Принципы обработки информационных ресурсов для оценки инновационного потенциала направлений научных исследований ♣

© И.М. Зацман

Институт проблем информатики РАН
im@a170.ipi.ac.ru

С.К. Шубников

Институт проблем информатики РАН
serg@a170.ipi.ac.ru

Аннотация

Доклад посвящен проблеме оценки инновационного потенциала направлений научных исследований с использованием информационных полнотекстовых ресурсов патентных электронных библиотек, доступ к которым предоставляется Роспатентом. Анализируется структура и наполнение информационных полнотекстовых ресурсов Роспатента. Показано, что современные исследования взаимосвязей технологических разработок с результатами научных исследований и вычисление индикаторов для количественной оценки этих взаимосвязей, включая оценку инновационного потенциала научных результатов, основаны на анализе и обработке массивов описаний изобретений к патентам, содержащих ссылки на научные публикации. Однако задача вычисления индикаторов предъявляет ряд требований к методологии обработки информационных ресурсов, а также к степени детализации схем ресурсов патентных электронных библиотек. Одно из этих требований заключается в дополнительной структуризации ссылок на цитируемые документы в полнотекстовых информационных ресурсах патентных электронных библиотек и баз данных.

1 Введение

Рассматриваемые в докладе методы мониторинга, анализа и индикаторной оценки инновационного потенциала направлений научных исследований (далее по тексту – методы оценки) основаны на использовании новой категории моделей. Эти модели были определены в работе [1] и их было предложено называть «полидоменными моделями».

Труды 9ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2007, Переславль-Залесский, Россия, 2007.

Полидоменные модели, состоят из нескольких видов компонентов, включая информационный, математический, алгоритмический и лексико-семантический компоненты. В работе [2] были рассмотрены несколько частных случаев наборов компонентов полидоменной модели. При этом каждый из рассмотренных частных случаев модели включал информационный компонент, с помощью которого специфицируются структура и наполнение информационных ресурсов. Примеры математического и алгоритмического компонентов полидоменных моделей рассматривались в работах [1, 2]. Задачи и функции лексико-семантического компонента рассматривались в работе [4].

Полидоменные модели были выбраны в качестве основы для разработки методов оценки в силу следующих причин.

Во-первых, по определению полидоменные модели позволяют эксплицировать широкий спектр связей и отношений (структурные, информационные, математические, алгоритмические и семантические), которые необходимо учитывать при определении значений индикаторов на основе накопленных информационных ресурсов систем мониторинга, анализа и индикаторной оценки (далее по тексту – системы оценки), а также область определения индикаторов. Эти модели позволяют эксплицировать связи между конкретными и абстрактными информационными объектами (например, между структурированными текстами и математическими формулами) [1, 2].

Во-вторых, полидоменные модели позволяют фиксировать изменение и/или пополнение во времени информационных ресурсов, а также эксплицировать эволюцию их семантики с помощью лексико-семантического компонента, включающего тезаурус [3].

В-третьих, полидоменные модели позволяют представить результаты эскизного проектирования системы оценки как «долговечной системы». Так как жизненный цикл систем оценки намного превышает период смены поколений используемых программно-аппаратных средств, то необходимо инвариантное по отношению к смене поколений представление результатов эскизного проектирования. Инвариантное представление

результатов дает возможность иметь описание функциональности, структуры системы и всего спектра связей и отношений между ее структурными компонентами, которое не зависит от используемых программно-аппаратных средств.

Инвариантное представление результатов эскизного проектирования систем оценки является основой обеспечения их эволюции в условиях многократной смены поколений программно-аппаратных средств. Иначе говоря, инвариантное представление результатов эскизного проектирования способно, с одной стороны, сохранять в течение долгого периода времени описание первоначальной функциональности, структуры системы, связей и отношений между структурными компонентами и, с другой стороны, позволяет модифицировать их в случае необходимости, эксплицируя эволюцию систем оценки в виде последовательности версий их полидоменных моделей.

Настоящий доклад посвящен вопросам проектирования информационного компонента полидоменной модели системы оценки. Отличие этого доклада от ранее опубликованных работ заключается в том, что для оценки инновационного потенциала предлагается использовать наследуемые и уже имеющиеся полнотекстовые информационные ресурсы патентных электронных библиотек и баз данных. При этом разработка самих индикаторов и вопросы создания лексико-семантического компонента модели здесь не рассматриваются, так как этим вопросам посвящен отдельный доклад [10].

Отметим, что для оценки инновационного потенциала необходимы именно полнотекстовые описания изобретений к патентам, включающие также краткие библиографические описания цитируемых документов (автор(ы) документа, его название, источник публикации, номер документа для случая объектов интеллектуальной собственности и т.п.).

Однако доступ к полнотекстовым информационным ресурсам патентных электронных библиотек является необходимым, но не достаточным условием. Причина заключается в том, что схемы имеющихся полнотекстовых информационных ресурсов Роспатента не могут обеспечить решение задач оценки инновационного потенциала, так как научные публикации и другие документы, которые цитируются в описаниях изобретений, в настоящее время недостаточно структурированы.

Поэтому предлагается дополнительно детализировать схемы полнотекстовых описаний изобретений в той их части, где цитируются научные публикации (в первую очередь, журнальные статьи и материалы конференций), в соответствии с международными стандартами. Затем предлагается провести дополнительное структурирование электронных форм описаний изобретений, подготовленных Роспатентом, в соответствии с детализированными схемами.

