УДК 347.77+007.51

А.С. Деревянко, О.С. Сомхиева

Национальный технический университет "ХПИ", Харьков

ПРИМЕНЕНИЕ СРЕДСТВ ИНТЕЛЛЕКТУАЛЬНОЙ ОБРАБОТКИ ДАННЫХ В ПАТЕНТНО-КОНЪЮНКТУРНЫХ ИССЛЕДОВАНИЯХ

Рассмотрен процесс проведения патентного поиска и проблемы, возникающие в ходе его, которые связаны с тем, что помимо патентной информации имеется еще информация патентно-ассоциированная, зачастую неструктурированная. К таким документам может быть применен широкий спектр алгоритмов Text Mining.

Ключевые слова: *Text Mining, Data Mining, патентно-конъюнктурные исследования, патентная информация.*

Введение

Постановка проблемы. Источниками, рассматриваемыми в патентно-конъюнктурных исследованиях (ПКИ), является патентная и патентно-ассоциированная информация (ПАИ) [1, 2]. Такая информация существует как в печатном виде, например, в виде книг, журналов, документов, отчетов, справочников, патентных документов, стандартов, спецификаций и каталогов, так и в других видах, например, аудиовизуальном и машиночитаемом, в виде передачи организационных и индивидуальных знаний на совещаниях, семинарах и мероприятиях по обучению.

Информация о промышленной собственности, известная как патентная информация, входит в систему научно-технической информации. Патентная информация является одним из главных составных элементов системы охраны промышленной собственности и включает в свою очередь патентную документацию.

Патентная информация получается из патентных баз данных (ПБД) различных ведомств и стран. Как правило, ПБД доступны через Интернет и имеют свои поисковые машины. При существенном различии в способах формулирования запросов и внешних форматах результатов, можно отметить, что результаты, полученные из ПБД, легко приводятся к единой общей структуре (в основной своей части описываемой стандартом Всемирной Организации Интеллектуальной Собственности ST.9 [3]), поскольку объект, являющийся результатом поиска, имеет один тип – патент. Следует, однако, отметить, что в составе этой структуры имеются атрибуты, рассматриваемые в ПБД как атомарные, но фактически представляющие собой полнотекстовые документы (полный текст патента, реферат), которые могут подвергаться дополнительному структурированию, выделению элементов и т.д.

Этого нельзя сказать о ПАИ. Патентная документация представляет собой широкий спектр документов, опубликованных или неопубликованных, содержащих данные о результатах исследований, конструирования, разработок, которые заявляют и признают в качестве изобретений, полезных моделей или промышленных образцов, а также об охране прав изобретателей, патентовладельцев и держателей дипломов и свидетельств о регистрации промышленных образцов и полезных моделей. Таким образом, ПАИ представляет собой публикации электронные или/и бумажные, содержащие полнотекстовые документы, которые не имеют явной структуры и не обязательно определяют все атрибуты, являющиеся необходимыми для патентной информации. Перед лицом, проводящим патентные исследования, стоит задача анализа содержания такого документа, фактически, аннотирования его и определения его релевантности.

Анализ литературы. Степень автоматизации ПКИ пока что невысока. Анализ соответствующей литературы [4 – 8] показывает, что главным направлением автоматизации здесь является автоматизированный поиск в ПБД или в Интернет-ресурсах общего назначения с использованием специализированных для конкретных ПБД или универсальных (Google, Yahoo) поисковиков. Даже поиск в ПБД дает обычно множество результатов, которые являются нерелевантными, а для поиска ПАИ используются поисковые системы общего назначения, в которых объем нерелевантных результатов еще больший. Технологии, позволяющие отсеивать нерелевантные результаты или автоматизировать формулирование запроса, не применяются. Вместе с тем, как мы попробуем показать, информационные технологии, позволяющие автоматизировать процесс ПКИ в значительно большей степени, существуют [9], и задача автоматизации ПКИ состоит не столько в изобретении новых методов и технологий, а в адаптации уже существующих технологий к предметной области и интеграции их. Отметим, что задача интеграции предполагает совместную работу в системе различных программных продуктов от различных производителей, и такая совместная работа может быть обеспечена только при том условии, что в основу их взаимодействия будут положены открытые стандарты. Ориентация на открытые стандарты также позволит легко заменять продукты, лежащие в основе тех или иных ее звеньев, что даст возможность оптимизировать систему по функциональным возможностям, по эффективности и по стоимости.

