

С.В. Клименко¹
 М.М. Шарнин²
 А.Х. Хакимова³
 О.В. Золотарев⁴
 А.Г. Мацкевич⁵

S.V. Klimenko
 M.M. Charnine
 A.Kh. Khakimova
 O.V. Zolotarev
 A.G. Matskevich

МЕТОДЫ ОЦЕНКИ КАЧЕСТВА И ВЛИЯНИЯ (ИМПАКТ) НАУЧНЫХ СТАТЕЙ ДЛЯ ПОВЫШЕНИЯ ОБЪЕКТИВНОСТИ ИНДЕКСА НАУЧНОГО ЦИТИРОВАНИЯ⁶

METHODS OF ASSESSING THE QUALITY AND INFLUENCE (IMPACT) OF SCIENTIFIC ARTICLES TO IMPROVE THE OBJECTIVITY OF THE SCIENCE CITATION INDEX

Важная современная тенденция в системе оценки работ ученых заключается не только в учете формальных библиографических ссылок на публикуемые статьи, но также в анализе содержания текстов научных работ. Предлагаемый подход к анализу текстов состоит в выявлении неформальных (неявных) ссылок в тексте на авторские идеи. Разработана методика выявления неявных ссылок, которая обеспечивает высокую корреляцию неявных ссылок с формальными. В процессе дальнейших исследований предполагается разработка мер по повышению корреляции явных и неявных ссылок. Это позволит более точно и объективно оценивать не только работы ученых, но и взаимовлияние идей в различных документах, даже в таких, где редко используются формальные ссылки, как, например, в интернет-текстах.

Ключевые слова: цитирование документов, явные ссылки, неявные ссылки, подобие документов, предметная область, значимые словосочетания.

An important current trend in the assessment of the work of scientists is not only to account the formal references in published articles, but also in the analysis of the content of scientific texts. The proposed approach to the analysis of texts is to identify the informal (implicit) references in text to the author's ideas. The developed method of detecting the implicit links provides a high correlation of implicit links with the formal links. In the course of further studies there is planned to develop measures to improve the correlation of the explicit and implicit links. This will allow you to more accurately and objectively assess not only the works of scientists, but also to the interaction between ideas in different documents, even in those that rarely use the formal links, such as, for example, in the online Internet texts.

Keywords: citation of documents, explicit links, implicit links, similarity of documents, subject area, meaningful word combinations.

1. Введение

В работе рассматриваются методы оценки качества научных статей и их влияния (импакт) на развитие предметной области. Работы выполняются в рамках грантов РФФИ:

16-07-00756 – Исследование и разработка семантических методов построения «Индекса контекстного научного цитирования»;

¹ Доктор физико-математических наук, профессор, генеральный директор АНО ИФТИ.

² Кандидат технических наук, ведущий научный сотрудник Института проблем информатики Российской академии наук Федерального исследовательского центра «Информатика и управление» Российской академии наук.

³ Кандидат биологических наук, ведущий научный сотрудник АНО ИФТИ.

16-29-09527 – Исследования и разработка методов тематического моделирования для мониторинга, прогноза и визуализации террористической активности в информационном поле Интернета с использованием виртуального окружения;

⁴ Кандидат технических наук, доцент, заведующий кафедрой информационных систем в экономике и управлении АНО ВО «Российский новый университет».

⁵ Старший научный сотрудник Института проблем информатики Российской академии наук Федерального исследовательского центра «Информатика и управление» Российской академии наук.

⁶ Работа выполнена при поддержке Российского фонда фундаментальных исследований, гранты 16-07-00756, 16-29-09527, 15-07-06586 и 16-37-50057.

15-07-06586 – Исследование и разработка лингвостатистических методов и алгоритмов автоматического формирования ассоциативно-иерархического портрета предметной области на основе онтологий;

16-37-50057 – Исследование корреляции библиографических и неявных ссылок, выявленных методом тематического моделирования коллекций документов, для определения значимых документов и перспективных направлений.

Предлагаются подходы к созданию более объективного индекса научного цитирования, учитывающего неявные ссылки между статьями. Исследование ведется на примере коллекции по тематике «Компьютерная графика, визуализация и виртуальное окружение в СССР и России», созданной авторским коллективом.

