УПРАВЛЕНИЕ В СОЦИАЛЬНЫХ И ЭКОНОМИЧЕСКИХ СИСТЕМАХ

УДК 004.91

ПРОБЛЕМА ДОСТОВЕРНОСТИ И ОБЪЕКТИВНОСТИ ИНФОРМАЦИИ ВНУТРИ НАУЧНОГО ИНТЕРНЕТ-СООБЩЕСТВА, ПОСТРОЕННОГО НА ПРИНЦИПАХ WEB 2.0

Статья поступила в редакцию 09.10.2013, в окончательном варианте 13.10.2013.

Зелепухина Виктория Андреевна, кандидат технических наук, доцент, Астраханский государственный университет, 414056, Российская Федерация, г. Астрахань, ул. Татищева, 20а, e-mail: viktoria_82@mail.ru

Структура научного сообщества может быть представлена в виде связного графа, узлами которого являются объекты и субъекты научной деятельности. Под субъектами понимаются как отдельные ученые, так и их коллективы. Под объектами – публикации, исследовательские проекты, гранты, патенты и др. Между узлами графа устанавливаются как прямые, так и транзитивные связи за счет наличия у субъектов совместных результатов научной деятельности, работы в одних и тех же организациях и др. Уровень достоверности информации, хранимой в базе данных научного сообщества, построенного согласно принципам Web 2.0, напрямую влияет на результаты наукометрического анализа, оценку ученого и всего коллектива, а также на эффективность информационного поиска в сети Интернет. Без привлечения специальных интеллектуальных алгоритмов, позволяющих идентифицировать связи между узлами сообщества, наукометрический анализ не будет объективен, а структура виртуального научного сообщества не будет соответствовать реальным связям между учеными. В статье выполнен аналитический обзор существующих в настоящее времени подходов и алгоритмов, позволяющих повысить достоверность и надежность данных, представленных в научном интернет-сообществе[1].

Ключевые слова: научное сообщество, научная сеть, электронные библиотеки, Web 2.0, достоверность информации, объективность информации, устранение неоднозначности

THE PROBLEMS OF THE ACCURACY AND OBJECTIVITY OF THE INFORMATION STORED IN THE SCIENTIFIC WEB 2.0 COMMUNITY

Zelepukhina Victoria A., Ph.D. (Engineering), Associated Professor, Astrakhan State University, 20a Tatishchev St., Astrakhan, 414056, Russian Federation, e-mail: viktoria_82@mail.ru

The structure of scientific community is represented by connected graph. Vertices of the graph are objects and subjects of scientific activity (actors). The objects are publications, grants, patents, etc. The actors are individual scientists and scientific groups. The graph vertices are connected, if the scientists have common objects of scientific activity. The relations between vertices can be direct or transitive and established between the vertices of different types. The accuracy of the information stored in the database of the scientific community Web 2.0, affect the results of scientometric analysis and the effectiveness of information retrieval. Virtual scientific community Web 2.0 will reflect the actual structure of the research groups, if it

[1] Работа выполнена при финансовой поддержке РФФИ в рамках научного проекта № 12-07-31145 мол_а («Разработка логико-концептуальной модели информационной системы для обмена информацией внутри научного интернет-сообщества»).

applies special intelligent algorithms that connect vertices of the graph. The paper describes the main approaches and algorithms for solving such problems.

Keywords: scientific network, scientific community, Web 2.0, the accuracy of the information, the objectivity of the information, disambiguation, digital libraries

