Поиск и рубрицирование ссылок на цитируемые публикации в электронных библиотеках полнотекстовых описаний изобретений

© В.А. Хаванскоv © С.К. Шубников
Институт проблем информатики Российской академии наук
Москва
chavanskov@yandex.ru sergeysh50@yandex.ru

Аннотация

В данной работе рассмотрены вопросы поиска и рубрицирования библиографических ссылок в полнотекстовых описаниях изобретений с целью исследования тематических взаимосвязей науки и технологий. На базе созданного в Институте проблем информатики РАН макета аналитической информационной системы был проведен анализ патентной информации по классу G06 Международной патентной классификации (Обработка данных; Вычисления; Счет), опубликованной Роспатентом в 2000–2012 гг.

Работа выполнена при частичной поддержке РГНФ (грант № 12-02-12019в).

1. Введение

Исследования взаимосвязей науки и технологий в последние десятилетия приобретает все большее значение в свете стратегического инновационного планирования развития науки и технологий. Вопрос заключается в поиске инструментов, которые могли бы с достаточной степенью достоверности указывать на направления фундаментальных научных исследований, влияющие на определенные технологии. Один из таких инструментов установления связей между технологическими областями и направлениями научных исследований предполагает использование информационных ресурсов Роспатента, представленных в электронном виде и доступных для автоматизированной обработки.

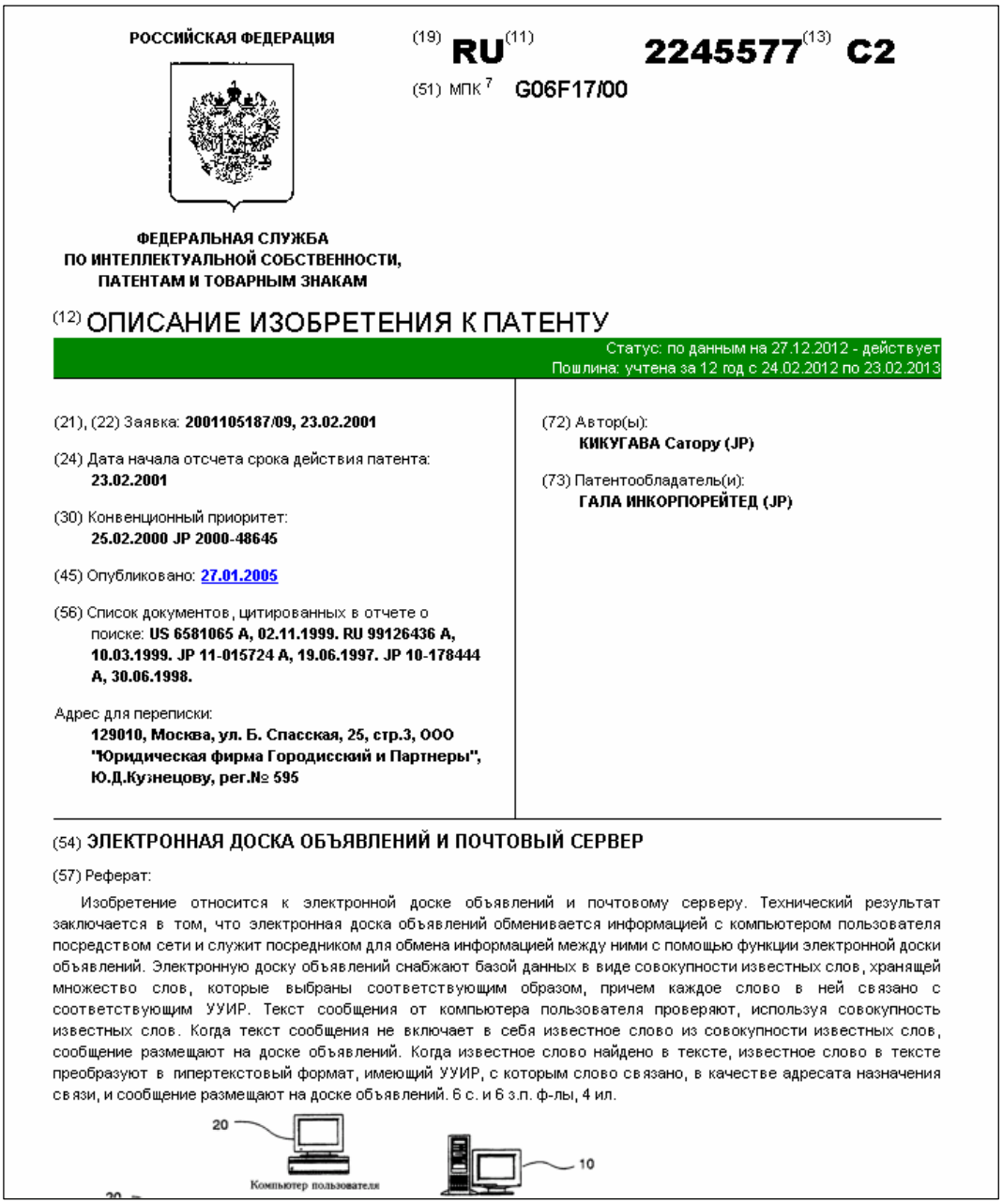
В соответствии со Страсбургским соглашением от 24 марта 1971 года о Международной патентной классификации (МПК) компетентные органы стран-участниц Союза по МПК при классифицировании патентных документов должны указывать «полные индексы МПК, присвоенные изобретению, описанному в документе» (Статья 4, пункт 3). Это означает, что публикуемые в Роспатенте сведения о выданных патентах на изобретения содержат индексы (МПК), которые можно использовать для описания исследуемых групп технологий [[4]]. С другой стороны, имеются полнотекстовые описания изобретений, представляющие собой неструктурированные тексты, в которых при изложении сути изобретения авторы ссылаются на публикации в научных изданиях. Таким образом, библиографические ссылки на публикации в описаниях изобретений, привязанные к одной или нескольким рубрикам направлений научных исследований, можно сопоставить с индексами МПК (технологиями).