Таким образом, далее рассматриваются три основных вопроса:

- исследования взаимосвязей технологических разработок с результатами научных исследований и информационное обеспечение решения задач оценки инновационного потенциала;
- анализ структуры и содержания (наполнения) ссылок на цитируемые документы в полнотекстовых информационных ресурсах патентных электронных библиотек и баз данных;
- основные положения методики дополнительного структурирования электронных форм описаний изобретений в той их части, где цитируются научные публикации и другие документы, например, стандарты.

2 Взаимосвязи научных исследований и технологических разработок

Задачи оценки инновационного потенциала направлений научных исследований в последнее время стали предметом научных исследований и их актуальность существенно возросла к началу 21 века [7, 9].

Согласно работе [7] причина усиления внимания к этим задачам заключается в том, что результаты финансирования фундаментальных исследований, ориентированных на развитие некоторой технологической сферы связаны, с одной стороны, с большим риском. С другой стороны, есть риск упустить новые технические решения и потерять конкурентоспособность в этой технологической сфере. Это касается конкурентоспособности в конкретной технологической сфере как отдельных предприятий, так и государства в целом. Поэтому и возникла потребность в решении задач мониторинга, анализа и количественной индикаторной оценки инновационного потенциала, а также в создании систем оценки как инструментов для стратегического управления финансированием фундаментальных исследований, ориентированных на технологическое развитие. При решении этих задач один из самых сложных вопросов заключается в том, как зафиксировать факт передачи и успешного использования в технологической сфере результатов фундаментальных исследований.

В силу указанных причин стали проводиться научные исследования различных мер «инновационности», видов индикаторов для ее оценки и методов для определения их значений. Один из предложенных подходов заключается в том, чтобы передачу знаний от науки к технологиям отслеживать с помощью документов, цитируемых в отчетах о патентном поиске. Для понимания рассматриваемых подходов и роли цитируемых документов необходимо сказать несколько слов о формуле изобретения. Формула изобретения

описывает область защиты технического решения для рассматриваемой экспертом патентной заявки и состоит из некоторого числа признаков патентуемого технического результата. Поскольку основной критерий для получения патента – новизна изобретения, то некоторые из этих признаков должны быть новыми или, по крайней мере, комбинация известных признаков должна быть новой. В процессе экспертизы задача эксперта состоит в том, чтобы найти и сослаться на уже существующие документы, которые описывают те же самые признаки или близкие тем, которые указаны в формуле изобретения рассматриваемой патентной заявки. Если нет документов с теми же признаками, то по этой заявке принимается решение о выдаче патента [7].

Следовательно, в результате экспертизы могут появляться цитируемые документы, которые описывают те же самые признаки или близкие тем, которые указаны в формуле изобретения рассматриваемой патентной заявки. Таким образом, результаты анализа роли, структуры и наполнения ссылок на научные публикации, цитируемые заявителями в патентных заявках и/или экспертами в отчетах о патентном поиске, являются исходными данными для разработки методологии решения задач оценки **инновационного** потенциала.

Отметим, что для анализа и оценки **научного** потенциала традиционно используются научные электронные библиотеки журнальных статей, а для анализа и оценки уровня **технологического** развития - патентные электронные библиотеки и базы данных. При этом, в каждом из этих двух случаев проведение анализа и оценки ограничено только одной сферой: в первом случае - сферой науки, во втором случае - сферой технологических разработок. Исследование взаимосвязей направлений научных исследований и технологических разработок относится одновременно к двум сферам и в процессе проведения такого исследования необходимо, в общем случае, использовать одновременно научные и патентные информационные ресурсы.

Однако результаты упомянутых в начале этого раздела исследований говорят о том, что достаточно часто при анализе и оценке инновационного потенциала направлений научных исследований ограничиваются только патентными электронными библиотеками и базами данных. Причина в том, что полнотекстовые описания изобретений включают краткие библиографические описания цитируемых научных документов (автор(ы) документа, его название, источник публикации), которых достаточно для определения целого ряда количественных индикаторов инновационного потенциала, примеры которых рассматриваются далее в этом разделе.

Используемые в настоящее время за рубежом подходы к определению количественных индикаторов инновационного потенциала направлений научных исследований основаны, как правило, на следующих положениях.

Во-первых, оценка инновационного потенциала некоторого направления научных исследований (области знаний) определяется как количественная оценка связи некоторой технологии с этой областью знаний, которая пропорциональна числу научных публикаций этой области знаний, цитируемых в описаниях изобретений к патентам, относящихся к соответствующей технологии. Так как инновационный потенциал некоторой области знаний может проявиться в разных технологических сферах, то оценка инновационного потенциала некоторой области знаний является векторной величиной, а нескольких областей знаний – матрицей (см. табл. 1). При получении этих оценок необходимо учитывать, что цитирование научных справочников и классических научных трудов в описаниях изобретений, скорее всего, будет указывать на хрестоматийные, а не на новые научные результаты. Кроме того, достоверность оценок во многом будет зависеть от представительности массивов описаний изобрете-ний, относящихся к этой новой технологии.

Во-вторых, оценки инновационного потенциала отражают влияние научных исследований на уровень технологического развития, но они не отражает других взаимосвязей (например, влияние результатов технологических разработок на инициирование проведения фундаментальных исследований и полученные новые научные результаты).