Современные научные интернет-сети и научные сообщества (НС) разрабатываются и эксплуатируются на основе концепции Web 2.0, согласно которой формирование информации об объектах научной деятельности (ОНД) производится силами самих пользователей. Структуру НС представляют в виде графа, в котором узлы – субъекты научной деятельности (СНД), т.е. ученые и коллективы ученых, и ОНД – связаны друг с другом либо прямым, либо транзитивным способами (за счет наличия у субъектов совместных публикаций, работы в одной организации, совместного участия в научных мероприятиях и др.) [2, 10, 14, 21]. Характерный шум и избыточность информации в НС Web 2.0 могут приводить к неверной идентификации связей между узлами. Таким образом, необходим контроль со стороны системы над поступающими от пользователей данными, а также применение интеллектуальных алгоритмов для повышения степени упорядоченности графа НС и синтеза виртуального сообщества на основе структуры НС и информации о СНД и ОНД. Так, по некоторым оценкам, лишь 20 % контента систем Web 2.0 предоставляется напрямую пользователями, а 80 % – результат интеллектуального анализа этого материала [27], осуществляемого с применением специальных алгоритмов. Настоящая статья представляет собой аналитический обзор алгоритмов, применяемых в НС на основе Web 2.0 для изучения информации об СНД и ОНД с целью установления связей между узлами НС.

Полнота, достоверность, объективность и актуальность информации, представленной в НС, а также системах для создания персональных страниц ученых и системах текущих исследований (Current Research Information Systems, CRIS), имеют большое значение для установления закономерностей динамики научной активности. К числу таких систем относятся ResearchGate [25], Academia.edu [12], ORCID [23], ResearcherID [24], Соционет [8], ИСТИНА [4] и др. Мировой опыт по способам оценивания эффективности науки представлен в виде различных руководств и правил (например, руководство Фраскати [16]) и использует различные показатели, такие как индекс Хирша, индекс цитирования, импакт-фактор и др. Может возникнуть ситуация, при которой информация об одном и том же ОНД представлена в базе данных (БД) системы несколько раз. В итоге в статистическом отчете о научной активности коллектива будет фигурировать, например, несколько статей в высокорейтинговом журнале вместо одной. Ситуация осложняется в случае автоматической интеграции данных об ОНД в профили вышестоящих сообществ и организаций. Отсутствие связи между узлами НС вызовет, наоборот, снижение рейтинга ученого и всего коллектива.

Описанные проблемы часто возникают в Scopus и РИНЦ: неверное объединение нескольких ученых в один профиль и, наоборот, дезинтеграция информации об авторе в несколько учетных записей. Проблемы усугубляются в случае отечественных ученых, публикующихся в переводных и международных изданиях: одна и та же фамилия может быть представлена различными способами в зависимости от выбранных правила транслитерации.

Для решения описанных проблем необходима разработка алгоритмов, позволяющих идентифицировать связи между узлами НС. Каждая связь характеризуется некоторым предикатом (отношением), что позволяет говорить о представлении знаний в НС в виде семантической сети.

Одним из видов отношений между узлами НС является тип связи, характеризующийся в соответствии с терминологией CERIF [15] как “same as”. Для идентификации такого отношения разработаны алгоритмы, позволяющие выявлять дубликаты среди записей об

ОНД [1, 3, 6, 9]. Эти алгоритмы учитывают наличие шума в записях, что позволяет производить нечеткое сопоставление атрибутов ОНД и вычислять «расстояние» между ними.

Для установления отношений между СНД и ОНД (например, “author of”, “editor of” и др.) сравнивают, в первую очередь, фамилии и имена авторов, указанных при описании ОНД, с фамилиями и именами СНД, зарегистрированных в НС. Таким образом, выполняется первичная кластеризация СНД по отношению к ОНД. Результат такой кластеризации не является достаточным условием при установлении связи между СНД и ОНД: синтез связей между узлами сообщества осложняется тем, что одному ученому могут соответствовать несколько различных фамилий, и тем, что одна фамилия может указывать на различных ученых. Кроме того, нам известны случаи, когда два ученых с одинаковыми фамилиями и инициалами занимаются родственными научными исследованиями. За рубежом подобная проблема именуется как «Author name disambiguation» (устранение неоднозначности имени автора) [29, 31] и весьма актуальна для области построения электронных библиотек (ЭБ) [18, 30, 32, 33]. В рамках поддержки деятельности отечественных ЭБ разработаны математические модели и алгоритмы, позволяющие идентифицировать связи между записями из авторитетных источников и библиографическими записями [5, 11]. Обычно в таких алгоритмах не уделяется особого внимания проблемам «грязных» данных в описании публикации. Это объясняется тем, что библиографические БД в ЭБ наполняются и поддерживаются профессиональными библиографами – это снижает вероятность появления ошибок.