Традиционные библиометрические показатели, такие, как количество цитирований и Индекс научного цитирования, часто используются, чтобы найти самые влиятельные и качественные научные статьи. Современные, более точные, библиометрические методы анализа графа цитирований учитывают также содержание статей. Различные методы были предложены для анализа библиографических ссылок, которые вместе образуют граф цитирований. Граф цитирования – это направленный граф, вершины которого соответствуют документам, рёбра – отношению цитирования. Для формирования графа цитирования необходимо выполнить связывание метаописаний документов с выделенными в пристатейных списках литературы библиографическими записями [6].

Научная публикация считается имеющей более высокие индекс и качество, если она имеет большее влияние (*impact*) на последующие работы. Значение научной работы на начальных этапах измерялось в основном через метрики, основанные только на количестве цитирований [1; 2]. Важной тенденцией современных исследований является анализ влияния как на граф цитирования, так и на тексты и идеи в последующих статьях.

Цитирование является неотъемлемой частью науки. Цитаты используются исследователями для обоснования своих аргументов или сравнения результатов с достижениями других авторов. Цитирование встречается в тексте статьи и может варьироваться от короткого упоминания до нескольких предложений. Чаще всего работа цитируется в одном предложении, а затем неявно вновь упоминается в последующих предложениях [3].

В этом проекте мы конкретно заинтересова-

ны в нахождении всех предложений, которые относятся к конкретной ссылке без использования формальной ссылки. Они будут называться неявным цитированием (в отличие от явного цитирования, где используется формальная ссылка). Такой подход позволит не только обнаружить связи между статьями, но и понять отношение автора к ссылке, поскольку можно ошибочно предположить, что цитаты – всегда положительные показатели. Атар и Тойфель обнаружили, что особенно отрицательные мнения о других работах редко появляются в явных цитатах [4]. Включение неявного цитирования увеличило количество негативных настроений более чем на 300%.

2. Описание коллекции

Для проведения исследований была создана коллекция научных статей по тематике «Компьютерная графика, визуализация и виртуальное окружение в СССР и России», свободно представленных в Интернете. При подготовке к исследованию был составлен список ключевых терминов по предметной области (ПО) «Компьютерная графика, визуализация и виртуальное окружение в СССР и России» (КГВ). Авторским коллективом из открытой информации в Интернете был собран большой объем текстов, связанных с КГВ, размером около 30 ГБ, из которого было выделено ядро, содержащее 350 МБ различных неповторяющихся фраз.

На основании списка ключевых терминов с помощью технологии семантического серфинга в рамках системы Keuwen была собрана коллекция научных статей по КГВ.

Коллекция статей была проанализирована вручную методом краудсорсинга на принадлежность к ПО КГВ и на наличие в них списка литературы (библиографии). Вся коллекция представлена на веб-сайте Ngraph.ru, разработанном авторским коллективом. Документы представлены в формате XML. На сайте представлены анализируемые научные работы, их индекс цитирования в коллекции и наиболее цитируемые авторы. Созданная коллекция является крупнейшей в мире коллекцией русскоязычных статей по компьютерной графике. Для сравнения, в eLibrary (РИНЦ) в настоящее время находится не более 500 статей на эту тему.

Коллекция содержит более 900 документов (статьи, диссертации, монографии), в которых выявлены ссылки еще на 6000 документов. Между полными документами коллекции имеется 119 прямых библиографических ссылок, которые связывают 141 документ из коллекции. В коллекции содержится 82 полнотекстовых до-

кумента, имеющих ссылки в других документах коллекции и ненулевой индекс цитирования. (Например, в коллекции из трех документов А, В и С, связанных двумя ссылками А->В и С->В, имеется только один документ В, имеющий ссылки со стороны других документов А и С.)

Для анализа статей коллекции использовался Лингвистический процессор BREF (Book Reference), разработанный авторским коллективом, который выделяет из статей формальные данные (название, авторы, библиографические ссылки и т.д.) и представляет их в формате XML. Структура XML-файла проста и самодокументируема.

Процессор обработки ссылок и списка литературы BREF реализован на базе лингвистического процессора Pullenti (Puller of Entities, www.pullenti.ru), который под псевдонимом Pink на соревновании FactRuEval конференции Диалог-2016 занял первые места на большинстве дорожек (см. <http://www.dialog-21.ru/evaluation/2016/ner/>). BREF позволяет по выделенной информации построить Граф цитирования и Граф соавторов, отражающих формальные связи в коллекции. Лингвистический процессор (ЛП) также предназначен для выявления похожих по смыслу фраз в других статьях и в документах из Интернета. Лингвистический процессор настраивается с помощью метода машинного обучения.