В-третьих, если в отдельно взятом описании изобретения нет ссылок на научные публикации, то это не должно интерпретироваться как отсутствие связей этого отдельно взятого изобретения с научными результатами, так как не все случаи передачи знаний обязательно эксплицируются в виде публикаций [6]. Кроме того, авторы «Третьего европейского отчета по научно-технологическим индикаторам» отмечают особое положение такой области знаний как математика [8, стр. 421]. С одной стороны, в описаниях изобретений к патентам (далее по тексту – в патентах) крайне редко встречаются ссылки на математические публикации (см. табл. 1), с другой стороны, очевидно, что математические методы и модели являются необходимыми для многих сфер технологий (производства).

Рассмотрим примеры количественных индикаторов, определенных с использованием информационных ресурсов патентных электронных библиотек.

Первый пример представляет собой матрицу чисел, характеризующих взаимосвязи сфер технологий (производства) и областей знаний, определенную на основе информационных ресурсов электронной библиотеки Европейского патентного ведомства. Эта матрица, приведенная в табл. 1, получена в результате обработки данных упомянутого европейского отчета [8, стр. 420].

В первой строке этой таблицы цифрами от 1 до 9 обозначены области знаний (их названия указаны в Примечании 1 под таблицей). В первой колонке перечислены 20 сфер технологий (производства). В соответствии с первым положением каждая строка табл. 1, кроме первой, содержит вектор частотностей научных публикаций одной из 9-ти перечисленных областей знаний. Учитывались только те публикации, которые были включены в патенты, относящиеся к указанной в начале строки сфере технологий (производства). Например, строка «Биотехнологии» содержит число 52.8, которое говорит о том, что в патентах по биотехнологиям 52.8% всех цитат являются ссылками к научным публикациям в науках о жизни.

Таблица 1.

Сферы технологий (производства) ↓	1	2	3	4	5	6	7	8	9
Биотехнологии	0.1	4.4	2.0	21.3	0.2	52.8	0.1	0.0	19.2
Фармацевтические и косметические средства	0.0	2.5	5.7	42.4	0.2	34.0	0.1	0.0	14.9
Технологии тонкого органического синтеза	0.0	2.8	10.8	28.8	0.2	40.6	0.1	0.0	16.5
Сельское хозяйство и химическое производство пищевых продуктов	0.1	33.4	2.1	4.2	0.8	48.1	0.0	0.0	11.2
Контрольно-измерительные технологии	0.4	1.5	6.3	29.0	6.6	32.4	10.3	0.0	13.3
Нефтехимическая промышленность и химическая переработка сырья	0.2	8.1	10.6	32.0	1.1	33.3	1.1	0.0	13.6
Технологии производства пищевых продуктов	0.0	15.6	3.9	15.6	3.9	41.7	0.6	0.0	18.9
Полупроводниковая промышленность	0.5	0.5	13.1	1.9	23.8	0.3	58.7	0.0	1.1
Телекоммуникационные технологии	0.6	1.9	1.0	2.5	77.0	0.7	15.7	0.2	0.3
Ядерная техника и технологии	1.6	0.0	8.1	17.7	37.1	4.8	24.2	0.0	6.5
Информационные технологии	1.2	1.0	1.2	6.8	71.1	5.4	11.4	0.2	1.7
Космическая техника	0.0	0.0	20.0	0.0	50.0	0.0	30.0	0.0	0.0
Оптическая промышленность	0.2	0.0	12.0	0.7	22.5	1.7	61.2	0.0	1.4
Медицинские технологии	0.0	1.1	2.8	51.7	4.0	23.4	6.8	0.0	9.9
Технологии обработки поверхностей и лакокрасочные технологии	1.7	0.6	32.8	1.7	16.9	2.8	39.0	0.0	4.5
Технологии химии полимеров	0.2	4.3	42.6	13.3	3.1	26.0	1.0	0.0	9.3
Аудиовизуальные технологии	0.0	0.0	0.0	2.9	63.8	0.0	32.6	0.0	0.7
Металлургическая промышленность и производство материалов	3.4	0.0	29.9	3.4	34.0	2.7	19.7	0.0	6.8
Электротехническая промышленность	0.0	0.4	23.3	3.6	25.7	0.4	42.2	0.0	4.4
Химическое машиностроение	2.9	2.9	42.7	9.7	18.4	9.7	4.9	0.0	8.7

Примечание 1. Цифрами в первой строке обозначены: 1 – науки о Земле и других планетах, 2- сельскохозяйственные науки, 3 – химия, 4 – медицинские науки, 5 – технические науки, 6 – науки о жизни, 7 – физика, 8 – математика, 9 – междисциплинарные проблемы (в этой таблице позиционируются как отдельная область знаний).

Примечание 2. Данные этой таблицы получены в результате обработки европейских патентных заявок, поданных в период времени 1992-1996гг.

Строка «Телекоммуникационные технологии» содержит числа 0.6 и 0.2, которые говорят о том, что в патентах этого вида технологий 0.6% всех цитат являются ссылками к научным публикациям в науках о Земле и 0.2% - ссылками к математическим публикациям.

Очевидно, что каждая из сфер технологий (производства) может быть также структурирована и для каждой сферы может быть получены своя матрица количественных индикаторов, определенных с использованием информационных ресурсов патентных электронных библиотек. Однако на разных уровнях структуризации используются разные подходы к определению матрицы индикаторов.

На верхнем уровне, примером которого является табл. 1, как правило, достаточно анализировать названия источника научной публикации (название журнала или материалов научной конференции). На следующем уровне дополнительно необходимо

анализировать название научной публикации. Чтобы перейти на следующий уровень анализа, необходимо дополнительно обращаться к электронным библиотекам научных публикаций и анализировать содержание аннотаций статей, а для наиболее детального анализа необходимо анализировать содержание полных текстов статей. Возникает естественный вопрос, зачем на разных уровнях использовать разные подходы, если матрицу индикаторов для любого уровня структуры можно получить с помощью универсального метода анализа содержание полных текстов статей.

