УДК 001.94 + 519.24
ББК 72.4 + 78.5

ЧТО МОЖНО УЛУЧШИТЬ В НАУКОМЕТРИЧЕСКОМ АНАЛИЗЕ – УЧЕТ НАЛИЧИЯ ДУБЛИКАТОВ И ЗАИМСТВОВАНИЙ В НАУЧНЫХ ПУБЛИКАЦИЯХ

Дербенёв Н. В.[1], **Толчеев В. О.**[2]
(*Национальный исследовательский университет «Московский энергетический институт», Москва*)

Дается общая характеристика наукометрических методов, отмечаются их недостатки, анализируются возможности применения и рассматриваются направления, по которым целесообразно разрабатывать новые подходы. Предлагается наряду с известными наукометрическими процедурами реализовать в информационно-аналитических системах оценки научной деятельности методы выявления дубликатов и плагиата. Приводятся результаты обработки больших массивов научных документов и примеры обнаруженных дубликатов.

Ключевые слова: наукометрические методы и индикаторы, российский индекс научного цитирования, информационно-аналитическая система, выявление дубликатов и нечетких (неполных) дубликатов, плагиат.

1. Общая оценка наукометрических методов

С момента своего появления наукометрические методы находятся в центре непрекращающейся и со временем только усиливающейся дискуссии, получая весьма неоднозначные оценки в научном сообществе. При этом как пропагандисты этих методов, так и их противники признают, что наукометрические пока-

[1] *Николай Викторович Дербенёв – старший преподаватель кафедры Управления и информатики НИУ МЭИ, nicvic@mail.ru.*
[2] *Владимир Олегович Толчеев – доктор технических наук, профессор кафедры Управления и информатики НИУ МЭИ, tolcheevvo@mail.ru.*

затели лишь косвенно свидетельствуют о качестве научных исследований, реальном вкладе ученого в развитие предметной области.

Основная критика ведется по нескольким направлениям:

– неточность наукометрических оценок, которая возникает из-за неполноты информационной базы, используемой для анализа;

– несовершенство применяемого инструментария (упрощенный аппарат математической статистики, введение ряда допущений, которые редко выполняются на практике);

– произвольная интерпретация результатов наукометрического анализа при составлении различных рейтингов (ученых, научных коллективов, вузов, институтов РАН) и выработке управленческих решений.

Это все создает наукометрии репутацию очень «лукавого зеркала науки», в котором каждый может увидеть то, что хочет увидеть.

Негативное отношение к наукометрии только усилилось после привязки карьерного роста и зарплаты ученых, возможности защитить докторскую диссертацию, получить грант РФФИ к наукометрическим показателям (в частности, «количеству журнальных публикаций» и «цитируемости») [8, 9, 14]. Так, в [9] справедливо отмечается, что использование показателя «количество журнальных публикаций» приводит к существенному сужению информационной базы исследования, исключает из нее доклады на конференциях, тематические сборники, монографии и учебники. Кроме того, хорошо известны и другие проблемы наукометрического анализа: можно ли считать равнозначными публикации в различных изданиях, насколько зависит число публикаций и индекс цитирования от отрасли знаний, как учесть отрицательное цитирование (самоцитирование), есть ли способы обнаружения «навязанного соавторства» (руководитель–подчиненный) или «мертвых душ», не участвовавших в исследованиях и т.п.

Вместе с тем, несмотря на активную критику наукометрии за контрпродуктивность, пока не удалось предложить более внятную, точную и реализуемую на практике систему «измерения» науки. Вряд ли можно считать, что использование прагматических показателей, процедур вебометрики, привлечение экс-

пертных оценок позволит заменить наукометрический анализ (хотя использование этих подходов совместно с наукометрическими методами, на наш взгляд, может улучшить результирующую точность получаемых оценок) [2, 7, 8, 13, 14].