Как правило, алгоритмы, идентифицирующие авторство научной работы, основаны на анализе текста публикации. При этом выделение ключевых терминов в полнотекстовой версии позволяет строить кластеры второго уровня, где остаются только те СНД, для которых было получено наибольшее совпадение по ключевым словам. Существующие алгоритмы можно условно разделить на следующие группы.

1. Алгоритмы, использующие совокупность характеристик публикации. В набор характеристик входят заголовок публикации, ключевые слова, аннотация, текст работы, список источников, ФИО авторов, их e-mail адреса, названия организаций, название журнала [17, 26]. Персональные «идентификационные коды ученых», присваиваемые им в различных системах, пока в текстах публикаций указывать не принято.

2. Публикации разбиваются на кластеры, каждый из которых соответствует одному автору. Математическая модель таких алгоритмов представляет собой, как правило, логистическое уравнение, для определения коэффициентов которого используются методы машинного обучения для задач классификации и регрессионного анализа. Например, метод опорных векторов [13], Random forest [33] и др.

3. Алгоритмы, производящие сравнение ключевых характеристик публикации с информацией из сети Интернет [20, 22, 28, 34]. В таких алгоритмах происходит генерация ряда запросов, текст которых основан на информации о публикации, и анализ «поисковой выдачи». Проведенное нами тестирование алгоритмов для отечественных авторов и публикаций показало большое количество ложных результатов в таких выдачах.

4. Алгоритмы, производящие нечеткое сопоставление фамилий, имен и/или инициалов авторов с последующим принятием решения об авторстве со стороны пользователя (администратора) ИС. Подходы, реализующие такие алгоритмы, получили название «наивных», так как они дают большое количество ложных результатов за счет наличия в БД систем нескольких ученых с одинаковыми именами. Для нечеткого сопоставления «фамилий-имен», как правило, используются методы, основанные на вычислении редакционного расстояния (например, Левенштейна, Дамерау – Левенштейна), сравнении n-грамм каждого имени. Использование n-грамм является более эффективным с точки зрения вычислительной эффективности – за счет того, что есть возможность заранее подготовить такие подстроки и сохранить их в БД системы.

Из недостатков существующих алгоритмов можно выделить отсутствие адаптации к российским национальным особенностям и требование наличия полнотекстовой версии публикации. Кроме того, НС на основе Web 2.0 присущи особенности, которые требуют адаптации и/или изменения алгоритмов, существующих в области использования ЭБ:

- ученые разных специальностей зачастую используют различную терминологию при описании одного и того же явления. Таким образом, сравнение ключевых слов может привести к тому, что участник НС не будет идентифицирован как потенциальный соавтор публикации;
- НС на основе Web 2.0, как правило, позволяют производить сбор и хранение информации об ОНД различных типов. Например, свидетельства об интеллектуальной собственности, сведения о научных наградах и премиях, информация о полученных грантах. Все такие ОНД являются частью структуры сообщества и требуют идентификации связей с другими узлами НС;
- алгоритмы устранения неоднозначности имени автора в ЭБ анализируют информацию, представленную в тексте самой публикации (ключевые слова, аннотация, список литературы, сведения об организации, email). Однако в случае НС на основе Web 2.0 нельзя гарантировать наличие загруженной пользователем полнотекстовой версии;
- в НС на основе Web 2.0, как правило, предусмотрено большое количество типов связей между узлами, что может компенсировать отсутствие полнотекстовых версий ОНД. Например, двое ученых уже имеют транзитивную связь за счет наличия совместного ОНД. Такая связь повышает вероятность соавторства данных участников при анализе другого ОНД, если в соответствующих его характеристиках указаны те же фамилии;
- НС на основе Web 2.0 характеризуются наличием «эталонного» набора авторов (пользователей), для которых выполняется идентификация связей с другими узлами сообщества. В случае с ЭБ такой набор есть, если сформированы соответствующие «авторитетные источники».