Необходимость в разработке разных подходов и методов диктуется существенными различиями в стоимости их реализации, которая существенно возрастает при переходе от первого к последнему из элементов следующего списка анализируемых видов информационных объектов:

- название источника научной публикации,
- название научной публикации,
- аннотация публикации,
- полный текст публикации.

Сначала определяется тот информационный объект, стоимость обработки которого вписывается в бюджеты создания и эксплуатации системы оценки, а вид выбранного информационного объекта во многом уже диктует максимально возможную глубину структуризации и методы определения матрицы индикаторов.

Отметим, что приведенный список не является полным, так как в системах оценки имеется возможность использования отчетов о патентном поиске, экспертных заключений и анкет как видов информационных объектов, а также структурированных описаний тех результатов научно-технической деятельности, которые должны регистрироваться в соответствии с постановлением Правительства РФ от 4 мая 2005 г. № 284 «О государственном учете результатов научно-исследовательских, опытно-конструкторских и технологических работ гражданского назначения».

Например, в информационной системе Роспатента хранятся отчеты о патентном поиске и экспертные заключения на поданные патентные заявки. В информационной системе Российского фонда фундаментальных исследований (РФФИ) хранятся экспертные анкеты по проектам, в том числе тех проектов, в рамках которых получены результаты, имеющие признаки патентоспособности. В соответствии с правилами РФФИ руководитель проекта обязан информировать РФФИ о результатах работ по проекту, имеющих признаки патентоспособности в двухмесячный срок с момента получения заключения (оценки) о патентоспособности.

Имеется потенциальная возможность анализа отчетов о патентном поиске и заключений экспертов Роспатента на поданную патентную заявку, а также экспертных анкет проектов РФФИ, в рамках которых получен патентуемый результат.

Однако в настоящее время содержание отчетов о патентном поиске, экспертных анкет и заключений, как правило, недоступно в процессе решения задач оценки инновационного потенциала. Кроме того, структура и наполнение экспертных анкет РФФИ не учитывает потребности задач оценки инновационного потенциала проектов, результаты которых имеют признаки патентоспособности. Поэтому отчеты о патентном поиске, экспертные анкеты и заключения не включены в список анализируемых видов информационных объектов.

В заключение раздела рассмотрим другой пример индикатора, определенного с использованием информационных ресурсов электронной библиотеки Патентного ведомства США, показывающего изменение во времени среднего числа ссылок на публикации по любым областям знаний, встретившихся в одном патенте США в период с 1987 по 2002 годы (см. рис. 1). Верхняя кривая показывает изменение среднего числа ссылок на все публикации, встретившиеся в одном патенте США, а нижняя пунктирная кривая показывает изменение только для научных статей, но не для всех публикаций [5, vol. 1, p. 5-51].

Очевидно, что для каждой области могут быть получены аналогичные кривые с использованием информационных ресурсов патентных электронных библиотек. Следовательно, данные, приведенные в табл. 1, скорее всего, являются результатом усреднения за период времени 1992-1996гг., то есть, значения элементов матрицы индикаторов в табл. 1 не являются постоянными, а зависят от времени.

Таким образом, ретроспективный анализ информационных ресурсов патентных электронных библиотек позволяет оценивать инновационный потенциал направлений научных исследований (областей знаний) и его изменение во времени, если в качестве меры «инновационности» принять частотность научных публикаций каждой области знаний, цитируемых в патентах, относящихся к некоторой сфере технологий.

3 Ссылки на цитируемые документы

Приведенные в предыдущем разделе примеры индикаторов вычислялись на основе информационных ресурсов электронных библиотек Европейского патентного ведомства и Патентного ведомства США.

Возникает естественный вопрос о возможности определения значений аналогичных отечественных индикаторов на основе информационных ресурсов электронной библиотеки Роспатента.

Ранее были перечислены те структурно выделенные виды информационных объектов, которые используются для определения рассмотренных индикаторов на разных уровнях структуризации сфер технологий, включая название источника научной публикации, название самой научной публикации, цитируемой в патенте, аннотацию этой публикации и ее полный текст.

При этом структурное выделение аннотации статьи в электронных библиотеках научных публикаций без разметки названий источников и публикаций в патентах недостаточно для определения значений индикаторов. Причина в том, что эти названия являются теми структурными полями, которые позволяют соотнести библиографическое описание публикации в патентной электронной библиотеке с аннотацией (полным текстом) цитируемой публикации в электронной научной библиотеке.