При проведении анализа объем отобранных полнотекстовых описаний патентов на изобретение может достигать нескольких сотен тысяч. Например, из работы [[2]], в которой описывается процесс обработки массива из 656 695 патентов на изобретения, выданных Патентным ведомством США. Для установления взаимосвязей между кодами МПК и кодами рубрик научных направлений исследований из описаний изобретений были выделены 1 147 160 ссылок на цитируемые публикации (ссылки на патенты были исключены). Затем из них для дальнейшей обработки были отобраны только те ссылки на журнальные статьи, для которых удалось идентифицировать название журнала и соотнести его с нормативным списком названий журналов, в котором каждому названию присвоена одна или несколько рубрик научных направлений исследований. Таким образом, было отобрано 106 636 ссылок, то есть менее 10% от выделенных ссылок на цитируемые публикации.

При реализации подобной методологии для исследования отечественных описаний патентов на изобретения возникает ряд дополнительных особенностей указанных в работе [[9]], а именно:

Отсутствие в «Административном регламенте исполнения Роспатентом приема заявок на изобретение, их рассмотрения и экспертизы» [[3]] требований к структурированию ссылок на цитируемые публикации (см. п. 10.11 (12) Регламента «*Библиографические данные источников информации указываются таким образом, чтобы источник информации мог быть по ним обнаружен*»).

Труды 16-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2014, Дубна, Россия, 13–16 октября 2014 г.



РОССИЙСКАЯ ФЕДЕРАЦИЯ

(19) **RU**(11) **2245577**(13) **C2**

(51) МПК [7] G06F17/00

ФЕДЕРАЛЬНАЯ СЛУЖБА
ПО ИНТЕЛЛЕКТУАЛЬНОЙ СОБСТВЕННОСТИ,
ПАТЕНТАМ И ТОВАРНЫМ ЗНАКАМ

(12) ОПИСАНИЕ ИЗОБРЕТЕНИЯ К ПАТЕНТУ

Статус: по данным на 27.12.2012 - действует
Пошлина: учтена за 12 год с 24.02.2012 по 23.02.2013

(21), (22) Заявка: **2001105187/09, 23.02.2001**

(24) Дата начала отсчета срока действия патента:
23.02.2001

(30) Конвенционный приоритет:
25.02.2000 JP 2000-48645

(45) Опубликовано: **27.01.2005**

(56) Список документов, цитированных в отчете о поиске: **US 6581065 A, 02.11.1999. RU 99126436 A, 10.03.1999. JP 11-015724 A, 19.06.1997. JP 10-178444 A, 30.06.1998.**

Адрес для переписки:
129010, Москва, ул. Б. Спасская, 25, стр.3, ООО "Юридическая фирма Городисский и Партнеры", Ю.Д.Кузнецову, рег.№ 595

(72) Автор(ы):
КИКУГАВА Сатору (JP)

(73) Патентообладатель(и):
ГАЛА ИНКОРПОРЕЙТЕД (JP)

(54) ЭЛЕКТРОННАЯ ДОСКА ОБЪЯВЛЕНИЙ И ПОЧТОВЫЙ СЕРВЕР

(57) Реферат:

Изобретение относится к электронной доске объявлений и почтовому серверу. Технический результат заключается в том, что электронная доска объявлений обменивается информацией с компьютером пользователя посредством сети и служит посредником для обмена информацией между ними с помощью функции электронной доски объявлений. Электронную доску объявлений снабжают базой данных в виде совокупности известных слов, хранящей множество слов, которые выбраны соответствующим образом, причем каждое слово в ней связано с соответствующим УУИР. Текст сообщения от компьютера пользователя проверяют, используя совокупность известных слов. Когда текст сообщения не включает в себя известное слово из совокупности известных слов, сообщение размещают на доске объявлений. Когда известное слово найдено в тексте, известное слово в тексте преобразуют в гипертекстовый формат, имеющий УУИР, с которым слово связано, в качестве адресата назначения связи, и сообщение размещают на доске объявлений. 6 с. и 6 з.п. ф-лы, 4 ил.

Рисунок 1. Фрагмент полнотекстового описание изобретения РОСПАТЕНТА

Отсутствие в опубликованных электронных версиях полнотекстовых описаний изобретений групп меток, выделяющих ссылки на цитируемые публикации согласно рекомендациям стандарта ВОИС ST.14 [[12]].

Отсутствие списка нормализованных и сокращенных названий журналов, используемых в ссылках на цитируемые публикации.