Тексты статей из БД были пропущены через ЛП, с помощью которого выявлены названия статей, авторы и библиографические ссылки, а также упоминания авторов в текстах статей. Составлен рейтинг упоминаний авторов (РУА). Лингвистический процессор (ЛП) получает на вход полный текст статьи, а на выходе строит формальную структуру на языке XML следующего вида:

```
<RESULT>
<ARTICLE>
<AUTHOR>описание автора</AUTHOR>...
может быть несколько или не быть совсем
<TITLE>наименование</TITLE>
<YEAR>год, если есть</YEAR>
<LANG>базовый язык: RU\EN </LANG>
<LINK>
<NUMBER>номер в списке литературы</
NUMBER>
<AUTHOR>...
<TITLE>...
<YEAR>...
</LINK>
<LINK> ... </LINK>
...
</ARTICLE>
```

```
<ARTICLE>
```

```
...
```

```
</ARTICLE>
```

```
</RESULT>
```

В данной структуре формальные библиографические ссылки помечены метками <LINK> ... </LINK>. При описании автора <AUTHOR> в таком файле: LAST-фамилия, FIRST-имя, MIDDLE-отчество. Из подобных XML-файлов строится, например, Граф цитирования.

При помощи BREF были построены схемы графов соавторов (см. рис. 2, 3), отражающие формальные связи в коллекции. Автоматизированный анализ тестовой БД позволяет строить кластеры, содержащие связанные друг с другом статьи. Кластерный анализ является статистическим методом и в основном применяется для построения тезауруса или детализации запросов к коллекции документов.

В данном проекте кластером называется файл, содержащий некоторую статью А, а также две или более статей, которые формально ссылаются на А. Анализ проводится на основе XML-файлов. При автоматизированной (и частично ручной) обработке подобных XML-файлов необходимо решить следующие задачи:

- найти дубли статей;
- найти кластеры авторов (соавторов);
- найти разделы выделенной предметной области (например, «Компьютерная графика» – КГ);
- построить рейтинг/индекс статей, равный количеству формальных ссылок;
- построить рейтинг/индекс авторов, равный суммарному количеству формальных ссылок;
- найти и устранить ошибки в XML;
- найти похожие названия статей;
- рассчитать корреляцию явных и неявных ссылок [22–24].

Объединение данных о таких формальных ссылках по всем статьям коллекции позволило определить наиболее цитируемые статьи и построить индекс коллекции КГВ (ИККГВ), который для каждой статьи в коллекции равен количеству формальных ссылок на нее из других статей коллекции. При этом для авторов, занимающих лидирующие позиции в рейтинге РУА, найдены наиболее цитируемые статьи, а также построены кластеры статей, содержащих формальные библиографические ссылки на эти наиболее цитируемые статьи.

Разработанные оригинальные программные средства позволяют провести анализ общих фраз в статьях каждого кластера. Был проанализирован лексический состав кластеров и выявлены

слова и фразы, которые встречаются в кластере чаще, чем в большой внешней коллекции на постороннюю тему. Такой анализ помогает понять, как устроены неявные ссылки, а именно: создается файл, где хранится список общих фраз для первой и второй статей; на первом месте этого файла стоит запись «виртуального окружения $v=3$ $V=9$ $Vm=0$ ». Это означает, что словосочетание «виртуального окружения» три раза встретилось в первой ($v=3$) и девять раз во второй статье ($V=9$) и ни разу – в большой коллекции по посторонней по отношению к КГВ теме ($Vm=0$). Фраза «виртуального окружения» является неявной ссылкой на «главную» статью «АВАНГО: система разработки виртуальных окружений».

Аналогичными методами – с помощью программы, анализирующей XML-файл, строятся кластеры авторов и соавторов (участники одного предметного кластера/научной школы часто бывают соавторами). При обработке списка авторов производится транслитерация имен на латинский алфавит с целью устранения написания одного имени несколькими вариантами; проводится сравнение с другими фамилиями с целью исключения разных написаний одного имени. В качестве меры близости используется расстояние Левенштейна. Ошибки приходится устранять ручным методом. Так, например, в библиографической ссылке вида «Ковалевский В.А. О корреляционном методе распознавания / В.А. Ковалевский // Читающие автоматы. – Киев, 1965. – С. 46–61» анализатор выделил двух авторов:

```
<AUTHOR          LAST=«Ковалевский»
FIRST=«В» MIDDLE=«А» />
<AUTHOR          LAST=«О»      FIRST=«В»
MIDDLE=«А» />
```

Очевидно, второй автор «А. В. О» является артефактом.