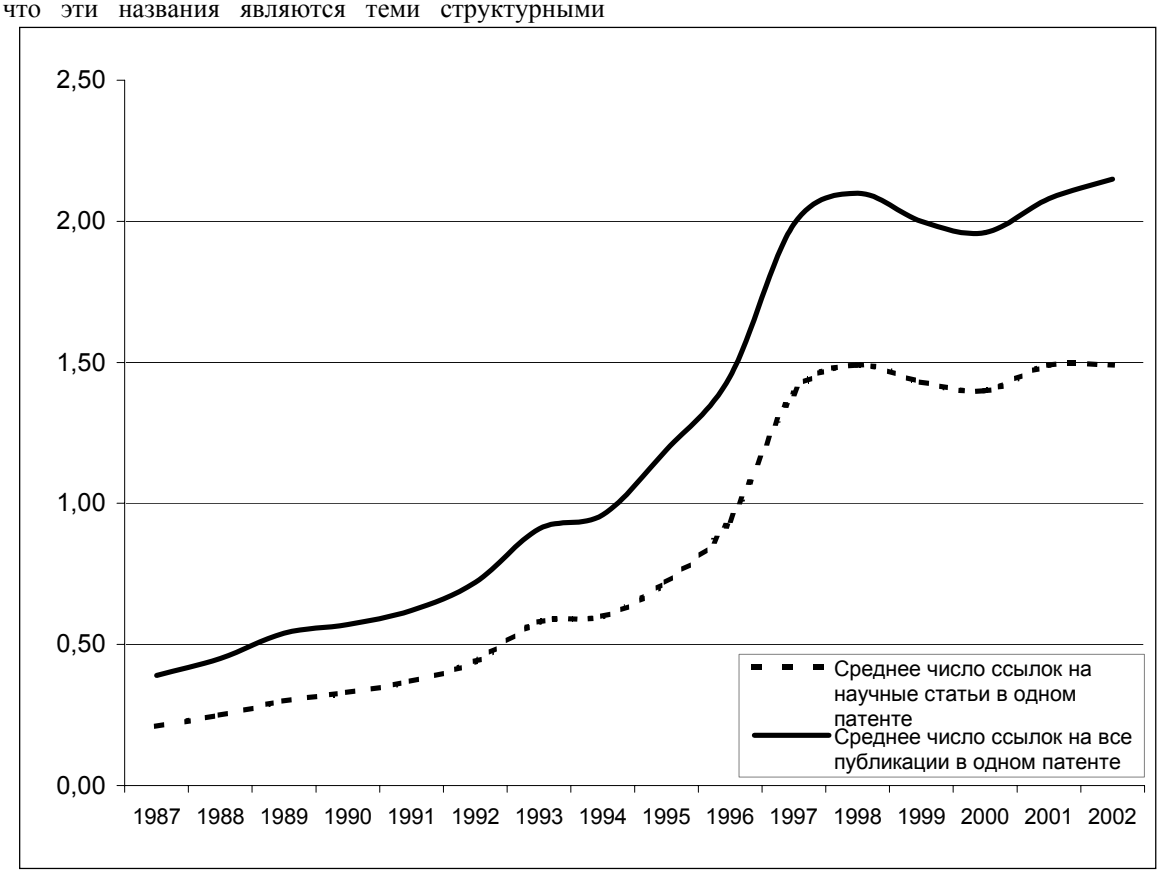

Рис. 1. Изменение среднего числа библиографических ссылок в одном патенте США за период 1987-2002гг.

Поэтому только структурное выделение аннотации статьи в электронных библиотеках научных публикаций без разметки названий источников и публикаций в патентной электронной библиотеке недостаточно для определения индикаторов. Так как в электронной библиотеке Роспатента эти названия не выделены как адресуемые поля, то нет возможности определить значения рассмотренных индикаторов из-за отсутствия меток для названий. Иногда названия публикаций и источников разделяются двумя подряд идущими символами «косая черта» (//), но это не является обязательным правилом.

Примером включения прямо в текст абзаца нескольких неразделенных ссылок может служить патент № 2256224, в котором есть следующий абзац с библиографическими описаниями трех публикаций:

"*Для решения задач обработки, анализа и распознавания изображений созданы алгоритмическо-программные системы, использующие современные вычислительные средства (Ю.И. Журавлев, И.Б. Гуревич, Распознавание образов и анализ изображений // Искусственный интеллект: В 3-х книгах, книга 2. Модели и методы: Справочник. М.: Радио и связь, 1990. С.149-191; Ю.И. Журавлев, И.Б. Гуревич, Методы и средства преобразования и обработки информации в задачах распознавания образов и анализа изображений // Параллельная обработка информации: В 5 т., т. 5: Проблемно-ориентированные и специализированные средства обработки информации. - Киев: Наукова думка, 1990. - С.218-318; И.Б. Гуревич, Ю.И. Журавлев, Д.М. Мурашов и др., Система автоматизации научных исследований в области анализа и понимания изображений на основе накопления и использования знаний. Ч.1 // Автометрия. - 1999. - №6. - С.23-50)*".

В качестве первого шага решения задачи определения значений индикаторов был проведен анализ номенклатуры информационных ресурсов, формируемых Роспатентом, структуры и содержания (наполнения) ссылок на цитируемые документы, включая научные публикации, в полнотекстовых информационных ресурсах отечественных патентных электронных библиотек и баз данных. Анализ проводился с целью определения тех изменений (детализации) схемы полнотекстовых описаний изобретений, которые необходимы для решения задач оценки инновационного потенциала направлений научных исследований.

Номенклатура информационных ресурсов, формируемых Роспатентом, определена Положением «Об официальных изданиях Федеральной службы по интеллектуальной собственности, патентам и товарным знакам» (далее по тексту - Положение). В номенклатуре значится Открытый реестр изобретений, размещенный на официальном сайте Роспатента (http://www.rupto.ru), на котором также можно найти все упоминаемые в докладе патентные нормативные документы и стандарты.

Содержание комплекта документов изобретения (титульный лист, полное описание и формула изобретения, чертежи и/или графические материалы) определяется п. 3.6 Положения. Отметим, что на титульном листе приводятся библиографические данные изобретения.

Структура и наполнение перечисленных документов определяется «Правилами составления, подачи и рассмотрения заявки на выдачу патента на изобретение» (далее по тексту – Правила).