Таким образом, с учетом указанных особенностей НС целесообразно проводить адаптацию разработанных для области ЭБ алгоритмов. В частности целесообразно следующее.

a. Использование информации об уже синтезированной структуре виртуального НС. Так, например, если двое ученых связаны посредством информации об обучении вида «Соискатель» – «Подготовка диссертации» – «Руководитель», то вероятность того, что именно эти ученые являются соавторами некоторой исследуемой публикации выше, чем для других ученых, имеющих те же фамилии.

b. Учет отечественных национальных особенностей, связанных с использованием различных способов транслитерации русских фамилий и имен авторами ОНД [6, 18], а также лицами, выполняющими ссылки на них в своих научных статьях, монографиях и пр.

c. Адаптацию алгоритмов для установления связей с узлами НС, соответствующих ОНД других типов (гранты, патенты, научные награды и др.), а также сведениями о научных мероприятиях, защитах на соискание ученых степеней, научном руководстве и др.

Решение проблемы достоверности и объективности информации в НС на основе Web 2.0 сводится не только к внедрению интеллектуальных алгоритмов анализа информации, заложенной в характеристиках ОНД и СНД. Так, сопряжение НС с внешними сервисами научной и наукометрической информации (например, с CrossRef, Scopus, eLIBRARY.RU, издательство Springer и др.), позволило бы ускорить процесс формирования информации об ОНД (в частности, о публикациях) и снизить вероятность появления неполных и недостоверных сведений.

Итак, выводы.

1. Проблема обеспечения достоверности и объективности информации внутри научных Интернет-сообществ, основанных на использовании технологии Web 2.0, в настоящее время является весьма актуальной.

2. Эта проблема может решаться за счет повышения интеллектуальности алгоритмов анализа информации по ОНД в отношении их авторства.

3. Неоднозначность используемых вариантов транслитераций фамилий и имен русскоязычных авторов на латиницу (как самими авторами в рамках библиографических описаний, так и лицами, дающими ссылки на ОНД в своих публикациях) приводит к дополнительным сложностям при определении авторства, делает целесообразным привлечение дополнительной информации из текстов статей и пр.

Список литературы

1. Барахнин В. Б. О задании меры сходства для кластеризации текстовых документов / В. Б. Барахнин, В. А. Нехаева, А. М. Федотов // Вестник Новосибирского государственного университета. Сер. Информационные технологии. – 2008.– № 1 (6). – С. 3–9.

2. Брумштейн Ю. М. Использование интернет-технологий в управлении научным имиджем регионального вуза / Ю. М. Брумштейн, А. А. Бондарев, И. А. Дюдиков // Прикаспийский журнал: управление и высокие технологии. – 2013. – № 2 (22). – С. 90–99.

3. Дербенев Н. В. Выявление нечетких дубликатов в наукометрическом анализе / Н. В. Дербенев, В. О. Толчеев // Информационные технологии. – 2011. – № 12. – С. 24–29.

4. Интеллектуальная Система Тематического Исследования Научно-технической информации. Режим доступа: http://istina.msu.ru/ (дата обращения 04.10.2013), свободный. – Загл. с экрана. – Яз. рус.

5. Князева А. А. Автоматическое связывание документов / А. А. Князева, И. Ю. Турчановский, О. С. Колобов // Электронные библиотеки: перспективные методы и технологии, электронные коллекции : тр. XIV Всерос. науч. конф. RCDL'2012. – Переславль-Залесский : Университет города Переславля, 2012. – С. 360–369.

6. Рубцов Д. Н. Выявление дубликатов в разнородных библиографических источниках / Д. Н. Рубцов, В. Б. Барахнин // Вестник Новосибирского государственного университета. Сер. Информационные технологии. – 2009. – № 3 (7). – С. 86–93.