Нельзя сказать, что проблемы наукометрии не известны тем, кто создает информационно-аналитические системы (ИАС) и использует наукометрические показатели для приятия управленческих решений. Шаги, предпринимаемые с целью повышения «правдоподобности и достоверности» оценок научной деятельности, представляются вполне правильными и уместными. Прежде всего, это относится к созданию ИАС «РИНЦ» (Российский индекс научного цитирования) и разработке дополнительных сервисов и инструментов «измерения» науки в ИАС «SCIENCE INDEX». В конце 2012 года Минобрнауки начало реализацию нового проекта по построению специализированной ИАС «Карта российской науки» для выявления активных и конкурентоспособных коллективов ученых, изучения научного задела и кадрового «ландшафта» в предметных областях [10].

Уже сейчас можно отметить определенные позитивные сдвиги:

– расширяется информационная база, доступная для исследований (в частности, в ИАС «SCIENCE INDEX» ведется работа по учету всех видов научных публикаций: монографий, сборников статей, материалов конференций, патентов, отчетов по НИОКР и т.п.) [3],

– появляются новые программные персонифицированные сервисы и инструменты анализа [3],

– предлагаются более комплексные и чувствительные индикаторы, которые комбинируют *показатели активности*, отражающие интенсивность публикации печатных трудов, и *показатели влияния*, характеризующие полезность научных идей для других специалистов [4, 11].

Разделяя мнение, что «измерять» научную деятельность нужно, и другой работоспособной альтернативы пока не предложено (и в перспективе принципиально новых вариантов решения задачи не просматривается), было бы правильно сместить акцент в дискуссии с критики современной наукометрии на вопросы улучшения ее инструментария. Прежде всего, надо со-

средоточиться на адаптации наукометрических методов к новым вызовам, появившимся на современном этапе лавинообразного роста числа научных журналов, конференций, других информационных ресурсов. Все эти данные (постепенно) заносятся в информационную базу и становятся доступными для анализа. Остается открытым вопрос: готовы ли разработчики ИАС и идеологи наукометрических оценок к тому, чтобы качественно обрабатывать столь большие и разнородные документальные потоки, совершенствовать имеющийся инструментарий?

2. Об одном из возможных направлений развития наукометрического анализа

Удивительным представляется то, что, несмотря на огромное внимание, которое уделяется в наукометрии анализу текстов статей, их тематик, размеров, библиографических ссылок и цитирования, авторства-соавторства, тем не менее в наукометрических исследованиях практически полностью игнорируется такой феномен, как тиражирование авторами одних и тех же материалов (издание статей-дубликатов) и не уделяется должного внимания борьбе с распространением документов с существенными заимствованиями (т.е. плагиатом – умышленном использовании фрагментов чужих работ без указания источника заимствования или изданием под своим именем чужой статьи).

Конечно же, можно отнести такие явления к неизбежному информационному шуму (мусору) и считать событиями крайне редкими и нехарактерными для российской науки. Однако это очень наивный взгляд на проблему, и он не соответствует, как будет показано ниже, реальной действительности. Здесь нельзя не согласиться с А.И. Орловым, который пишет, что «если лет двадцать назад надо было перепечатывать текст, вставлять формулы, то сейчас с помощью текстового редактора, интернета и/или принтера технические сложности снимаются – статьи можно печь как блины» [9]. И такие «блины» уже пекутся, пока, к счастью, не в массовом порядке. Однако при расширении информационной базы и включении в наукометрический анализ материалов научных конференций количество документов-дубликатов (статей-клонов) может существенно возрасти. Это

приведет к тому, что индикатор «количество публикаций» автора будет давать «завышенные» результаты, включая как уникальные, так и неуникальные работы, сделанные путем переиздания одних и тех же результатов в разных журналах и конференциях (часто даже без каких-либо изменений в названиях и аннотациях).

Можно констатировать, что тиражирование научных статей-клонов и публикаций с высоким уровнем заимствований при очевидной простоте реализации имеют ничтожную вероятность выявления. И этот «намеренный» пропуск информационного шума связан не с отсутствием средств обнаружения (таких средств много, их сравнительный анализ приведен, например в [5, 6, 12]), а с тем, что в рамках существующих ИАС уделяется очень мало внимания идентификации таких публикаций.