Таким образом, при исследовании тематических взаимосвязей технологий и направлений научных исследований возникает задача анализа десятков и сотен тысяч полнотекстовых описаний изобретений и поиска в тексте на естественном языке (ЕЯ) ссылок на публикации, с последующей их структуризацией и привязкой ссылок к рубрикам направлений научных исследований. Как следствие, возникает задача автоматизация данного процесса. При этом необходимо учитывать, что библиографическая информация является структурированным информационным объектом, состоящим из нескольких полей, который может размещаться внутри неструктурированного текста, а разные поля библиографической информации могут быть в общем случае на разных языках.

Целью данной работы является рассмотрение технических вопросов решения задачи автоматизации обработки массивов описаний патентов, а также:

Выбор механизма поиска и выделения слабоструктурированного текста (ссылки на цитирование публикации) в произвольном тексте на ЕЯ;

Выбор механизма рубрицирования выделенных ссылок публикаций статей в периодических

изданиях по заданным классификаторам направлений научных исследований.

Дальнейшее изложение будет вестись на основании результатов обработки с использованием макета аналитической информационной системы [[10]] 6665 патентов, опубликованных Роспатентом в 2000–2012 гг. и относящихся к классу G06 Международной патентной классификации (Обработка данных; Вычисления; Счет).

2. Поиск и выделение ссылок на цитируемые публикации

В виду указанного выше отсутствия требований к оформлению ссылок на цитируемые публикации и указаний к месту их размещения их поиск в тексте полнотекстового описания становится не тривиальной задачей.

В ряде работ [[1], [5], [6]] уже предложены механизмы поиска ссылок на цитируемые публикации в текстах статей и библиографических данных патентов. Общим подходом для них является то, что исходный текст представляется как неким образом структурированный текст – научная статья, которая имеет название, авторов, пристатейный список литературы [5, 6] или текст описания патента [1] (**Ошибка! Источник ссылки не найден.**), размеченный определенными метками. Проведенный анализ возможности использования данных методов дал следующие результаты.

В отличие от принятой в научной среде практики размещения списка используемой литературы в конце научной публикации, в полнотекстовых описаниях изобретений ссылки на цитируемую публикацию могут встретиться в любом месте. Использование метода, рассматривающего полнотекстовое описание патента как научную статью и выделяющий в тексте ключевые слова (литература, список литературы и пр.), позволяет обработать менее 14% патентов от их общего числа.

Использование метода анализа текста, обозначенного заданными метками (в частности, метка с кодом 56 – Список документов, цитированных в отчете о поиске), показывает, что в данных полях содержится менее 12% не патентных ссылок на цитируемые публикации.

Таким образом, использование этих методов значительно снижает ресурсную базу для вычисления индикаторов взаимосвязи научных исследований и технологий.

В тоже время анализ текстов описаний изобретения показывает, что подавляющая часть ссылок на цитируемые публикации в целом соответствует общепринятым правилам составления библиографических ссылок, хотя в них и существует определенные различия.

В работе [[8]] подробно рассматривается механизм использования для автоматизации процесса поиска ссылок на цитируемые публикации настраиваемых шаблонов объекта поиска.

Суть данного механизма заключается в следующем:

Предложена структура объекта поиска – ссылки на цитируемую публикацию, которая состоит из выделенных библиографических элементов и выгладит следующим образом:

[автор{S_1}] [название публикации] [{S_2}название источника] {S_3}атрибуты публикации

(наличие квадратных скобок говорит о необязательности присутствия данного элемента в искомом фрагменте текста). Данная структура объекта поиска опирается на требования соответствующего стандарта [[7]], описывающего правила составления библиографических ссылок.

В качестве описания библиографических элементов структуры поиска используются регулярные выражения [[11]].

Регулярные выражения предоставляют мощный, гибкий и эффективный метод обработки текста. Обширные возможности сопоставления шаблонов, предоставляемые регулярными выражениями, позволяют быстро анализировать большие объемы текста, отыскивая в них определенные символьные шаблоны, проверять текст на соответствие определенным заранее шаблонам (например, формату адреса электронной почты) и добавлять извлеченные строки в коллекцию для формирования отчетов.

Например, шаблон поиска библиографического элемента ***[автор]*** может быть представлен следующим образом:

```
\p{Lu}\p{Ll}?[.][\s]?(\p{Lu}[.])?[\
s]?\p{Lu}+(\p{Ll}+)?
```

В данном случае:

- \p{ *name* } соответствует любому одиночному символу в общей категории Юникода, указанном в параметре *name*;

- Lu,Ll – прописная буква, строчная буква;

- [*character_group*] соответствует любому одиночному символу, входящему в *character_group*;

- квантор ? указывает, что предыдущий элемент не обязательный и может присутствовать не более одного раза;

- квантор + указывает, что предыдущий элемент обязательный и должен присутствовать не менее одного раза.

Таким образом, приведенный пример шаблона элемента направлен на выделение в тексте следующих вариантов написания автора публикации:

Таблица 1

	[автор]	*[название публикации]*	*[название источника]*	*атрибуты публикации*
1.	(\p{Lu}\p{Ll}?[.][ ]?(\p{Lu}[.])?[ ]?\p{Lu}+\p{Ll}+([.,])+)+	([\D]+){1}([\w\s,.]+		[\d{0,3}]?)\d{4}
2.	(\p{Lu}?\p{Ll}?\p{Lu}+(\p{Ll}+)?[ ](\p{Lu}\p{Ll}?[.][ ]?)?(\p{Lu}[.])[,]?)+	([\D]+){1}([\w\s,.]+		[\d{0,3}]?)\d{4}
3.			['«"]\D+['»"](.)+[\s.,]	(\d{4}[).,]){1}

- И.О. Фамилия;
- Им.О. Фамилия;
- И. О. Фамилия;
- И. Фамилия;
- Им. О. Фамилия

и другие варианты, когда инициалы имени и отчества автора указаны перед его фамилией.

