2017).

14. Valko D.V., Koltashev A.S. *Modul agregatsii naukometricheskikh dannykh otkrytykh servisov v seti Internet* [Module for Aggregating Scientific Data of Open Services on the Internet]. Patent RF 2016619028, no. 2016616489.