Несомненно, представляется правильным и своевременным использование в ИАС «SCIENCE INDEX» для выявления оригинальности научных статей системы «Антиплагиат» [1]. Возможно, использование «Антиплагиата» закроет одну из «брешей» наукометрии и повысит эффективность борьбы с материалами, содержащими большое число заимствований.

Однако эта мера вряд ли позволит эффективно выявлять статьи-клоны (хотя задачи плагиата и дублирования информации весьма близки, тем не менее, они не решаются с помощью одного универсального метода). Фактически обнаружение дубликатов в ИАС «SCIENCE INDEX» перекладывается на самого автора путем предоставления «возможности удаления из списка своих работ или цитирований ошибочно попавшие туда публикации» (если, конечно, под такими «статьями для редактирования» разработчики ИАС понимают в том числе и дубликаты).

Для дальнейшего изложения материала и обоснования необходимости тщательного анализа научных публикаций с целью выявления в них неуникальных статей дадим определение дубликатам (нечетким дубликатам). К *дубликатам* (*неуникальным публикациям*) принято относить документы с идентичным (полностью совпадающим) содержанием. *Нечеткими* (*неполными*) *дубликатами* или *почти дубликатами* считаются документы, в содержательную часть которых внесены незначительные изменения.

Отметим, что дубликаты могут появляться не только из-за желания автора (авторов) растиражировать одни и те же результаты и сведения, но и по независящим от него причинам:

– ошибочное добавление статьи в базу данных из-за опечаток в названии (или номере) журнала, фамилиях авторов, названиях и аннотациях, неправильного указания номеров страниц и их количества;

– «недобросовестных» действий соавторов, которые самостоятельно (без согласования с другими авторами) осуществляют переиздание документов, изменяют последовательность фамилий соавторов и т.п.

Для более формального определения дубликатов и нечетких дубликатов введем меру близости $\rho(\vec{X}_j, \vec{X}_l)$, значения которой изменяются в интервале [0, 1]. Здесь $\vec{X}_j$, $\vec{X}_l$ – публикации, представленные в виде векторов, элементы векторов – термины, использованные в статье. Мера близости должна равняться единице в случае, если документы $\vec{X}_j$, $\vec{X}_l$ – дубликаты, и стремиться к нулю, если нет. Тогда:

- два документа $\vec{X}_j$, $\vec{X}_l$ считаются *полными дубликатами*, если мера близости $\rho(\vec{X}_j, \vec{X}_l)$ равна единице;
- два документа $\vec{X}_j$, $\vec{X}_l$ считаются *нечёткими дубликатами*, если мера близости $\rho(\vec{X}_j, \vec{X}_l)$ превосходит экспериментально установленный порог θ ($\rho(\vec{X}_j, \vec{X}_l) \geq \theta$).

Пороговое значение θ выбирается экспертно (или экспериментально) и зависит от цели исследования, специфики документального массива и используемой меры близости. Субъективность определения порога является одним из наиболее «уязвимых» мест всех процедур выявления нечетких дубликатов (и заимствований).

Аналогично можно определить плагиат, отличие будет заключаться в том, что в случае поиска дубликатов анализируются два документа одного автора (группы авторов), в случае обна-

ружения плагиата $\vec{X}_j$ и $\vec{X}_l$ – статьи разных авторов, и происходит сравнение новой работы с текстами, уже имеющимися в хранилище информации, на предмет наличия общих фрагментов. Отсюда очевидны существенные различия между дублированием информации («растиражировал самого себя») и плагиатом («украл у другого»).

Необходимо отметить, что окончательное суждение о том, являются ли статьи (нечеткими) дубликатами и содержат ли они некорректно оформленные заимствования, можно сделать только с помощью экспертов на основе изучения и оценки полнотекстовых вариантов статей.

Далее основное внимание будет уделено вопросу выявления (нечетких) дубликатов, как наименее разработанному (но, на наш взгляд, очень востребованному) направлению в области обработки и анализа текстовых документов.