7. Система стандартов по информации, библиотечному и издательскому делу. Правила транслитерации кирилловского письма латинским алфавитом. – Режим доступа: http://protect.gost.ru/document.aspx?control=7&id=130715 (дата обращения 14.08.2013), свободный. – Загл. с экрана. – Яз. рус.

8. Соционет. – Режим доступа: http://socionet.ru/ (дата обращения 04.10.2013), свободный. – Загл. с экрана. – Яз. рус.

9. Толчеев В. О. Анализ проблемы и разработка процедуры выявления нечетких дубликатов научных статей по библиографическим описаниям / В. О. Толчеев // Информационные технологии. – 2011. – № 2. – С. 17–21.

10. Умаров А. С. Некоторые аспекты создания информационных систем для сбора и хранения научной и наукометрической информации / А. С. Умаров, Н. В. Попова, В. А. Зелепухина // Прикаспийский журнал: управление и высокие технологии. – 2013. – № 3 (23). – С. 111–118.

11. Федотов А. М. Проблемы авторитетного контроля для распределенных электронных библиотек и библиографических баз данных / А. М. Федотов, О. Л. Жижимов, А. А. Князева и др. // Вестник Новосибирского государственного университета. Сер. Информационные технологии. – 2011. – № 1 (9). – С. 89–101.

12. Academia.edu. – Available at: http://www.academia.edu/ (accessed 04.10.2013).

13. Dendek P. J. Evaluation of Features for Author Name Disambiguation Using Linear Support Vector Machines / P. J. Dendek, L. Bolikowski, M. Lukasik // Document Analysis Systems (DAS) : 10th IAPR International Workshop. – 2012. – P. 440–444.

14. Ebel H. Dynamics of social networks / H. Ebel, J. Davidsen, S. Bornholdt // Complexity – Complex Adaptive systems. – 2002. – Vol. 2 (8). – P. 24–27.

15. Features of CERIF. Available at: http://www.eurocris.org/Index.php? page=featuresCERIF&t=1 (accessed 04.10.2013).

16. Frascati Manual. Available at: http://www.tubitak.gov.tr/tubitak_content_files/ BTYPD/kilavuzlar/Frascati.pdf (accessed 04.10.2013).

17. Gurney T. Author Disambiguation Using Multi-Aspect Similarity Indicators / T. Gurney, E. Horlings, P. Van den Besselaar // Scientometrics. – 2012. – Vol. 2 (91). – P. 435–449.

18. Huynh T. Vietnamese Author Name Disambiguation for Integrating Publications from Heterogeneous Sources / T. Huynh, K. Hoang, T. Do, D. Huynh // Intelligent Information and Database Systems: 5th Asian Conference, ACIIDS 2013, Kuala Lumpur, Malaysia, March 18–20, 2013. – Springer Berlin Heidelberg, 2013. – Part I. – P. 226–235.

19. ISO 9:1995 – Information and documentation – Transliteration of Cyrillic characters into Latin characters – Slavic and non-Slavic languages. Available at: http://www.iso.org/iso/ home/store/catalogue_tc/catalogue_detail.htm?csnumber=3589 (accessed 14.08.2013).

20. Lu Y. Name Disambiguation Using Web Connection / Y. Lu, Z. Nie, T. Cheng, Y. Gao, J. R. Wen // Workshop on Information Integration on the Web : Proceedings of AAAI 2007. – 2007. – P. 56–61.

21. Newman M. E. J. Scientific collaboration networks. I. Network construction and fundamental results / M. E. J. Newman // Phys. Rev. E. – 2001. – № 1 (64). – P. 64–71.

22. Nicolas M. Focused Crawling Using Name Disambiguation on Search Engine Results / M. Nicolas, K. Khelif // Intelligence and Security Informatics Conference (EISIC), 2011 European.– Institute of Electrical and Electronics Engineers, 2011. – P. 340–345.