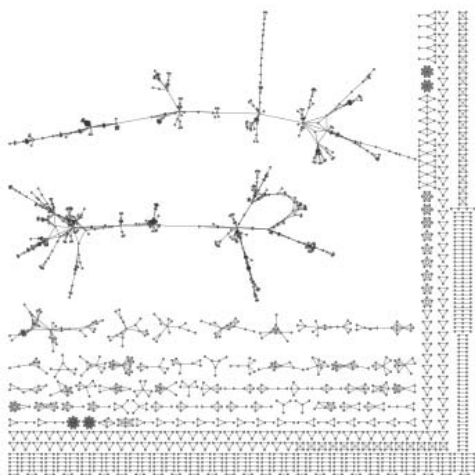


Рис. 1. 2937 вершин (авторов), 3422 ребра

Приведем примеры полученных кластеров авторов и соавторов (рис. 1–3). Большие круги (вершины графа) – это кластеры авторов (научные школы); соавторы хотя бы одной общей статьи соединены ребрами. Подобные результаты в дальнейшем планируется использовать в качестве инструментов исследования, например сравнивая графы авторов/соавторов с графом цитирования. Таким образом можно будет выделить значимые связи между авторами/соавторами, выявить самоцитирование.

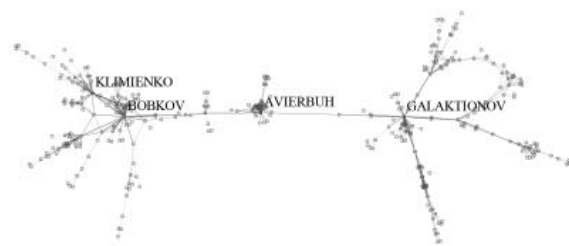


Рис. 2. Наибольшая компонента связности: 338 авторов, 769 рёбер

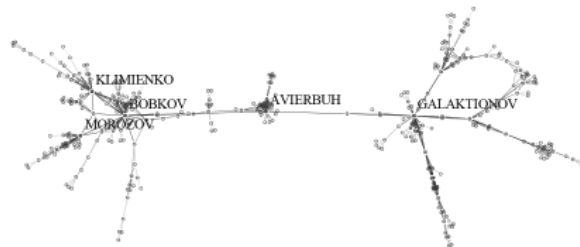


Рис. 3. Граф соавторов (выделены авторы, имеющие не менее 18 соавторов)

3. Ассоциативный портрет предметной области (АППО) и значимые словосочетания (ЗС)

Для изучения статистических свойств коллекции используется методология построения Ассоциативного портрета предметной области (АППО), разработанная авторским коллективом. Руководителем проекта РФФИ 13-07-00272 «Методика автоматического формирования ассоциативных портретов предметных областей на основе естественно-языковых текстов больших объемов для систем извлечения знаний» является руководитель данного исследования М.М. Шарнин [21; 25–27; 29].

Значимые словосочетания/термины (ЗС/ЗТ) – это отдельные слова, словосочетания и фразы, которые определяются на основе частотного словаря коллекции в соответствии с методологией построения АППО. К значимым словосочетаниям в частности относятся значимые термины и все короткие фразы коллекции [28; 30].

Для расчета ассоциативной связи (семантической близости) между ЗС/терминами предметной области используется косинусная мера, сравнивающая параметры контекста ЗС (контекстные векторы). Следует отметить, что посредством ЗС могут быть представлены как отдельные слова, словосочетания и термины, так и более сложные конструкции – объекты и именованные сущности. Иерархические связи между ЗС выбираются из числа ассоциативных связей таким образом, чтобы более общие ЗС имели большее количество ассоциативных связей, при этом учитываются лексико-синтаксические шаблоны, объединяющие данные ЗС.

Для структуризации текстов предметной области и построения иерархии категорий используется методика построения Ассоциативно-иерархического портрета предметной области (отчет РФФИ), в которой для расчета иерархических связей между ЗС/ЗТ используются методы тематического моделирования, такие, как LDA и hLDA.

Выделенные по указанным методикам ассоциативные и иерархические связи между значимыми словосочетаниями и терминами позволяют разрабатывать более совершенные методы и метрики/меры подобия научных текстов.