Остается без ответа важный вопрос – действительно ли в больших массивах научных статей российских ученых встречаются дубликаты и нечеткие дубликаты или это отдельные артефакты, редкие события, на обнаружение которых не стоит тратить силы и средства. В следующем разделе мы постараемся ответить на этот вопрос на основе экспериментальных исследований.

3. *Исследования на выборках, примеры (нечетких) дубликатов*

В данной работе для получения выборок документов и проведения экспериментальных исследований были использованы возможности, предоставляемые научной электронной библиотекой eLibrary.ru. С помощью авторского указателя ИАС «РИНЦ» было отобрано 1070 авторов, которые занимаются исследованиями в области автоматики и вычислительной техники и имеют больше 10 печатных работ (в первую очередь выбирались публикации, вышедшие в различных изданиях в течение последних нескольких лет). Таким способом была сформирована выборка из 7257 *библиографических документов* (название, аннотация, ключевые

слова, инициалы авторов, место издания). Обратим внимание, что все экспериментальные исследования осуществлялись на основе анализа именно библиографических документов, имеющихся в свободном доступе в ИАС «РИНЦ».

В нашей статье приводятся результаты, полученные при использовании в качестве меры близости $\rho(\vec{X}_j, \vec{X}_l)$ коэффициента ассоциативности Джаккарда (аналогичные результаты были получены для метода шинглов и расстояния Джаро–Винклера [5]). Коэффициент ассоциативности Джаккарда рассчитывается по формуле

$$J(\vec{X}_j, \vec{X}_l) = \frac{\left|\vec{X}_j \cap \vec{X}_l\right|}{\left|\vec{X}_j \cup \vec{X}_l\right|} = \frac{A}{A + B + C}.$$

Здесь A – число совпавших терминов в двух документах $\vec{X}_j$ и $\vec{X}_l$; B – число терминов, имеющихся в $\vec{X}_j$ и отсутствующих в $\vec{X}_l$; C – число терминов, имеющихся в $\vec{X}_l$ и отсутствующих в $\vec{X}_j$.

Результаты экспериментальных расчетов позволили сделать следующие выводы по распределению значений меры близости в интервале [0, 1]. Так, в интервале [0,6; 0,8] оказалось 264 документа, в интервале [0,8; 0,9] – 114 документа, в интервале [0,9; 1] – 72 документа (в интервале [0,5; 0,6] содержалось 80 документов, остальные статьи попали в диапазон [0; 0,5]). Было также выявлено 178 полных дубликатов, которые имели коэффициент Джаккарда равный единице из-за допущенных в библиографических описаниях опечаток (чаще всего для одной и той же статьи в eLibrary.ru указывались разные страницы или номера журналов, см. пример 1).

Для обсуждаемых в данной работе проблем важно проанализировать полные дубликаты и нечеткие дубликаты, попадающие в интервал [0,8; 1]. Приведем пример полных дубликатов, для которых коэффициент Джаккарда равнялся единице (полужирным шрифтом выделены различающиеся фрагменты документов, из-за этических соображений в примерах используются только инициалы авторов).

Пример 1.