Разработан алгоритм поиска, основанный на приложении к тексту на ЕЯ совокупности заданных шаблонов поиска и интеграции результатов поиска каждого из них. При этом исходный текст представляется сплошным массивом абзацев без деления на структурные элементы текста (ячейки таблицы, ссылки, сноски и пр.).

Данный механизм был использован при анализе массива патентов на изобретения по коду МПК G06 за период 2000–2012 годы. При программной обработке в течении 5 часов[1] из этого массива было выделено 2758 патентов, в которых найдено 8847 фрагментов текста соответствующих критериям, заданных шаблонами поиска. Список и структура шаблонов поиска приведена в таблице 1.

Как видно из таблицы, в первом и втором шаблонах начало искомого фрагмента задано элементом структуры ***[автор]***, а в третьем – элементом ***[название источника]***. Последнее обстоятельство отражает ситуацию некорректности требований к предоставлению заявителем сведений об источниках информации (п. 10.11 (12) Регламента [**Ошибка! Источник ссылки не найден.**]).

После анализа выделенных фрагментов текста обнаружилось порядка 30% случаев, когда используемые шаблоны выделили фрагмент текста полнотекстового описания, не являющийся ссылкой на цитируемую публикацию, но отвечающий критериям поиска, так называемый «шумовой эффект» (**Ошибка! Источник ссылки не найден.** патент 2259020). Данное обстоятельство не влияет на конечный результат, так как при дальнейшей структуризации ссылки исключается из рассмотрения.

В тоже время, среди полнотекстовых описаний, в которых программой не обнаружено ссылок на цитируемые публикации, при визуальном просмотре обнаружены пропущенные программой ссылки на цитируемые публикации (**Ошибка! Источник ссылки не найден.**). Как видно из рисунка, список используемых для поиска шаблонов необходимо дополнить несколькими шаблонами:

- начало искомого текста задавать элементом ***[название публикации]*** (1-я публикация);
- в элементе ***[название источника]*** требуется указывать наличие строчных букв (2-я и 3-я публикации).

Данная процедура (добавление и редактирование шаблонов поиска) заложена в используемую программу. После добавления и/или редактирования шаблонов возможен повторный анализ всего массива полнотекстовых описаний патентов.

3. Рубрицирование ссылок на цитируемые публикации

Второй задачей используемого механизма поиска и структуризации ссылок цитирования публикаций является задача рубрицирования найденной публикации в контексте используемого классификатора рубрик научных направлений исследований. При этом, следует отметить, что для рубрицирования из общего массива найденных ссылок на цитируемые публикации отбираются только ссылки на опубликованные статьи в научных периодических изданиях и в трудах конференций. Данный отбор осуществляется с использованием той же приведенной ранее структуры описания ссылки на цитируемую публикацию.

Само рубрицирование заключается в реализации ряда последовательных этапов:

Выделение в ссылке на публикацию статьи названия периодического издания или конференции.

Поиск в нормализованной базе данных найденного названия периодического издания (конференции).

Присвоение анализируемой ссылке на публикацию кодов рубрик направлений научных исследований, соотнесенных с данным научным периодическим изданием (данной конференцией).

Рассмотрим более подробно каждый из перечисленных этапов.

[1] Без учета времени загрузки описаний из открытого электронного реестра полнотекстовых описаний патентов на изобретения Федерального института промышленной собственности (ФИПС).