4. ЗС, Неявные ссылки и Мера подобия научных текстов

Неявная ссылка – это упоминание в тексте авторов, их работ и идей. Неявные ссылки представляются в виде множества значимых словосочетаний (ЗС), т.е. неявная ссылка может быть представлена одним словом/фразой/ЗС или набором близких по смыслу фраз/словосочетаний/ЗС. Разработан метод, выявляющий неявные ссылки между двумя статьями, связанными библиографическими ссылками. При этом неявные ссылки находятся в виде ЗС, получивших наибольшие веса, которые есть в обеих статьях, но отсутствуют в большой внешней коллекции текстов. Весовая функция подбирается так, чтобы наиболее значимые ЗС получили наибольший вес и попали в начало списка.

Для подхода, описанного выше, подобие текста играет важную роль. Тенденция неявного цитирования обычно простирается за пределы одного предложения.

Авторами разработана новая Мера подобия двух научных текстов – это отношение суммы весов их общих ЗС к сумме весов ЗС в меньшем тексте. Мера подобия текста самого на себя равна 100%. В простейшем случае вес ЗС принимается равным количеству символов в ЗС.

Чем больше Мера подобия текстов – тем выше вероятность наличия библиографической ссылки (связи) между этими текстами. Это измерено экспериментально и представлено на следующем графике, демонстрирующем зависимость меры связанности от меры подобия.

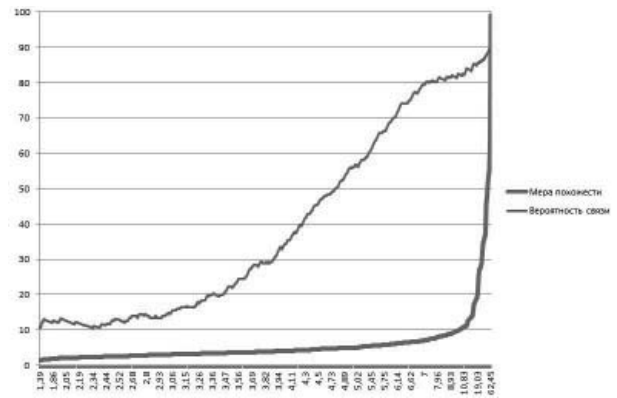


Рис. 4. Зависимость связанности пар документов от их меры подобия (нелинейная шкала подобия с равномерным размещением пар)

Для построения графика взяты 119 напрямую связанных пар статей (мера связанности 100%), 112 косвенно связанных пар (мера связанности 80%) и 350 несвязанных пар (мера связанности 0). Точки на горизонтальной оси соответствуют парам статей (всего отобрана 581 пара), и значения на горизонтальной оси отражают меры подобия этих пар. Мера подобия выбранных пар меняется от 1,39% до 99%. По вертикали для каждой пары отложена мера подобия (жирная линия) и вероятность наличия связи в паре (тонкая линия). Из графика видно, что большинство рассматриваемых пар имеют меру подобия менее 4% и вероятность наличия связи менее 30%.

Можно считать, что между двумя текстами есть неявная ссылка/связь, если их мера подобия составляет более 8%. В этом случае вероятность наличия связи по библиографическим ссылкам между этими текстами более 70%.

Использование данной меры помогает находить пропущенные библиографические ссылки между статьями.

Таким образом, наличие неявной ссылки между статьями можно определить с помощью предложенной Меры подобия текстов.

Дальнейшее развитие Меры подобия связано с совершенствованием функции, определяющей меру веса значимых словосочетаний.

Разработка системы семантических мер сходства может улучшить общие результаты алгоритмов поиска неявного цитирования. Так, например, Murray (2015) описывает введение трех

функций, связанных с текстовым сходством, а именно: TITLE_SIMILARITY (сходство с названием цитируемого документа), CONTENT_SIMILARITY (сходство с содержанием цитируемой работы) и CITE_SIMILARITY (сходство по отношению к другим предложениям, на которые имеются явные ссылки) [3].

Разработана программа, которая выявляет неявные ссылки между двумя статьями, связанными формальными ссылками, например имеющими общую библиографическую ссылку. Такие неявные ссылки находятся в виде фраз и словосочетаний, получивших наибольшие веса, которые есть в обеих статьях, но отсутствуют в большой внешней коллекции текстов на постоянную тему размером в десятки гигабайт.

Для примера рассмотрим результаты анализа программой двух статей, имеющих ссылку на третью статью «С.В. Клименко, И.Н. Никитин, Л.Д. Никитина. АВАНГО: система разработки виртуальных окружений»:

– МЕТОД НАУЧНОЙ ВИЗУАЛИЗАЦИИ как инструмент исследования развития патологии челюстно-лицевой ...