Цель статьи. Предложить архитектуру системы, работа которой основана на использовании методов интеллектуального анализа данных, применяемых к структурированной и неструктурированной части патентной информации. Система должна обеспечивать автоматизацию (отчасти, при участии человека исследователя) таких звеньев процесса ПКИ как формулирование запроса, оценка релевантности результата, сохранение результатов и онтологий. Мы намерены предложить информационные технологии, которые могут использоваться в отдельных звеньях процесса патентно-конъюнктурных исследований и для интеграции звеньев системы.

Решение проблемы

Концепция работы предлагаемой архитектуры системы поддержки ПКИ вкратце состоит в следующем:

1. ПКИ начинаются с поиска в ПБД по эмпирически выбранным (подобранным экспертами) ключевым словам.

2. Результаты этого поиска сохраняются в локальном хранилище данных и используются как *обучающая выборка*, на основании которой строится онтология предметной области.

3. Полученная онтология является исходным средством для формулирования запросов на поиск ПАИ и оценки релевантности этого поиска.

4. Результаты, полученные на каждом этапе (или на промежуточных стадиях каждого этапа) сохраняются в локальном хранилище данных и могут визуализироваться, корректироваться лицом, производящим исследование, и обновленная онтология может служить исходной для повторения процесса с любой предшествовавшей точки.

Укрупненно процесс обработки данных в системе представлен на рис. 1 (использована нотация BPMN).

Взаимодействие с ПБД является уже достаточно отработанной технологией, например, в поисковой системе Укрпатента [10] предусматривается портал

поиска, обеспечивающий единую точку доступа к информации из нескольких наиболее популярных ПБД на CD. ПБД на CD представляют собой специализированные хранилища данных в пропертиарных форматах, доступ к которым (поиск в которых) осуществляется при помощи специализированных программ-шлюзов, каждая из которых обеспечивают доступ к большему или меньшему множеству форматов. Не отрицая необходимости хранения патентной информации на CD (именно в такой форме происходит обмен между патентными ведомствами различных стран) мы все же предлагаем продублировать это хранилище данных (или хотя бы витрину этого хранилища) в универсальной реляционной базе данных, приведя данные из разных источников к единому формату. В состав системы, следовательно, должны входить программы шлюзы, обеспечивающие процесс ETL для данных, поступающих на CD-носителях. Шлюзы эти, однако, применяются эпизодически – только при поступлении новых носителей.