23. ORCID. – Available at: http://orcid.org/ (accessed 04.10.2013).

24. ResearcherID. – Available at: http://www.researcherid.com/ (accessed 04.10.2013).

25. ResearchGate. – Available at: http://researchgate.net (accessed 04.10.2013).

26. Rui Z. Author Name Disambiguation for Citations on the Deep Web / Z. Rui, D. Shen, Yu. Kou, T. Nie // Web-Age Information Management: WAIM 2010 International Workshops: IWGD 2010, XMLDM 2010, WCMT 2010, Jiuzhaigou Valley, China, July 15–17, 2010. Revised Selected Papers. – Springer Berlin Heidelberg, 2010. – P. 198–209.

27. Segaran T. Programming Collective Intelligence. O'Reilly Media / T. Segaran. – 2007. – 362 p.

28. Shin D. Automatic Method for Author Name Disambiguation Using Social Networks / D. Shin, T. Kim, H. Jung, J. Choi // Advanced Information Networking and Applications (AINA) : 24th IEEE International Conference. – 2010. – P. 1263–1270.

29. Smalheiser Neil R. Author name disambiguation / Neil R. Smalheiser, Vetle I. Torvik // Annual Review of Information Science and Technology. – 2009. – Vol. 1 (43). – P. 1–43.

30. Strotmann A. Author name disambiguation for collaboration network analysis and visualization / A. Strotmann, D. Zhao, T. Bubela // Proc. Am. Soc. Info. Sci. Tech. – 2009. – Vol. 46. – P. 1–20.

31. Strotmann A. Author name disambiguation: What difference does it make in author-based citation analysis? / A. Strotmann, Z. Dangzhi // Journal of the American Society for Information Science and Technology. – 2012. – Vol. 9 (63). – P. 1820–1833.

32. Torvik Vetle I. A probabilistic similarity metric for Medline records: A model for author name disambiguation / Vetle I. Torvik, Marc Weeber, Don R. Swanson, Neil R. Smalheiser // Journal of the American Society for Information Science and Technology. – 2005. – Vol. 2 (56). – P. 140–158.

33. Treeratpituk P. Disambiguating authors in academic publications using random forests / P. Treeratpituk, C. Lee Giles // JOINT CONFERENCE IN DIGITAL LIBRARIES (JCDL '09). – New York, 2009. – P. 39–48.

34. Yang K.-H. Author Name Disambiguation for Citations Using Topic and Web Correlation / K.-H. Yang, H.-T. Peng, J.-Y. Jiang, H.-M. Lee, J.-M. Ho // Research and Advanced Technology for Digital Libraries : 12th European Conference, ECDL 2008, Aarhus, Denmark, September 14–19, 2008. – Springer Berlin Heidelberg, 2008. – P. 185–196.

35. Yang Kai-Hsiang. Author Name Disambiguation in Citations / Kai-Hsiang Yang, Yi-Hsuan Wu // WI-IAT '11 Proceedings of the 2011 IEEE/WIC/ACM : International Conferences on Web Intelligence and Intelligent Agent Technology. – 2011. – Vol. 3. – P. 335–338.

References

1. Barakhnin V. B., Nekhaeva V. A., Fedotov A. M. O zadanii mery skhodstva dlya klasterizatsii tekstovykh dokumentov [Similarity determination for textual documents clusterization]. *Vestnik Novosibirskogo gosudarstvennogo universiteta. Seriya. Informatsionnye tekhnologii* [Bulletin of Novosibirsk State University. Series. Information Technologies], 2008, no. 1 (6), pp. 3–9.

2. Brumshteyn Yu. M., Bondarev A. A., Dyudikov I. A. Ispolzovanie internet-tekhnologiy v upravlenii nauchnym imidzhem regionalnogo vuza [Internet technologies usage in managementof regional universities scientific image]. *Prikaspiyskiy zhurnal: upravlenie i vysokie tekhnologii* [Caspian Journal: Management and High Technologies], 2013, no. 2 (22), pp. 90–99.

3. Derbenev N. V., Tolcheev V. O. Vyyavlenie nechetkikh dublikatov v naukometricheskom analize [Using a Method of Detecting Near Duplicates in Sciencemetric Analysis]. *Informatsionnye tekhnologii* [Information Technologies], 2011, no. 12, pp. 24–29.