– ИСПОЛЬЗОВАНИЕ СИСТЕМЫ ВИРТУАЛЬНОГО ОКРУЖЕНИЯ ДЛЯ ВИЗУАЛИЗАЦИИ ИЗОБРАЖЕНИЙ ЗЕМНОЙ ПОВЕРХНОСТИ ...

По расчетам, эти две статьи связаны наиболее значимой неявной ссылкой в виде словосочетания «виртуального окружения». Это словосочетание получило наибольший вес при расчете меры подобия этих статей.

Аналогичный алгоритм был применен для поиска всех неявных ссылок в одной статье, для этого в качестве второй статьи был использован корпус интернет-текстов по компьютерной графике размером 350 МБ из неповторяющихся фраз.

Для достижения основной цели – построения Индекса контекстного научного цитирования (ИКНЦ) – необходимо совершенствование этого алгоритма выявления неявных ссылок, с использованием корреляции явных и неявных ссылок в качестве критерия качества алгоритма.

Основные параметры при исследовании корреляции – это длины/размеры фраз/ЗС, количество вхождений их в статью, в корпус текстов по теме и во внешний посторонний корпус. Также в качестве параметров можно использовать замены на синонимы, перевод и грамматические трансформации. В дальнейшем планируется достичь максимизации корреляции путем подбора значений параметров.

5. Корреляция между явными и неявными ссылками

В рассматриваемой коллекции корреляция между мерой подобия 95% статей и их мерой связанности через библиографические ссылки составляет более 50%. Из расчета убраны 5% наиболее похожих пар статей, которые снижают общую корреляцию до 30%. Статьи считаются связанными, если между ними есть прямая библиографическая ссылка (мера связанности равна 5), или если они связаны через другую статью (мера связанности равна 4). Данные величины меры связанности выбраны при максимизации корреляции.



Рис. 5. Максимизация корреляции в зависимости от минимальной длины словосочетаний

На рис. 5 представлен график зависимости корреляции от минимальной длины значимых словосочетаний, которые рассматриваются при расчете корреляции. Из графика видно, что максимум корреляции 55,7% достигается, если мы игнорируем словосочетания с длиной меньше восьми букв. Игнорирование более длинных словосочетаний приводит к уменьшению корреляции. Корреляция рассчитана при игнорировании 5% наиболее похожих пар статей (где наиболее сильна нелинейная зависимость согласно предыдущему графику). Эти наиболее похожие пары статей безусловно должны быть связаны между собой формальными ссылками (несмотря на то, что авторы это не всегда делают), т.к. содержат слишком много общих фраз и идей.

6. Корреляция Меры подобия с Индексом цитирования

Был рассчитан индекс цитирования для каждой статьи по количеству прямых библиографических ссылок на нее из других статей рассматриваемой коллекции.

Также был составлен корпус интернет-текстов по компьютерной графике, из которой было выделено ядро 350 МБ уникальных фраз.

Была измерена мера подобия каждой статьи коллекции на этот интернет-корпус. Корреляция между индексом цитирования и мерой подобия статьи на интернет-корпус – более 15%. Это доказывает возможность осуществлять оценку качества новых статей без библиографических ссылок, пользуясь большими объемами дополнительной информации из Интернета.

7. План дальнейших исследований

Авторский коллектив планирует в дальнейшем сосредоточить усилия в следующих направлениях:

- создание индекса контекстного научного цитирования (ИКНЦ), учитывающего неявные ссылки в других публикациях и корпусе интернет-текстов с использованием предложенной меры подобия статей;

- увеличение размера коллекции до нескольких тысяч статей с помощью автоматического поиска в Интернете и автоматического определения их принадлежности к предметной области «Компьютерная графика, визуализация и виртуальное окружение в СССР и России»;

- определение графа цитирования и количества библиографических ссылок на каждую статью по данным XML, построенного лингвистическим процессором;

- выявление преемственности научных идей для определения их архитектуры в выбранной предметной области;

- сопоставление полученных индексов цитирования (ИЦ) для статей коллекции с ИЦ других баз цитирования (РИНЦ, Scopus, WoS, Google Scholar);

- расчет ИКНЦ-статей коллекции для разных параметров неявных ссылок и максимизация корреляции ИКНЦ с ИЦ по базе данной коллекции и по другим базам цитирования.

Литература

1. Garfield, E., Sher, I.H., Torpic, R.J. The Use of Citation Data in Writing the History of Science. – Philadelphia: Institute of Scientific Information, 1964. – 76 p.