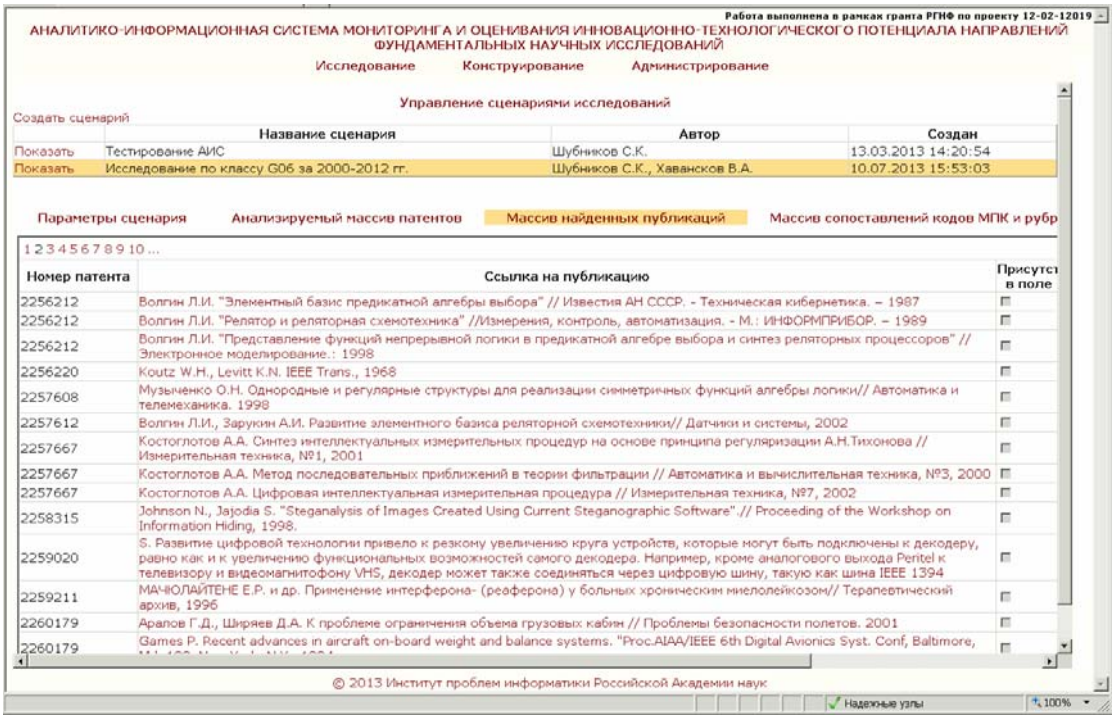


Рисунок 2. Список найденных публикаций

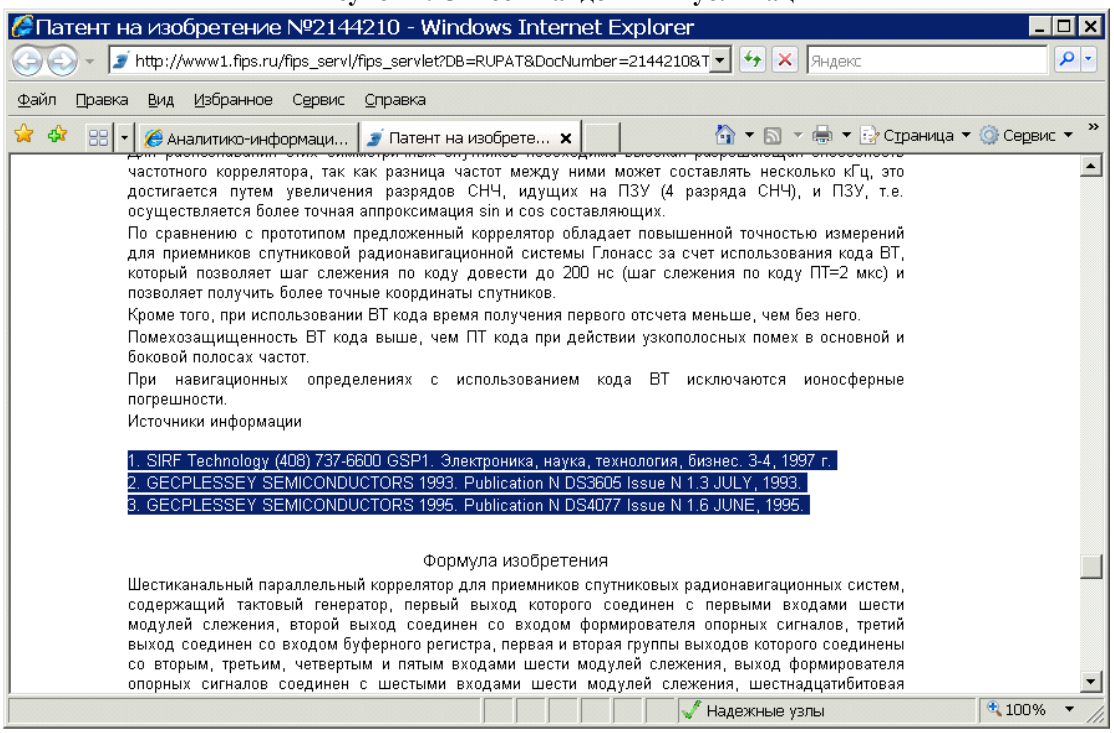


Рисунок 3. Полнотекстовое описание патента 2144210 с пропущенными ссылками на цитируемые публикации

Выделение названия периодического издания или конференции.

Для выделения в анализируемом фрагменте текста периодического издания или конференции, в котором опубликована статья, используются шаблоны описания разделителей между библиографическими элементами структуры описания ссылки на цитируемую публикацию, а именно ***{S₂}*** (разделитель – издание) и ***{S₃}*** (разделитель – атрибуты). Разделитель ***{S₃}***, который достаточно очевиден и, как показала практика, вопросов не возникает, представлен следующими шаблонами описаний:

([,.] - \w+[.]:)

([,.] \d{4})

([,.] - \d{4})

([,.] [(V.)(v.)(vol.)(Vol.)(VOL.)])

([,.] \d+[,.])

([,.] [№N][\d]?)

([,.] [Pp]{1,2}[.])

([,.] [(Т.)(т.)(том)(Том)(ТОМ)])

([,.] [(В.)(в.)(вып.)(Вып.)(ВЫП.)])

Данный набор шаблонов максимально полно отражает возможные варианты представления начала библиографического элемента ***атрибуты публикации***.

С разделителем ***{S_2}*** дело обстоит несколько иначе. Если опираться на действующий государственный стандарт «Библиографическая запись. Библиографическое описание» [**Ошибка! Источник ссылки не найден.**], то таковым разделителем является двойной слеш “//”. Как показал анализ обработанного массива найденных ссылок на цитируемые публикации статей в периодических изданиях, данный шаблон разделителя используется только в ~ 60% из них. Поэтому, к данному шаблону разделителя, в качестве разделителей были добавлены описания разделителя в виде ключевых слов:

//

Ж[\w-]+л[ ]

IEEE

[(]\d{4}[)]

Последний шаблон использует год публикации заключенный в круглые скобки. Подобное отделение название статьи от названия периодического издания встречается в научных публикациях в некоторых англоязычных периодических изданиях.