Автор(ы): Ш.А.М., Г.С.В., К.В.А., **Б.В.Р.**, М.Н.М., О.С.И. Место публикации: Автоматизация в промышленности. 2009. **№2.** С. 4–7. Название: СРАВНЕНИЕ РАБОТЫ АДАПТИВНОГО ПИД-РЕГУЛЯТОРА С ОПТИМАЛЬНЫМ НЕЛИНЕЙНЫМ В РЕЖИМАХ ИНТЕГРАЛЬНОГО НАСЫЩЕНИЯ Аннотация: Интегральное насыщение возникает в линейных ПИД(ПИ)-регуляторах при выходе управляющего сигнала за пределы линейной зоны. Существуют различные алгоритмы устранения отрицательного эффекта интегрального насыщения применительно к ПИД(ПИ)-регуляторам с постоянными настройками. В работе предлагается алгоритм работы адаптивного ПИД(ПИ)-регулятора в условиях интегрального насыщения.
Автор(ы): **Б.Н.М.**, Ш.Л.М., Г.С.В., К.С.И., М.В.А., О.В.Р. Место публикации: Автоматизация в промышленности. 2009. **№1.** С. 4–7. Название: СРАВНЕНИЕ РАБОТЫ АДАПТИВНОГО ПИД-РЕГУЛЯТОРА С ОПТИМАЛЬНЫМ НЕЛИНЕЙНЫМ В РЕЖИМАХ ИНТЕГРАЛЬНОГО НАСЫЩЕНИЯ Аннотация: Интегральное насыщение возникает в линейных ПИД(ПИ)-регуляторах при выходе управляющего сигнала за пределы линейной зоны. Существуют различные алгоритмы устранения отрицательного эффекта интегрального насыщения применительно к ПИД(ПИ)-регуляторам с постоянными настройками. В работе предлагается алгоритм работы адаптивного ПИД(ПИ)-регулятора в условиях интегрального насыщения.

Очевидно, что появление этих двух одинаковых документов в «РИНЦ» связано с опечатками в номерах журнала и изменении последовательностей фамилий авторов.

Рассмотрим два других примера, которые относятся к почти-дубликатам (значение меры близости находится в интервале [0,8; 1]).

Пример 2.

Автор(ы): А.Т.Г., М.Е.П., Я.В.П.

Место публикации: **Известия Российской академии наук. Теория и системы управления. 2007. №5. С. 5–10.**

Название: УКЛОНЕНИЕ **ОТ ОБНАРУЖЕНИЯ** В ТРЕХМЕРНОМ ПРОСТРАНСТВЕ

Аннотация: Рассматривается дифференциальная игра одного преследователя против группы из истинной и ложной целей в трехмерном пространстве, в которой согласованно действующие цели решают задачу уклонения истинной цели от обнаружения преследователем.

Автор(ы): А.Т.Г., М.Е.П., Я.В.П.

Место публикации: **Автоматика и телемеханика. 2008. №5. С. 1–14.**

Название: УКЛОНЕНИЕ **ГРУППОВОЙ ЦЕЛИ** В ТРЕХМЕРНОМ ПРОСТРАНСТВЕ

Аннотация: Рассматривается дифференциальная игра одного преследователя против группы из истинной и ложной целей в трехмерном пространстве, в которой согласованно действующие цели решают задачу уклонения истинной цели от обнаружения преследователем.

Пример 3.

Автор(ы): Л.В.И.

Место публикации: **Безопасность в техносфере. 2009. №2. С. 4–5**

Название: **ЛОГИКО-АВТОМАТНОЕ** МОДЕЛИРОВАНИЕ БЕЗОПАСНОСТИ ОКРУЖАЮЩЕЙ СРЕДЫ

Аннотация: Предложена **логико-автоматная** модель безопасности окружающей среды. В ней зависимость состояния u (безопасное: $u = 0$, опасное: $u = 1$) среды от аналогичных состояний отдельных ее факторов x_i описывается булевой функцией $u = f(x_i)$. Дается аналитическое решение с использованием логической теории динамических автоматов.

Автор(ы): Л.В.И.

Место публикации: **Вестник Тамбовского университета. Серия: Естественные и технические науки 2008. Т. 13, №5. С. 395–396.**

Название **АВТОМАТНО-ЛОГИЧЕСКОЕ** МОДЕЛИРОВАНИЕ БЕЗОПАСНОСТИ ОКРУЖАЮЩЕЙ СРЕДЫ

Аннотация: Предложена **автоматно-логическая** модель безопасности окружающей среды. В ней зависимость состояния u (безопасное: $0 = u$, опасное: $1 = u$) среды от аналогичных состояний отдельных ее факторов x_i описывается булевой функцией $u = f(x_i)$. **Известны процессы $x_i(t)$, где t – время. Требуется найти процесс $y(t)$.** Дается аналитическое решение с использованием логической теории динамических автоматов.