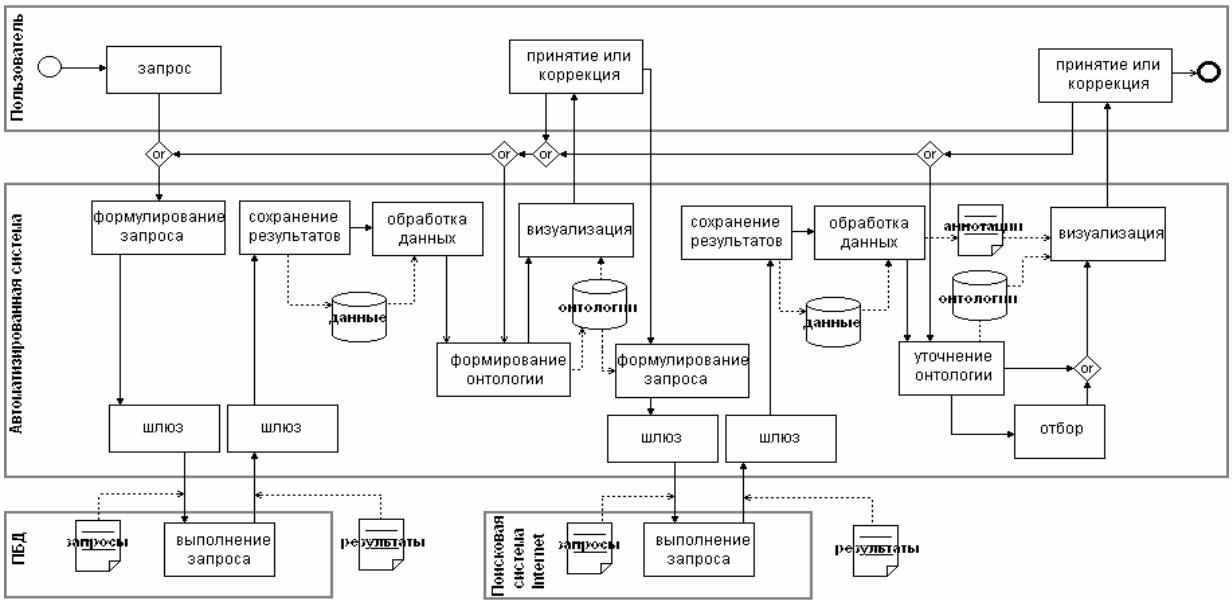

Рис. 1. Система поддержки патентно-конъюнктурных исследований

Наряду с локальным хранилищем информации ПБД источником такой информации должен быть Internet. Для этого источника выполнение запроса к ПБД и получение результата также должно быть унифицировано. Программный шлюз между различными ПБД и единым хранилищем системы для этой задачи должен быть единственным, но он должен выполнять преобразование запроса в формат конкретной ПБД и результата – в формат общего хранилища по шаблонам XSLT, связанным с конкретными ПБД [11]. Система должна иметь расширяемый репозиторий таких шаблонов для различных ПБД. Перенося различие в форматах интерфейсов ПБД в шаблоны, мы переносим их из алгоритмов в данные и тем самым уменьшаем сложность системы.

Хранилищем данных для хранения результатов запросов и результатов их обработки (оно же может использоваться и как хранилище шаблонов) должна быть современная СУБД. По этим мы имеем в виду то, что это хранилище должно обеспечивать широкую функциональность, связанную с формулированием запросов поиска и аналитикой и обеспечивать высокую производительность и надежность хранения данных и работе с ними. Эти требованиям не представляется возможным реализовать на хранилище с пропертиарными форматами хранения данных.

Некоторая проблема возникает при выборе типа СУБД для такого хранилища. С одной стороны, как мы отмечали, информация ПБД является хорошо структурированной и естественным представляется хранить ее в реляционной модели данных, что может также обеспечить дополнительную функциональность при ее обработке. С другой стороны, как мы покажем ниже, результаты обработки в большинстве случаев представляются в форматах, базирующихся на языке XML, и для них было бы естественным сохранение в базе данных XML. К счастью, ведущие современные реляционные СУБД дают пользователю возможность работать с данными XML и формулировать запросы как на языке SQL, так и на языке XQuery или на комбинации этих двух языков. Имеется даже пример (IBM DB2 v.9) полностью гибридной СУБД, сочетающей в себе как реляционное, так и XML-хранилище данных.

После начального заполнения базы данных к полученным данным применяются методы интеллектуальной обработки, включающие в себя как методы OLAP по выделенным атрибутам реляционной структуры – ранжирование, определение статистических показателей [12], так и методы Data Mining (по структурированной части) и Text Mining (по полнотекстовым полям) – классификация и кластеризация [13, 14]. Здесь также стоит отметить, что патентный анализ в данной области свидетельствует о положительной динамике изменений, а, следовательно, и финансовых инвестиций в соответствующую технологию. Так, абсолютным чемпионом среди компаний-патентообладателей является корпорация IBM. У нее

265 патентов. Далее идет Microsoft с 65 патентами.

К Text Mining в настоящее время относят следующие способы обработки текстов: тематическое индексирование, поиск по ключевым словам, кластеризация, классификация, выделение феноменов, понятий, устойчивых характеристик, ответ на вопросы, составление аннотаций или резюмирование текстов.

Конечной целью интеллектуальной обработки данных на этом этапе должно являться построение онтологии прикладной области запроса – таксономии объектов предметной области и описания свойств объектов и взаимодействия их друг с другом. В открытом проекте semantic Web ведется разработка средств формального описания знаний, в частности языка RDF (resource description framework) и на его основе – языка OWL (web ontology language) [15]. Использование этих языков (они строятся на базе XML), во-первых, даст нам мощное семантическое средство для представления онтологий, во-вторых, позволит использовать разработанные Сообществом инструменты создания онтологий и работы с ними, в-третьих, предоставит в наше распоряжение базовые онтологии, уже разработанные Сообществом для некоторых прикладных областей.

На основании полученных онтологий строится запрос к поисковым системам общего назначения. Здесь опять могут использоваться несколько поисковых систем со шлюзом, управляемым шаблоном XSLT. Результаты, полученные от поисковых систем, будут представлять патентно-ассоциированную информацию, полнотекстовые документы. К этим документам может быть применен широкий спектр алгоритмов Text Mining с целью автоматического аннотирования, кластеризации, выделения понятий и, в конечном счете, уточнения онтологий. Главной целью такой обработки является сегментация полученных данных по двум основным сегментам – релевантные и нерелевантные и формирование и уточнение атрибутов этих двух сегментов.

Визуализация результатов (как на предварительном, так и на окончательном этапах) состоит в:

- представлении пользователю ранжированных аннотированных результатов поиска и отбора с возможностью редактирования их (перемещения в ранжированном списке, удаления);
- представлении пользователю самих онтологий с возможностью их редактирования (несмотря на то, что имеются дружественные для пользователя средства такого представления/редактирования, эта опция может быть более сложной для пользователя).