При этом следует отметить, что в отличие от процедуры поиска ссылок на цитированные публикации, использующий единый алгоритм обработки выделенного текста фрагмента для всех используемых шаблонов поиска, в данном случае для каждого шаблона, использующего в качестве разделителя ключевое слово, действует отдельный алгоритм выделения названия периодического издания. Это вызвано тем, что в некоторых случаях ключевое слово может являться частью названия периодического издания (**Ошибка! Источник ссылки не найден.** патент 2277257), а в других – нет (**Ошибка! Источник ссылки не найден.** патент 2277261 вторая запись). Но данное обстоятельство может приводить и к ложным выделениям названия периодического издания (**Ошибка! Источник ссылки не найден.** патент 2277261 первая и третья записи). В этом случае в работу включается оператор, задачей которого является устранение подобных ложных записей.

Поиск в нормализованной базе данных найденного названия периодического издания (конференции).

На данном этапе используется нормализованная база данных названий периодических изданий и конференций. В ней, кроме полного официального названия периодического издания или конференции, указывается и иные возможные названия (не полные, сокращенные и пр.) – «псевдонимы». Именно по псевдонимам ведется автоматический поиск периодического издания после выделения названия на предыдущем этапе. В случае отсутствия искомого названия в базе данных в работу включается оператор, задачей которого является либо привязка искомого названия к существующему периодическому изданию и добавление его в список псевдонимов, либо поиск и ввод в базу данных нового полного официального названия периодического издания.

Каждая нормализованная запись названия периодического издания содержит перечень рубрик заданных классификаторов направлений научных исследований.

Присвоение анализируемой ссылке на публикацию кодов рубрик направлений научных исследований, соотнесенных с данным научным периодическим изданием.

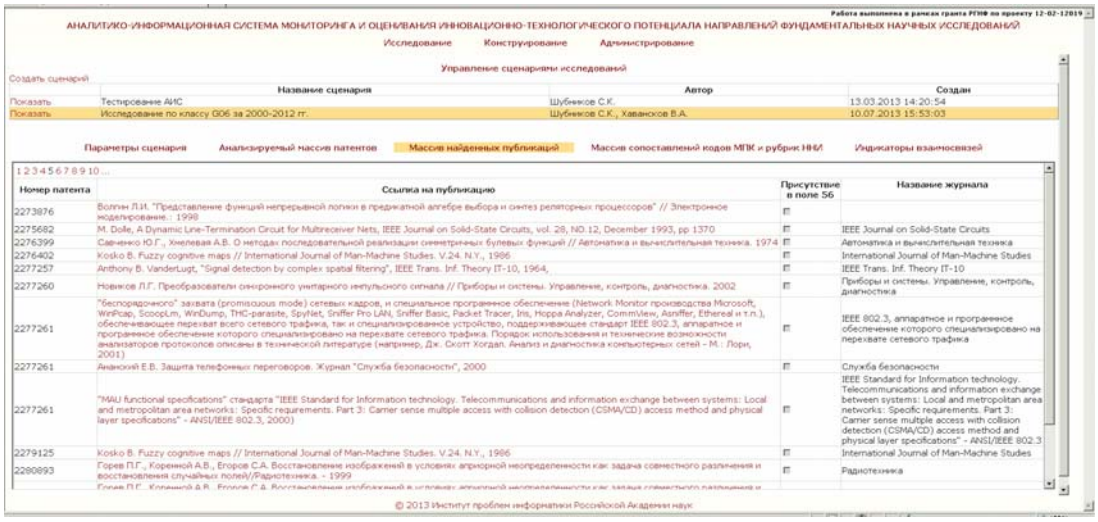


Рисунок 4. Список найденных публикаций и выделенных названий периодических изданий