Каждый может делать свои собственные заключения и предположения по поводу приведенных примеров 2 и 3. Как указывалось выше, мы не можем назвать эти публикации дубликатами без дополнительного экспертного изучения полнотекстовых документов. На основе нашего анализа (когда удавалось найти полнотекстовые версии в открытом доступе) можно сделать вывод, что не всегда практически полное совпадение названий и аннотаций соответствует одинаковым полнотекстовым статьям (чаще всего такие «дубликаты» выявляются на уровне анализа библиографических описаний из-за небрежного составления аннотаций, например, переписывания с предыдущей статьи).

Важно также отметить другую, наиболее типичную причину «ложного» обнаружения нечетких дубликатов по библиографическим описаниям. Эта причина обуславливается самой спецификой научной деятельности, когда ученый после опубликования первых результатов изучения и осмысления проблемы продолжает исследования и предлагает оригинальные (модифицирует известные) модели и методы, излагая их в новых статьях. Это приводит к появлению тематически близких *«связанных» публикаций*. Такие публикации, несомненно, являются уникальными и отражают последовательные этапы проведения НИОКР. Однако очень часто этим документам соответствуют практически одинаковые названия и аннотации (не из-за дублирования

информации, а из-за использования общей для предметной области терминологии). Именно разделение работ на дубликаты и «связанные» публикации является наиболее сложной и нетривиальной задачей в области автоматизированного выявления статей-клонов.

Таким образом, мы не можем сделать корректного заключения о том, являются ли статьи, представленные в примерах 2 и 3, (нечеткими) дубликатами. К сожалению, этого не делает и ИАС «SCIENCE INDEX» (или ИАС «РИНЦ»), в которой можно (при желании) достаточно просто реализовать анализ как библиографических, так и полнотекстовых версий статей. При этом использование хорошо известных и апробированных решений позволило бы избежать высокой ресурсоемкости и трудозатратности обработки текстов.

4. Заключительные предложения

Как представляется, разработка и внедрение процедур выявления нечетких дубликатов способны существенным образом расширить инструментарий наукометрии и стать барьером для информационного шума, в частности предотвратить тиражирование статей-клонов. На практике для обнаружения нечетких дубликатов из потока научной периодики достаточно дополнить, например, ИАС «РИНЦ» (или ИАС «SCIENCE INDEX») программно-алгоритмическим модулем, позволяющим анализировать публикации автора на совпадения. Процесс поиска и идентификации дубликатов может быть достаточно легко формализован: на первом этапе на основе анализа коротких библиографических описаний определяются публикации – кандидаты в дубликаты, на втором этапе проверяются на идентичность их полнотекстовые версии (на третьем этапе, если необходимо, проводится экспертиза с привлечением специалистов-предметников). Для документов, у которых обнаружено практически полное совпадение названий, аннотаций и текстов статей, можно предусмотреть специальную категорию – «нечеткие дубликаты». Количество таких публикаций следует указывать для каждого автора наряду с показателями самоцитирования, цитирования соавторами, коэффициентом Хирша и т.п. Возможно,

наряду с индикатором «количество публикаций» правильно будет вычислять скорректированный показатель, при расчете которого не используются имеющиеся (нечеткие) дубликаты.

Конечно же, введение новых показателей может вернуть нас обратно к началу дискуссии – то́ ли и правильно ли мы мерим; надо ли, понимая грубость измерений, что-то уточнять и усложнять, вводить дополнительные индикаторы, которые обязательно вызовут споры и, к тому же потребуют экспертного (экспериментального) определения ряда значений, например, величины порога θ?

Все-таки, на наш взгляд, в условиях существенного роста потока научных документов, неизбежного увеличения информационной базы «РИНЦ» разработка упреждающих средств отсечения информационного шума была бы весьма своевременна и важна для совершенствования инструментария наукометрического анализа и повышения доверия к получаемым оценкам.