В п. 3.2.4.2 Правил, указано, что «…непосредственно в тексте приводятся библиографические данные источника информации…». Требования к библиографическим данным приводится в п. 6.9 Правил и они сформулированы следующим образом «Библиографические данные источников информации указываются таким образом, чтобы источник информации мог быть по ним обнаружен».

В Правилах не содержится иных указаний на требования, предъявляемые к составлению библиографических данных (ссылок) на источники информации или документы (далее по тексту – ссылки на цитируемые документы). Однако если обратиться к международному опыту, то Всемирной организацией интеллектуальной собственности (ВОИС) был разработан ряд документов, направленных на стандартизацию электронного представления патентных документов, включая ссылки на цитируемые документы.

Требования к составлению ссылок, включая ссылки на научные статьи, заданы стандартом ВОИС ST.14, который носит рекомендательный характер. Проведенный выборочный анализ патентов РФ показывает, что, как правило, содержание ссылок на цитируемые документы в выборочном массиве описаний, сформированным на основе Открытого реестра изобретений, ему соответствуют. Однако для решения рассматриваемых задач важно не только содержание ссылок на цитируемые документы, но и структура их электронного представления.

Состав и структуру меток для электронного представления патентных документов определяет стандарт ВОИС ST.32 «Рекомендации по разметке патентных документов с использованием SGML». В этом стандарте есть группа меток для разметки ссылок на цитируемые документы, содержание которых определяется стандартом ST.14.

Стандарт ВОИС ST.36 «Рекомендации по обработке патентных документов с использованием XML» сохраняет преемственность ST.32 в части использования группы меток для ссылок на цитируемые документы.

Из всего вышеизложенного можно сделать вывод о том, что, если бы электронные формы отечественных патентных документов соответствовали бы стандартам ВОИС ST.14, ST.32 и ST.36 в части использования группы меток для разметки ссылок, то на основе отечественных патентных электронных библиотек и баз данных можно было бы вычислить ряд индикаторов оценки инновационного потенциала направлений научных исследований.

Однако в настоящее время электронные формы отечественных патентных документов в электронных библиотеках и базах данных не соответствуют рекомендованным стандартам ВОИС в части использования группы меток для разметки ссылок.

Поэтому была предпринята попытка разработать методику детализации схемы электронных форм патентных документов и дополнительного структурирования уже имеющихся информационных ресурсов патентных электронных библиотек и баз данных. При разработке методики использовались следующие исходные положения.

1. Правила, на основании которых отечественный заявитель готовит описание изобретения, определяют требования к содержанию ссылок на цитируемые документы, но не стандартизируют структуру представления библиографических ссылок.

2. В отечественных патентных документах не используется группа меток для разметки ссылок на цитируемые документы, содержание которые определяется стандартом ST.14.

3. Стандарт ВОИС ST.14 разрешает в ссылках на цитируемые документы использовать сокращения в названиях журналов (см. пп. 12.d.ii и 12.d.iii этого стандарта на официальном сайте Роспатента), а именно, разрешается использовать сокращенные названия периодических изданий, согласующиеся с общепризнанной международной практикой. При этом в самом стандарте указано только 234 рекомендованных сокращенных названий. Для остальных журналов отсутствует список нормализованных и сокращенных названий. Как следствие, один и тот же журнал может быть назван по-разному разными заявителями.

4. Отсутствуют национальная версия стандарта ST.36, которая могла бы определить сокращенные названия отечественных периодических изданий, используемые в ссылках на цитируемые документы.

Для разработки и опробования экспериментального варианта методики детализации схемы электронных форм описаний изобретений и дополнительного их структурирования, из Открытого реестра изобретений Роспатента было сделано две выборки[1)].

Для первой выборки случайным образом извлечено из реестра 100 патентов разной технологической направленности. Их анализ дал следующие результаты. На отобранные патенты всего приходится 996 ссылок на цитируемые документы, включая 542 ссылки непосредственно в текстах абзацев описания и 91 ссылку, которые выделены в виде отдельного списка и приведены в конце описания изобретения, а в самом тексте описания указывается номер из этого списка (оставшиеся ссылки относятся к титульному листу изобретения, где приводятся библиографические данные изобретения). Распределение 996 ссылок по видам документов приведено во втором и третьем столбцах табл. 2.

Таблица 2

Вид документов	**Случайная выборка**		**Тематическая выборка**	
	Число ссылок	**Процент ссылок**	**Число ссылок**	**Процент ссылок**
Патенты	509	51,1%	1169	46,8%
Авторские свидетельства	90	9,0%	292	11,7%
Заявки на изобретения	42	4,2%	550	22,0%
Книги	144	14,5%	291	11,7%
Статьи в энциклопедиях	5	0,5%	2	0,1%
Статьи в журналах или сборниках на русском языке	30	3,0%	79	3,2%
Статьи в журналах или сборниках на иностранном языке	125	12,6%	24	0,96%
Материалы конференций	7	0,7%	15	0,6%
Рефераты диссертаций	3	0,3%	1	0,04%
Диссертации	0	0%	4	0,16%
Отчеты	1	0,1%	13	0,5%
Стандарты	32	3,2%	49	2,0%
Web-ссылки	8	0,8%	7	0,3%
Сумма	996	100%	2496	100%

Как видно из таблицы, для первой выборки доля ссылок на непатентные цитируемые документы составляет около 36% от общего количества ссылок, из которых 15% приходится на книги и статьи в энциклопедиях и 16,3% - на статьи в журналах, сборниках и на материалы конференций. При этом, доля статей в журналах или сборниках на русском языке составляет 19,3% от общего числа статей в журналах или сборниках.