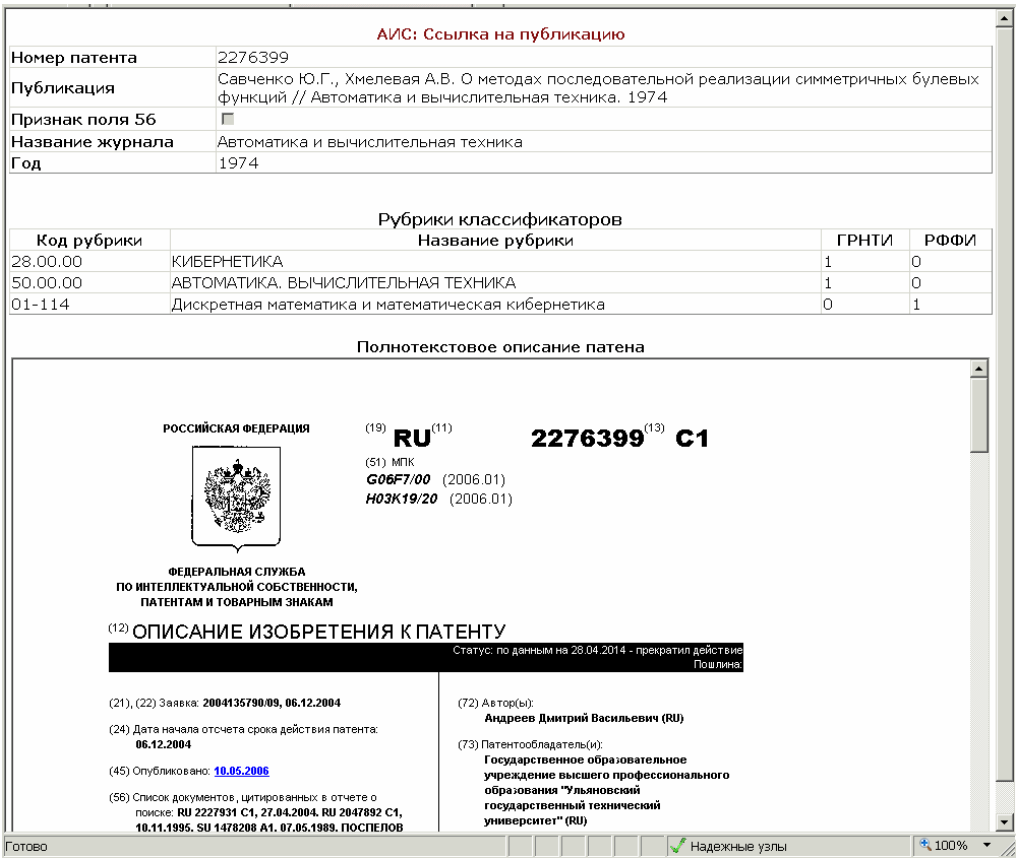


Рисунок 5. Ссылка на публикацию с кодами рубрик заданных классификаторов

Таблица 2

	Всего выделенных ссылок на публикацию статей	**Ошибочно выделенных ссылок**	**«Слипшихся» ссылок**	**Выделенных ссылок с лишним текстом**	**Правильно выделенных ссылок**
		(в %)	**(в %)**	**(в %)**	**(в %)**
Периодические издания	365	3,01	5,75	9,86	70,68
Конференции	147	30,61	34,69	38,10	38,10
Всего	512	10,94	14,06	17,97	61,33

При привязке ссылки на цитируемую публикацию к нормализованному названию периодического издания коды рубрик периодического издания копируются в запись ссылки на цитируемую публикацию.

В результате выполнения всех этапов для каждой ссылки на цитируемую публикацию устанавливается привязка к кодам рубрик заданных классификаторов направлений научных исследований (**Ошибка! Источник ссылки не найден.**).

Таким образом, в результате обработки всего массива полнотекстовых описаний патентов на изобретения 512 выделенных фрагментов текста были автоматически идентифицированы как ссылки на публикации статей в периодических изданиях и трудах конференций. Последующий анализ этих записей дал результаты, приведенные в таблице 2.

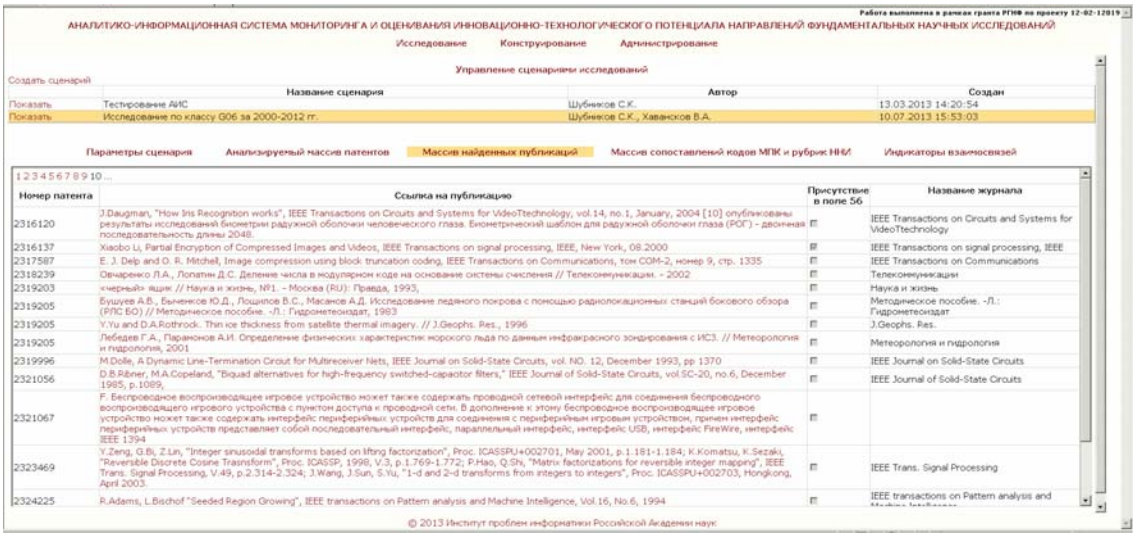


Рисунок 6. Пример слипшихся ссылок и ссылок с лишним текстом

Таблица 3

	Объем просматриваемого текста			**Трудозатраты**	
				4 у-п.л./чел.	**2 у-п.л./чел.**
	в знаках	**в у-п.л.**	**в станд. стр.**	**в чел./дн.**	**в чел./дн.**
Все выделенные фрагменты	3 069 293	76,73	1 918,31	19,18	38,37
Фрагменты публикаций статей	100 181	2,50	62,61	0,63	1,25
Рубрицированные публикации статей	50 934	1,27	31,83	0,32	0,64

К ошибочно выделенным относятся записи, которые могут быть представлены как, например, на **Ошибка! Источник ссылки не найден.** для патента 2259020, «слипшиеся» ссылки могут выглядеть как на рисунке 6 для патента 2323469, а выделенные ссылки с лишним текстом – как на том же рисунке для патента 2316120.