4. Intellectual System of Thematical Research of Scientific-Technical Information. Available at: http://istina.msu.ru/ (accessed 4 October 2013).

5. Knyazeva A. A., Turchanovskiy I. Yu., Kolobov O. S. Avtomaticheskoe svyazyvanie dokumentov [Automatic Documents Linkage]. *Elektronnye biblioteki: perspektivnye metody i tekhnologii, elektronnye kollektsii: trudy XIV Vserossiyskoy nauchnoy konferentsii RCDL'2012* [Digital Libraries: Advanced Methods and Technologies. Digital collections (RCDL)]. Pereslavl-Zalesskiy, 2012, pp. 36–369.

6. Rubtsov D. N., Barakhnin V. B. Vyyavlenie dublikatov v raznorodnykh bibliograficheskikh istochnikakh. [Duplicate Detection in Heterogenous Bibliographic Sources]. *Vestnik Novosibirskogo gosudarstvennogo universiteta. Seriya. Informatsionnye tekhnologii* [Bulletin of Novosibirsk State University. Series. Information Technologies], 2009, no. 3 (7), pp. 86–93.

7. The system of standards for information, librarianship and publishing. Terms of transliteration of the Cyrillic letters using the Latin alphabet. Available at: http://protect.gost.ru/document.aspx?control=7&id=130715 (accessed 4 October 2013).

8. Socionet. Aviable at: http://socionet.ru/ (accessed 4 October 2013).

9. Tolcheev V. O. Analiz problemy i razrabotka protsedury vyyavleniya nechetkikh dublikatov nauchnykh statey po bibliograficheskim opisaniyam [Analysis of problem and development of method of detection of fuzzy dublicates of scientific articles on the base of bibliographic descriptions]. *Informatsionnye tekhnologii* [Information Technologies], 2011, no. 2, pp. 17–21.

10. Umarov A. S., Popova N. V., Zelepukhina V. A. Nekotorye aspekty sozdaniya informatsionnykh sistem dlya sbora i khraneniya nauchnoy i naukometricheskoy informatsii [Some aspects of the development of information systems for the collection and storage of scientific and scientometric information]. *Prikaspiyskiy zhurnal: upravlenie i vysokie tekhnologii* [Caspian Journal: Management and High Technologies], 2013, no. 3, pp. 111–118.

11. Fedotov A. M., Zhizhimov O. L., Knyazeva A. A. et al. Problemy avtoritetnogo kontrolya dlya raspredelennykh elektronnykh bibliotek i bibliograficheskikh baz dannykh [Problems of authority control for distributed digital libraries and bibliographic databases]. *Vestnik Novosibirskogo gosudarstvennogo universiteta. Seriya. Informatsionnye tekhnologii* [Bulletin of Novosibirsk State University. Series. Information Technologies], 2011, no. 1 (9), pp. 89–101.

12. Academia.edu. Available at: http://www.academia.edu/ (accessed 4 October 2013).

13. Dendek P. J., Bolikowski L., Lukasik M. Evaluation of Features for Author Name Disambiguation Using Linear Support Vector Machines. *Document Analysis Systems (DAS): 10th IAPR International Workshop*, 2012, pp. 440–444.

14. Ebel H., Davidsen J., Bornholdt S. Dynamics of social networks. *Complexity – Complex Adaptive systems*, 2002, no. 2 (8), pp. 24–27.

15. Features of CERIF. Available at: http://www.eurocris.org/Index.php?page=featuresCERIF&t=1 (accessed 4 October 2013).

16. Frascati Manual. Available at: http://www.tubitak.gov.tr/tubitak_content_files/BTYPD/kilavuzlar/Frascati.pdf (accessed 4 October 2013).

17. Thomas Gurney, Edwin Horlings, and Peter Van Den Besselaar. Author disambiguation using multi-aspect similarity indicators. *Scientometrics*, 2012, no. 91 (2), pp. 435–449.