Литература

1. *Антиплагиат: Интернет-ресурс* [Электронный ресурс] URL: http://www.antiplagiat.ru (дата обращения 28.07.2013)
2. АНТОПОЛЬСКИЙ А.Б. *Использование информационных ресурсов для оценки эффективности научных исследований* // Межотраслевая информационная служба. – 2011. – №1. – С. 40–53.
3. АРЕФЬЕВ П.Г., ЕРЕМЕНКО Г.О., ГЛУХОВ В.А. *Российский индекс научного цитирования – инструмент для анализа науки* // Библиосфера. – 2012. – №5. – С. 66-71.
4. ГЕРМАШЕВ И.В., СИЛИНА А.Ю., ВАСИЛЬЕВА В.Д., ДЕРБИШЕР В.Е. *Обработка нечетких данных для оценки активности научной деятельности* // Информационные технологии. – 2008. – №12. – С. 12–14.
5. ДЕРБЕНЁВ Н.В., ТОЛЧЕЕВ В.О. *Выявление нечетких дубликатов в наукометрическом анализе* // Информационные технологии. – №12. – 2011. – С. 24–29.
6. ЗЕЛЕНКОВ Ю.Г., СЕГАЛОВИЧ И.В. *Сравнительный анализ методов определения нечетких дубликатов для Web–*

документов // Труды 9-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции». – Переславль-Залесский: Изд-во ИПС РАН, 2007. – С. 166–174.

7. КАЛЕНОВ Н.Е., СЕЛЮЦКАЯ О.В. *Некоторые оценки качества российского индекса научного цитирования на примере журнала «Информационные ресурсы России»* // Информационные ресурсы России. – 2010. – №6. – С. 2–13.
8. КОТЛЯРОВ И.Д. *Новый метод оценки продуктивности научной деятельности* // Библиосфера. – 2010. – №2. – С. 60–66.
9. ОРЛОВ А.И. *Два типа методологических ошибок при управлении научной деятельностью* // Управление большими системами. – 2013. – № 44 – С. 32–54.
10. *ПОИСК* №51 от 21 декабря 2012. – С. 5.
11. СИЛИНА А.Ю., ВАСИЛЬЕВА В.Д., ДЕРБИШЕР В.Е., ГЕРМАШЕВ И.В. *Систематизация наукометрических показателей эффективности научной деятельности* // Информационные технологии. – 2009. – №6. – С. 53–56.
12. ТОЛЧЕЕВ В.О. *Анализ проблемы и разработка процедуры выявления нечетких дубликатов научных статей по библиографическим описаниям* // Информационные технологии. – 2011. – №2. – С. 17–21.
13. ФЕДОРЕЦ О.В. *Коллективная экспертиза научных журналов: методика агрегирования экспертных оценок и построения рейтинга* // Управление большими системами: сборник трудов. – 2009. – №27. – С. 18–35.
14. ЭПШТЕЙН В.Л. *О контрпродуктивности использования наукометрического показателя результативности научной деятельности для будущего России* // Control Science. – 2007. – №3. – С. 70–72.

WHAT CAN BE IMPROVED IN SCIENCEMETRICS – NEAR-DUPLICATES AND PLAGIARISM DETECTION IN SCIENTIFIC PUBLICATIONS

Nicolay Derbenev, National Research University "Moscow Power Engineering Institute", assistant professor (nicvic@mail.ru)
Vladimir Tolcheev, National Research University "Moscow Power Engineering Institute", Doctor of Science, professor (tolcheevvo@mail.ru)

Abstract: *We conduct the analysis of scientometrics methods, consider their advantages and disadvantages*, *discuss a new*, *rather promising*, *line of development. We suggest applying the methods of near-duplicates and plagiarism detection in scientometric analysis. We give some results of processing a big array of research papers, and provide examples of uncovered near-duplicates.*

Key words: scientometric methods and indicators, Russian index of scientific citation, information-analytical system, duplicates (near-duplicates) detection, plagiarism.

Поступила в редакцию 26.01.2013.
Опубликована 31.07.2013.