4. Заключение

Основным результатом данной работы можно считать проведенный впервые в России эксперимент по анализу полнотекстовых описаний изобретений, опубликованных Роспатентом в период с 2000 по 2012 гг. и относящихся к классу G06 МПК (Обработка данных; Вычисления; Счет) с целью выделения и рубрицирования библиографических ссылок на научные публикации для определения научных направлений связанных с данным классом. Был исследован относительно небольшой массив отобранных изобретений, однако предлагаемая технология может быть использована как на длительных временных интервалов выборки, так и для более представительных типов технологий.

Как можно оценить предлагаемую технологию?

Если представить весь объем текста для всех 8847 выделенных фрагментов как сплошной текст, то получим результаты, приведенные в таблице 3.

Иными словами, для обработки всего массива выделенных фрагментов текста, в зависимости от стадии просмотра и анализа выделенного текста, а также от поставленных целей и заданной производительности, трудозатраты составят от 0,32 до 38,37 чел./дней. Напомним, для поиска и анализа всего массива полнотекстовых описаний изобретений (6665) при производительности 4 у-п.л./день потребуется 1675 человеко/дней.

Также, следует отметить, что обработке подвергается весь массив описаний патентов на изобретения, а не только те, которые имеют специально оговоренные словарные метки (список литературы, список публикаций, литература и пр.) и выделяют в тексте пристатейные списки цитируемых публикаций.

Дальнейшим развитием данного подхода может являться лингвистический анализ контекста упоминания конкретной ссылки на цитируемую публикацию с вычислением некоторой меры ее веса в списке всех ссылок цитирования, используемых для анализа взаимосвязей науки и технологий.

Литература

[1] INCENTIM (2003), Linking Science to Technology – Bibliographic References in Patents, Project Report (http://www.cordis.lu/indicators).

[2] Verbeek A., Debackere K., Luwel M., Andries P., Zimmermann E., Deleus D. Linking science to technology: Using bibliographic references in patents to build linkage schemes // Scientometrics, 2002. Vol. 54, No. 3. P. 399–420.

[3] Административный регламент исполнения Роспатентом приема заявок на изобретение, их рассмотрения и экспертизы. ФИПС, 2008 [электронный ресурс]. – http://www1.fips.ru/wps/wcm/connect/content_ru/ru/documents/russian_laws/order_minobr/administrative_regulations/test_8/.

[4] Архипова М.Ю., Зацман И.М., Шульга С.Ю. Индикаторы патентной активности в сфере информационно-коммуникационных технологий и методика их вычисления // Экономика, статистика и информатика. Вестник УМО. 2010. №4. С. 93–104.

[5] Васильев А., Козлов Д., Самусев С., Шамина О. Извлечение метаинформации и библиографических ссылок из текстов русскоязычных научных статей // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды Девятой Всероссийской научной конференции RCDL'2007. – Переславль: Университет города Переславля, 2007. С. 175–184.

[6] Васильев А., Козлов Д., Самусев С., Шамина О. Создание электронной библиотеки русскоязычных научных статей. // Сборник работ стипендиатов гранта «Интернет-математика 2007». – Екатеринбург: Изд-во Уральского ун-та, 2007. – С. 37–45.

[7] ГОСТ 7.1–2003 Библиографическая запись. Библиографическое описание. Общие требования и правила составления. [электронный ресурс]. – http://lib.usfeu.ru/index.php/gost-7-1-2003.

[8] Зацман И.М., Хаванcков В. А., Шубников С.К. Метод извлечения библиографической информации из полнотекстовых описаний изобретений // Информатика и ее применения. – 2013. – Т. 7, вып. 4. – С. 52–65.

[9] Зацман И.М., Шубников С.К. Принципы обработки информационных ресурсов для оценки инновационного потенциала направлений научных исследований // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды Девятой Всероссийской научной конференции RCDL'2007. – Переславль: Университет города Переславля, 2007. – С. 35–44.

[10] Минин В.А., Зацман И.М., Хаванcков В.А., Шубников С.К. Архитектурные решения для систем вычисления индикаторов тематических взаимосвязей науки и технологий // Системы и средства информатики. – 2013. – Т. 23, № 2. – С. 260–283.

[11] Регулярные выражения в .NET Framework // MSDN. Библиотека – http://msdn.microsoft.com/ru-ru/library/hs600312.aspx

[12] Стандарт ВОИС ST.14 «Рекомендации по включению ссылок, цитируемых в патентных документах» [электронный ресурс]. – http://www.rupto.ru/rupto/nfile/52b8dfc1-1049-11e1-a520-9c8e9921fb2c/03_14_01.pdf.

Identification and Classification of Citation References in Digital Libraries of Full-text Patent Descriptions

Valeriy A. Khavanskov, Sergey K. Shubnikov

This paper considers the problems of identification and classifications of bibliographic references in full-text patent descriptions for the purposes of investigating relationships between science and technology fields. The paper also reports on analysis of patents classified in G06 class (*Computing, Calculating, Counting* according to the International Patent Classification) published by Rospatent agency between 2000 and 2012. The analysis was completed with the use of the draft analytical information system created in the Institute for Informatics Problems of the Russian Academy of Sciences.