2. Hirsch, J.E. An index to quantify an individual's scientific research output // Proceedings of the National Academy of Sciences of the United States of America. – 2005. – Vol. 102. – No. 46. – Pp. 16569–16572.

3. Murray, J. Finding Implicit Citations in Scientific Publications: Improvements to citation context detection methods. Degree Project, in Computer Science, second level. – Stockholm, Sweden. – 2015. – September 7. – 55 p.

4. Athar, A., Teufel, S. Detection of implicit citations for sentiment detection // Proceedings of the Workshop on Detecting Structure in Scholarly Discourse, Association for Computational Linguistics. – 2012. – Pp. 18–26.

5. Золотарев О.В., Шарнин М.М., Клименко С.В. Семантический подход к анализу террористической активности в сети Интернет на основе методов тематического моделирования // Вестник Российского нового университета. Серия «Сложные системы: модели, анализ и управление». – 2016. – Выпуск 3, в печати.

6. Полежаев В.А. Задачи и методы автоматического построения графа цитирований по коллекции научных документов // Компьютерные исследования и моделирование. – 2012. – Т. 4. – № 4. – С. 707–719.

7. Charnine, M., Klimenko, S. Measuring of “Idea-based” Influence of Scientific Papers // Proceedings of the 2015 International Conference on Information Science and Security (ICISS 2015), December 14–16, 2015. – Seoul, South Korea, 2015. – Pp. 160–164.

8. Ezhela, V.V., Klimenko, S.V., Raikov, A.N., Sharnin, M.M. A Semantic Approach to the Evaluation of the Quality of Academic Publications // Automatic Documentation and Mathematical Linguistics. – 2015. – Vol. 49. – No. 4. – Pp. 117–121.

9. Charnine, M., Somin, N., Klimenko, S., Ezhela, V. Linguistic Approach to Scientometrics // Proceedings of the 2015 International Conference on Artificial Intelligence (ICAI 2015), WORLDCOMP'15, July 27–30, 2015. – Las Vegas, Nevada, USA. – V. II. – Pp. 812–817.

10. Charnine, M., Somin, N., Nikolaev, V. Conceptual Text Generation Based on Key Phrases // Proceedings of the 2014 International Conference on Artificial Intelligence (ICAI 2014). – Vol. I, WORLDCOMP'14, July 21–24, 2014. – Las Vegas, Nevada, USA: CSREA Press. – Pp. 639–643.

11. Михеев М.Ю., Сомин Н.В., Галина И.В., Золотарев О.В., Козеренко Е.Б., Морозова Ю.И., Шарнин М.М. Фальштексты: классификация и методы опознания текстовых имитаций и документов с подменой авторства // Информатика и ее применения. – 2014. – Т. 8. – Вып. 4. – С. 70–77.

12. Шарнин М.М., Сомин Н.В., Галина И.В., Родина И.В., Николаев В.Г. Осмысленные генеранты: метод порождения текстов повышенной информативности из релевантных фраз // Труды Международной конференции и Школы по физико-технической информатике СРТ2014, 11–18 мая 2014 г., Ларнака, Республика Кипр; 17–21 июня 2014 г., Протвино, Московская об-

ласть. – Протвино – М. : Изд. ИФТИ, 2015. – С. 87–91.

13. Charnine, M. Keywen: Automated Writing Tools : монография. – Booktango, USA, 2013.

14. Charnine, M., Charnine, V. Keywen Category Structure : монография. – Wordclay, USA, 2008. – Pp. 1–60.

15. Шарнин М.М., Кузнецов И.П. Автоматическое формирование электронных энциклопедий и справочных пособий по информации из сети Интернет // Системы и средства информатики. – ИПИ РАН. – 2004. – Вып. 14. – С. 210–223.

16. Шарнин М.М., Солин Н.В., Кузнецов И.П., Морозова Ю.И., Галина И.В., Козеренко Е.Б. Статистические механизмы формирования ассоциативных портретов предметных областей на основе естественно-языковых текстов больших объемов для систем извлечения знаний // Информатика и её применения. – 2013. – Т. 7. – № 2. – ИПИ РАН. – С. 92–99.

17. Zolotarev, O., Charnine, M., Matskevich, A. A Conceptual Business Process Structuring by Extracting Knowledge from Natural Language Texts // Proceedings of the 2014 International Conference on Artificial Intelligence (ICAI 2014). – Vol. I, WORLDCOMP'14, July 21–24, 2014. – Las Vegas, Nevada, USA: CSREA Press, 2014. – Pp. 82–87.