Для второй выборки был составлен тематический запрос на поиск патентов за несколько месяцев 2005 года одной тематической направленности (вычислительные устройства и системы). В результате было отобрано 324 патента. Их анализ дал следующие результаты. На отобранные патенты всего приходится 2496 ссылок на цитируемые документы.

Распределение этих ссылок по видам документов приведено в четвертом и пятом столбцах табл. 2. Как видно из таблицы, доля ссылок на непатентные цитируемые документы составляет около 20% от общего количества ссылок, из которых 11,8% приходится на книги и статьи в энциклопедиях и 4,8% - на статьи в журналах, сборниках и на материалы конференций. При этом, доля статей в журналах или сборниках на русском языке составляет 76,7% от общего числа статей в журналах или сборниках.

Следовательно, доля научных публикаций от общего числа цитируемых документов и доля иностранных статей существенно зависят от тематики изобретений.

4 Предлагаемая методика структурирования

На основе результатов анализа описаний изобретений в двух выборках с учетом четырех исходных положений, перечисленных в предыдущем разделе, была разработана методика дополнительного структурирования электронных форм описаний изобретений в той их части, где цитируются научные публикации и другие документы. Выборка описаний изобретений для дополнительного структурирования осуществлялась с помощью информационно-поисковой системы (ИПС) «МИМОЗА», которая поставляется Роспатентом вместе с библиографическими данными и рефератами патентов на DVD-диске. Для обеспечения процесса дополнительного структурирования и обработки электронных форм описаний изобретений был разработан комплекс программ в ИПИ РАН[2)].

В методике учитывается тот факт, что DVD-диск содержит ссылки на сайт Роспатента в сети Интернет, обеспечивающие бесплатный доступ к полным описаниям изобретений. Таким образом, пользователь ИПС «МИМОЗА», после выполнения поиска на DVD-диске библиографических данных и рефератов нужных ему изобретений, с помощью компьютера, подключенного к сети Интернет, сразу же может получить доступ к электронным формам полнотекстовых описаний изобретений, используя ссылки на соответствующие страницы сайта Роспатента.

Основные положения разработанной методики заключаются в следующем.

1. Используя DVD-диск, с помощью ИПС «МИМОЗА» по заданному запросу выбираются библиографические данные и рефераты описаний изобретений (например, по автору изобретения, дате публикации или

по коду Международной патентной классификации (МПК)).

2. Используя ссылки на сайт Роспатента, выбираются электронные формы полнотекстовых описаний изобретений, соответствующие найденным библиографическим данным и рефератам описаний изобретений.
3. Из каждого описания изобретения автоматически копируется часть библиографических данных (код МПК, название, автор(ы) изобретения и другие поля, необходимые для вычисления индикаторов).
4. С помощью разработанного в ИПИ РАН комплекса программ в каждом найденном описании изобретения выделяются, копируются и сохраняются ссылки на цитируемые документы. При этом для каждой ссылки указывается:
   a. вид публикации в соответствии с табл. 2;
   b. область в документе, где встретились ссылки (поле библиографических данных, абзацы текста описания или список использованной литературы в конце текста);
   c. для ссылок из текста – адрес в тексте описания.
5. Текст каждой ссылки на цитируемый документ структурируется.

Доступные в настоящее время информационные ресурсы Роспатента и разработанная методика позволяют формировать массивы структурированных ссылок на цитируемые документы для найденных описаний изобретений. Иначе говоря, задавая тематические запросы и используя предлагаемую методику, можно формировать массивы структурированных ссылок на цитируемые документы, соответствующие заданному запросу.

5 Заключение

Предлагаемая методика дополнительного структурирования электронных форм описаний изобретений была опробована с использованием DVD-диска Роспатента «Справочно-поисковый аппарат к описаниям изобретений за 1994-2005 годы» и сайта Роспатента. В тематическом запросе на поиск было заполнено два поля: «Дата публикации» и «Код МПК», соединенные оператором «И». Для первого поля было задано значение «2005», для второго – «G06», что является кодом рубрики, к которой относятся дискретные, аналоговые и гибридные вычислительные устройства и системы.

В результате выполнения этого запроса была сформирована выборка описаний изобретений, относящихся к рубрике G06, на которые были выданы патенты РФ и которые были опубликованы в 2005 году. С использованием разработанной методики дополнительного структурирования электронных форм описаний изобретений был сформирован массив структурированных ссылок на цитируемые документы для описаний изобретений, относящихся к рубрике G06. На основе этого массива и была сформирована тематическая выборка и получены данные для четвертого и пятого столбцов в табл. 2.

Так как в этих структурированных ссылках выделены названия журналов и научных статей, то это уже позволяет увидеть те направления исследований в информатике и других областях знаний, на результаты которых имеются ссылки в изобретениях, относящихся к рубрике G06. Так как выделен год публикации цитируемого документа, то также можно получить распределение цитируемых документов по годам их публикации.

На основе распределения документов по годам была построена диаграмма на рис. 2. Первый столбик диаграммы отображает долю документов (45%), опубликованных не ранее чем за пять лет до даты подачи заявки на выдачу патента. Последний столбик отображают долю документов (10%), опубликованных более чем за 20 лет до даты подачи заявки.