18. Tin Huynh, Kiem Hoang, Tien Do, Duc Huynh: Vietnamese Author Name Disambiguation for Integrating Publications from Heterogeneous Sources. *Intelligent Information and Database Systems: 5th Asian Conference, ACIIDS 2013*, Kuala Lumpur, Malaysia, March 18–20, 2013, part I, pp. 226–235.

19. ISO 9:1995 – Information and documentation – Transliteration of Cyrillic characters into Latin characters – Slavic and non-Slavic languages. Available at: http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=3589 (accessed 4 October 2013).

20. Lu Y., Nie Z., Cheng T., Gao Y., Wen J. R. Name Disambiguation Using Web Connection. *Workshop on Information Integration on the Web: Proceedings of AAAI 2007*, 2007, pp. 56–61.

21. Newman M. E. J. Scientific collaboration networks. I. Network construction and fundamental results. *Phys. Rev. E*., 2001, no. 1 (64), pp. 64–71.

22. Nicolas M., Khelif. K. Focused Crawling Using Name Disambiguation on Search Engine Results. *Intelligence and Security Informatics Conference (EISIC)*, 2011, pp. 340–345.

23. ORCID. Aviable at: http://orcid.org/ (accessed 4 October 2013).

24. ResearcherID. Available at: http://www.researcherid.com/ (accessed 4 October 2013).

25. ResearchGate. Available at: http://researchgate.net (accessed 4 October 2013).

26. Rui Z., Shen D., Kou Yu., Nie T. Author Name Disambiguation for Citations on the Deep Web. *Web-Age Information Management: WAIM 2010 International Workshops: IWGD 2010, XMLDM 2010, WCMT 2010*, Jiuzhaigou Valley, China, July 15–17, 2010 Revised Selected Papers. Springer Berlin Heidelberg, 2010, pp. 198–209.

27. Segaran T. *Programming Collective Intelligence*. O'Reilly Media, 2007. 362 p.

28. Shin D., Kim T., Jung H., Choi J. Automatic Method for Author Name Disambiguation Using Social Networks. *Advanced Information Networking and Applications (AINA)*: 24th IEEE International Conference, 2010, pp. 1263–1270.

29. Smalheiser Neil R., Torvik Vetle I. Author name disambiguation. *Annual Review of Information Science and Technology*, 2009, no. 1 (43), pp. 1–43.

30. Strotmann A., Zhao D., Bubela T. Author name disambiguation for collaboration network analysis and visualization. *Proc. Am. Soc. Info. Sci. Tech*, 2009, no. 46, pp. 1–20.

31. Strotmann A., Dangzhi Z. Author name disambiguation: What difference does it make in author-based citation analysis? *Journal of the American Society for Information Science and Technology*, 2012, no. 9 (63), pp. 1820–1833.

32. Torvik Vetle I., Weeber Marc, Swanson Don R., Smalheiser Neil R. A probabilistic similarity metric for Medline records: A model for author name disambiguation. *Journal of the American Society for Information Science and Technology*, 2005, no. 2 (56), pp. 140–158.

33. Treeratpituk P., Lee Giles C. Disambiguating authors in academic publications using random forests (2009). *JOINT CONFERENCE IN DIGITAL LIBRARIES (JCDL '09)*, New York, 2009, pp. 39–48.

34. Yang K.-H., Peng H.-T., Jiang J.-Y., Lee H.-M., Ho J.-M. Author Name Disambiguation for Citations Using Topic and Web Correlation (2008). Research and Advanced Technology for Digital Libraries: 12th European Conference, ECDL 2008, Aarhus, Denmark, September 14–19, 2008. Springer Berlin Heidelberg, 2008, pp. 185–196.

35. Yang Kai-Hsiang, Wu Yi-Hsuan. Author Name Disambiguation in Citations (2011) WI-IAT '11. *Proceedings of the 2011 IEEE/WIC/ACM: International Conferences on Web Intelligence and Intelligent Agent Technology*, 2011, vol. 3, pp. 335–338.