Коррекция, выполненная на списке результатов, в конечном итоге также превращается в коррекцию онтологий. После коррекции обработка может быть повторена с формированием запроса к ПБД в соответствии с уточненной онтологией; формированием запроса к поисковым системам общего назначения в соответствии с уточненной онтологией; повторением обработки уже имеющихся данных в соответствии с уточненной онтологией.

Представленный на рисунке процесс включает в себя большое число компонентов, работа которых должна координироваться. (Отметим также, что представленная на рисунке схема является весьма укрупненной, многие ее компоненты являются, в свою очередь, составными, в частности, обработка данных включает последовательное выполнение целого ряда модулей, реализующих различные алгоритмы трансформации и анализа.) Существует несколько архитектур построения таких систем, из которых наиболее развитым нам представляется открытый проект фирмы IBM UIMA (Unstructured Information Management Architecture) [16]. Эта архитектура предусматривает наличие «движков», управляющих в соответствии с заданным сценарием выполнением модулей обработки (прежде всего – интеллектуального анализа данных) документов или коллекций документов. Обмен данными между модулями происходит через гибкую структуру данных, называемую CAS (Comon Analysis System) и представляемую средствами языка XML. В CAS содержится «аннотация» документа, включающая в себя его разметку, выполненную модулями анализа, и любую другую информацию, в частности, это может быть и построенная для документа онтология. Модули UIMA имеют строго специфицированные интерфейсы и могут быть производителями CAS, потребителями CAS или совмещать в себе обе эти функции. Для сторонних модулей, включаемых в систему, могут быть сделаны «обертки», обеспечивающие для них интерфейс модулей UIMA. Архитектура UIMA поддерживается открытыми программными средствами, включающими в себя помимо движков и библиотек для разработки модулей и «оберток», также и модули преобразования форматов документов и выполнения некоторых наиболее общих алгоритмов Text Mining.

Предложенная архитектура будет иметь наиболее эффективное применение в корпоративной среде, когда возникает возможность накапливания и повторного использования как данных, так и онтологий. В этом случае возможно выполнение процесса обработки данных в распределенной среде, для чего представленный на рисунке процесс должен быть детализирован, переведен (возможно, автоматически) в описание на языке BPEL и установлен на каком-либо сервере приложений, имеющий в своем составе контейнер, поддерживающий хореографию бизнес-процессов (например, IBM WebSphere).

В качестве готового программного продукта, который также можно использовать для решения данной задачи можно порекомендовать IBM DB2 Data Miner for Text. С его помощью возможно решение следующих задач Data Mining: классификация данных, кластеризация, анализ данных в целях определения важности атрибута, построение ассоциативных правил. IBM DB2 Data Miner является одним из немногих продуктов на рынке, позволяющим собирать данные результатов для последующего анализа с помощью других продуктов, например инструментов OLAP. Она обладает средствами визуализации данных, а также поддерживает несколько ал-

горитмов добычи информации.

Что касается рынка программных продуктов Text Mining, то тут существенная часть разработок и готовых коммерческих модулей приходится на долю компаний, хорошо зарекомендовавших себя на рынке создания средств управления базами данных и интеллектуальных методов их обработки. Это, прежде всего, такие лидеры рынка средств Data Mining, как IBM и Oracle. Наряду со своими мощными средствами по обработке числовых и категориальных данных в пакетах Data Mining, эти компании развивают направление анализа информационных массивов, используя как свои технологические разработки в области нейронных сетей и деревьев решений, так и специфические инструменты анализа текстов.

Существует свободно распространяемый продукт, который содержит модули, могущие служить основой для создания собственных цепочек Text Mining в среде UIMA. Это система TextAnalyst Российской компании Мегапьютер Интеллидженс. Она решает следующие задачи Text Mining: создание семантической сети большого текста, автоматическое аннотирование текста, поиск по тексту, классификацию документов, кластеризацию текстов. Преимущество данной системы заключается в том, что она рассматривает технологию Text Mining в качестве отдельного математического аппарата, который разработчики программного обеспечения могут встраивать в свои продукты, не опираясь на платформы информационно-поисковых систем или СУБД.

Выводы

Современное развитие средств ИТ предоставляет полный набор технологий и программных продуктов, позволяющих в значительной степени автоматизировать и облегчить процесс поиска и анализа патентной и патентно-ассоциированной информации, задача в настоящий момент состоит в интеграции и адаптации этих средств к нуждам названной прикладной области и преодолении некоторой инертности пользователей.