18. Charnine, M.M., Kuznetsov, I.P., Kozerenko, E.B. Semantic Navigator for Internet Search // Proceeding of International Conference on Machine Learning, 27–30, 2005. – Las Vegas, USA: CSREA Press, 2005. – Pp. 60–65.

19. Charnine, M., Protasov, V. Optimal Automated Method for Collaborative Development of University Curricula // Proceedings of the 2013 International Conference on Artificial Intelligence (ICAI 2013). – V. I. WORLDCOMP'13. July 22–25, 2013. – Las Vegas, Nevada, USA: CSREA Press, 2013. – Pp. 96–100.

20. Bacon, E., Hagel, G., Charnine, M., Foggie, R., Kirk, B., Schagaev, I., Kravtsov, G. WEDUCA: Web-enhanced design of university curricula // Proceedings of the FECS'13: The International Conference on Frontiers in Education: Computer Science and Computer Engineering, July 22–25, 2013. – Las Vegas, Nevada, USA: CSREA Press, 2013. – Pp. 288–294.

21. Золотарев О.В. Методы выделения процессов, объектов, отношений из текстов естественного языка // Проблемы безопасности российского общества. – Смоленск : Свиток, 2014. – № 3–4. – С. 276–283.

22. Золотарев О.В. Инновационные решения в формировании функциональной структуры предметной области // Вестник Российского

нового университета. – 2013. – Выпуск 4. – С. 82–84.

23. Золотарев О.В. Методы и инструменты моделирования предметной области // Цивилизация знаний: проблемы и перспективы социальных коммуникаций : труды Тринадцатой Международной научной конференции, г. Москва, 20–21 апреля 2012 г. – М. : РосНОУ, 2012.

24. Золотарев О.В. Новые подходы в построении функциональной структуры предметной области // Сб. трудов по материалам конференции «20 лет постсоветской России: кризисные явления и механизмы модернизации». – Екатеринбург : Гуманитарный университет, 2011.

25. Золотарев О.В. Формализация знаний о предметной области на основе анализа естественно-языковых структур // Цивилизация знаний: проблема человека в науке XXI века : труды Двенадцатой Международной научной конференции, г. Москва, 22–23 апреля 2011 г. – М. : РосНОУ, 2011.

26. Золотарев О.В. Средства анализа информации в системах, основанных на семантических сетях // Цивилизация знаний: проблемы модернизации России : труды Одиннадцатой Международной научной конференции, Москва, 23–24 апреля 2010 г. – М. : РосНОУ, 2010.

27. Золотарев О.В., Козеренко Е.Б., Шарнин М.М. Проведение аналитической разведки на основе анализа неструктурированной информации из различных источников, включая Интернет и средства массовой информации // Вестник Российского нового университета. Серия «Сложные системы: модели, анализ и управление». – 2015. – Выпуск 1. – С. 49–54.

28. Шарнин М.М., Золотарев О.В., Солин Н.В. Извлечение и обработка знаний из неструктурированных текстов деловой сферы и социальных сетей // Труды IV Международной научно-практической конференции «Социальный компьютинг: основы, технологии развития, социально-гуманитарные эффекты», Москва, МПГУ, 22–24 октября 2015 г. – М., 2015.

29. Zolotarev, O.V., Charnine, M.M., Matskevich, A.G., Kuznetsov, K.I. Business Intelligence Processing on the Base of Unstructured Information Analysis from Different Sources Including Mass Media and Internet // Proceedings of the 2015 International Conference on Artificial Intelligence (ICAI 2015). – Vol. I, WORLDCOMP'15, July 27–30, 2015. – Las Vegas, Nevada, USA, 2015. – V. I. – Pp. 295–299.

30. Золотарев О.В., Шарнин М.М. Методы извлечения знаний из текстов естественного языка и построение моделей бизнес-процессов

на основе выделения процессов, объектов, их связей и характеристик // Труды XIX Международной конференции СРТ2014. Ларнака, Кипр, 12–18 мая 2014. – М. : Изд-во Института физико-технической информатики (ИФТИ), 2015. – С. 92–98.

31. Золотарев О.В., Козеренко Е.Б., Шарнин М.М. Принципы построения моделей бизнес-процессов предметной области на основе обработки текстов естественного языка // Вестник Российского нового университета. – 2014. – Выпуск 4. – С. 82–88.