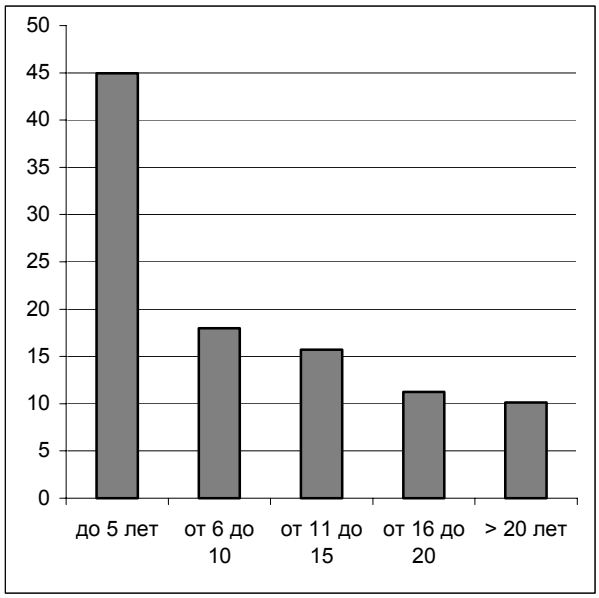

Рис. 2. Распределение цитируемых документов по времени их публикации

Первые результаты опробования методики дополнительного структурирования электронных форм описаний изобретений показали, что наследуемые информационные ресурсы патентных электронных библиотек могут быть использованы для оценки инновационного потенциала направлений научных исследований, если в качестве меры «инновационности» принять частотность цитируемых научных публикаций в каждом из направлений.

В заключение отметим, что возможны и другие меры «инновационности», использующие сопоставление содержания патентов и аннотаций (полных текстов) всех научных статей в достаточно

представительной электронной научной библиотеке, а не только содержания цитируемых в патентах статей. При этом в процессе сопоставления содержания патентов и научных статей необходимо учитывать, что в научных исследованиях и технологических разработках часто используются совершенно разные системы терминов.

Сопоставление систем терминов в широком спектре технологических разработок с учетом эволюции семантики во времени является весьма трудоемкой задачей. Однако использование цитируемых в патентах статей в интересах решения задачи сопоставления систем терминов позволяет уменьшить размерность и сложность этой задачи.

Литература

[1] Зацман И.М. Полидоменные модели в системах оценки инновационного потенциала и результативности научных исследований // Труды международной конференции Диалог-2006 "Компьютерная лингвистика и интеллектуальные технологии".- М.: Изд-во РГГУ, 2006.- С. 178-183.

[2] Зацман И.М. Полидоменные модели электронных библиотек систем мониторинга сферы науки // Труды Восьмой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2006 (Суздаль, 17-19 октября 2006 г.). – Ярославль: Ярославский государственный университет, 2006. - С. 75-81.

[3] Зацман И.М. Семантическое, информационное и знаковое кодирование патентных документов электронных библиотек // Труды Седьмой Всероссийской конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2005 (Ярославль, 4-6 октября 2005 г.).– Ярославль: Ярославский госуниверситет, 2005. С. 112-121.

[4] Кожунова О.С., Зацман И.М. Прагматические аспекты создания семантического словаря терминов информационного мониторинга // Труды международной конференции Диалог-2007 "Компьютерная лингвистика и интеллектуальные технологии".- М.: Изд-во РГГУ, 2007.- С. 278-285.

[5] National Science Board, Science and Engineering Indicators – 2004. Two volumes. Arlington, VA: National Science Foundation, 2004 (volume 1, NSB 04-1; volume 2, NSB 04-1A).

[6] Nonaka I., Takeuchi H. The knowledge-creating company. N. Y.: Oxford University Press, 1995 (перевод на русский язык: Нонака И., Такеучи Х. Компания – создатель знания. М.: ЗАО «Олимп-бизнес», 2003).

[7] Schmoch, U. Tracing the knowledge transfer from science to technology as reflected in patent indicators // Scientometrics. Vol. 26, 1993, pp. 193-211.

[8] Third European Report on Science & Technology Indicators. - Luxembourg: Office for Official Publications of the European Communities, 2003. - 451 pp.

[9] Tijssen, R.J.W., Buter, R.K., Van Leeuwen, Th.N. Technological relevance of science: an assessment of citation linkages between patents and research papers // Scientometrics. Vol. 47, 2000, pp. 389-412.

[10] Zatsman I., Kozhunova O. Evaluation System for Russian Academy of Sciences: Clarification Tools // Atlanta Conference on Science, Technology, and Innovation Policy 2007 (ATLC 2007), October 18-20, 2007. Proceedings. - Atlanta: Georgia Institute of Technology, 2007 (in print).

Processing Principles of Information Resources for Evaluation of Innovative Potential of Science Fields

I. Zatsman, S. Shubnikov

The paper is devoted to a problem of an evaluation of innovative potential of science fields with use of information of patent electronic libraries, access to which is given by Rospatent. The structure and filling of patent documents is analyzed. It is shown, that modern researches of S&T interaction (which science fields are most cited in patents?) are based on a processing of patent documents, containing references to scientific papers. Researches of S&T interaction clarify a number of requirements to methodology of processing of patent documents, and also to a degree of elaboration of schemes for patent electronic libraries. One of these requirements consists in additional structurization of references to scientific papers in patent documents.

♣ Работа выполнена при частичной поддержке РГНФ, грант № 06-02-04043а.

[1)] Анализ отобранных описаний изобретений выполнен С.В. Базилевич и Н.В. Луневой.

[2)] Программное обеспечение разработано С.К. Шубниковым, Г.И. Зацманом и М.Г. Кружковым.