Список литературы

1. Основи правової охорони інтелектуальної власності в Україні / За заг. ред. О.А. Підопригори, О.Д.Святоцького. – К.: Видавничий Дім "Ін Юре", 2003. – 236 с.

2. Кузнєцов Ю.М. Основи патентознавства та авторського права / Ю.М. Кузнєцов. – К.: ЗМОК, 2001. – 206 с.

3. Руководство по информации и документации в области промышленной собственности Стандарты – ST.9 [Електронний ресурс] / Укрпатент. – Режим доступу: http://www.ukrpatent.org/atachs/ sst9.doc.

4. Косенко С. Патентна інформація [Електронний ресурс] / Юридичний журнал. Вип. 8. – К.: Європейський університет, 2006. – Режим доступу: http://www.justinian.com.ua/article.php?id=2354.

5. Рагойша А. А. Патентные базы данных в Интернете [Електронний ресурс] / Режим доступу: http://www.abc.chemistry.bsu.by/patent/default.htm.

6. Аветисов А.Р. Современный патентный поиск: использование традиционных источников и возможностей сети Интернет [Електронний ресурс] / Белорусский медицинский журнал. – 2004. – Вип. 3. – Режим доступу: http://bsmu.by/bmm/03.2004/41.html.

7. Брагарник О. Використання Internet як засобу патентного пошуку [Електронний ресурс] / Державний департамент інтелектуальної власності. – Режим доступу: http://www.sdip.gov.ua/ukr/help/statti/ dopuintei/bragarnyk.

8. Клейман А.М. Проведение патентных исследований в Интернете и оформление отчёта о поиске. Метод. указания. [Електронний ресурс]. – Режим доступу: http://mgtu-sistema.ru/triz/patent_metodika.pdf.

9. Деревянко А.С. Технологии и средства консолидации информации / А.С. Деревянко, М.Н. Солощук. – Х.: НТУ «ХПИ», 2007. – 224 с.

10. Інструкції по використанню інформаційно-пошукової системи [Електронний ресурс] / Укрпатент. – Режим доступу: http://www. ukrpatent. org/instruct.html.

11. Альошкіна Ю.А. Консолідація патентної інформації з різних джерел / Ю.А. Альошкіна, О.С. Дерев'янко // Мат-лы VIII науч.-техн. конф. "Проблемы информатики и моделирования". – Х.: НТУ "ХПИ", 2008. – С. 11.

12. Барсегян А.А. Технологии анализа данных / А.А. Барсегян и др. – СПб: БХВ-Петербург, 2007. – 384 с.

13. Data Mining. Теория и практика / Под ред. И.Н. Брянцева. – М.: БДЦ-пресс, 2006. – 208 с.

14. Чубукова И.А. Data Mining: учебное пособие / И.А. Чубукова. – М.: БИНОМ – 382 с.

15. OWL Web Ontology Language Overview [Електронний ресурс] / W3C Recommendation 10 February 2004. – Режим доступу: http://www.w3.org/ TR/owl-features/.

16. Gotz T., Suhre O. Design and Implementation the UIMA Common Analisys System / T. Gotz, O. Suhre // IBM System Journal. – 2004. – Vol.43, № 3. –P. 476-489.

Надійшла до редколегії 15.09.2009

Рецензент: д-р техн. наук, проф. П.А. Качанов, Национальный технический университет "ХПИ", Харьков.

ЗАСТОСУВАННЯ ЗАСОБІВ ІНТЕЛЕКТУАЛЬНОЇ ОБРОБКИ ДАНИХ В ПАТЕНТНО-КОН'ЮНКТУРНИХ ДОСЛІДЖЕННЯХ

А.С. Деревянко, О.С. Сомхиева

Розглянутий процес проведення патентного пошуку і проблеми, що виникають в ході його, які пов'язані з тим, що окрім патентної інформації є ще інформація патентно-асоційована, часто неструктурована. До таких документів може бути застосований широкий спектр алгоритмів Text Mining.

Ключові слова: *Text Mining, Data Mining, патентно-кон'юнктурні дослідження, патентна інформація.*

APPLICATION OF FACILITIES OF INTELLECTUAL PROCESSING OF DATA IS IN PATENT-CONJUNCTURE RESEARCHES

À.Ñ. Derevyanko, O.S. Somkhieva

A process is considered leadthroughs of patent search and problem, arising up during his, which are related to that besides patent information there is yet information patent-associated, frequently unstructured. To such documents the wide spectrum of algorithms of Text Mining can be applied.

Keywords: *Text Mining, Data Mining, patent-conjuncture researches, patent information.*