Обробка інформації в складних організаційних системах

УДК 519.7 + 347.77

О.И. Король

Национальный технический университет «ХПИ», Харьков

ПРЕДСТАВЛЕНИЕ И КЛАССИФИКАЦИЯ НЕСТРУКТУРИРОВАННЫХ ПАТЕНТНО-КОНЪЮНКТУРНЫХ ДАННЫХ НА ОСНОВЕ ИСПОЛЬЗОВАНИЯ АЛГЕБРЫ КОНЕЧНЫХ ПРЕДИКАТОВ

Осуществлен анализ существующих систем интеллектуальной обработки данных. Рассмотрены вопросы представления и классификации неструктурированной патентно-конъюнктурной (ПКИ) и патентно-ассоциативной информации (ПАИ), на основе метода компараторной идентификации теории Булевой алгебры.

Ключевые слова: *интеллектуальная обработка данных, патентно-конъюнктурная и патентно-ассоциативная информация, метод компараторной идентификации.*

Введение

Цель статьи: предложить новый метод представления, интеллектуального поиска и обработки ПКИ и ПАИ, основанный на теории алгебры Буля и алгебры конечных предикатов.

Анализ последних исследований показал, что для облегчения патентно-конъюнктурных исследований необходимо автоматизировать процесс поиска и применить методы интеллектуальной обработки информации (ИОИ). А.С. Деревянко и О.С. Сомхиева – исследователи кафедры «Информатики и интеллектуальной собственности» НТУ «ХПИ» – в своих работах предложили систему, работа которой основана на использовании методов интеллектуального анализа данных, применяемых к структурированной и неструктурированной части патентной информации. Концепция работы предлагаемой архитектуры вкратце состоит в том, что: сначала производят поиск ПКИ в патентных базах данных по эмпирически выбранным (подобранным экспертами) ключевым словам; затем результаты этого поиска сохраняются в локальном хранилище данных и используются для построения онтологий предметной области, которые, в свою очередь, являются исходным средством для формулирования запросов на поиск ПАИ и оценки релевантности этого поиска. Кроме того, 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 [1].

Однако, стоит отметить, что в подобной системе не учтены лингвистические технологии, которые позволят провести более полное патентно-конъюнктурное исследование, дополненное ПАИ, представленной на разных языках. Потому что абсолютная новизна объектов промышленной собственности заключается в мировой неизвестности данного объекта. Из этого следует, что необходимо учитывать ПКИ и ПАИ, которая встречается на совершенно не знакомых и не понятных языках.

Отметим, что одними из наиболее эффективных технологических средств реализации функциональности систем ИОИ на сегодняшний момент являются: средства поиска закономерностей и нетривиального анализа данных Data Mining, оперативный анализ информации OLAP, специализированные средства их создания работы с текстовыми документами Text Mining. Вопросами работы данных систем занимаются такие ученые как: О.И. Петренко, С.В. Маклаков, В.Е. Туманов [2 – 4].

Однако, Data Mining, OLAP, Text Mining системы в задачах интеллектуального поиска и обработки ПКИ и ПАИ применены еще не были. Использование данных систем ИОИ должно предоставить возможность решать целый ряд актуальных для анализа ПКИ и ПАИ задач:

– консолидировать информацию из разнородных источников (патентные базы данных различных ведомств и стран, опубликованные или неопубликованные источники о научных конференциях, внешние источники и т.д.), в том числе иностранных, в хранилище данных, с предварительной очисткой, преобразованием данных и приведением информации к общей корпоративной модели данных;

– формировать наглядные графические и табличные представления имеющейся информации (визуализация данных);

– формировать в автоматическом режиме произвольные отчеты по созданной модели данных, на основе которых проводить тот или иной анализ.

Кроме того, наличие неструктурированной информации, такой как ПАИ, порождает массу проблем: их нельзя индексировать, невозможно организовать поиск по их содержимому или автоматическое сравнение этой информации. Чтобы избежать подобного дублирования данных, необходимо представить патентные данные таким образом, чтобы по ним можно было организовывать поиск (с надлежащими скоростью, точностью и полнотой) и выдачу данных пользователю.

Как отмечается в [5], неструктурированные данные содержат информацию о своей структуре внутри себя, и именно благодаря этому пользователь однозначно распознает и классифицирует тексты по внутреннему содержанию. Такое хранение информации о структуре данных внутри самих данных дает возможность разработать способы их классификации и методы их обработки и хранения. Таким образом, семантику данных можно хранить в виде знаний, а для адресации к данным, основанным на их содержимом, использовать базу знаний предметной области, содержащую понятия и отношения между ними. На основе этой базы можно произвести формальное описание содержания документов, используя ее понятия в качестве ключевых слов при индексировании данных.

Основная часть

Используем теорию алгебры конечных предикатов для представления онтологий, по которым будем осуществлять поиск, а также ограничительные текстовые поля в виде формул вида: $F(x_1, x_2, \ldots, x_n)=1/0$.

Для того, чтобы отобрать информацию, касающуюся зарегистрированных разработок (в качестве примера рассмотрим изобретение), закодируем такие словосочетания: «формула изобретения», «устройство», «процесс» / «способ» / «метод», «вещество» / «композиция», «применение», «приоритет»:

$$x_1^{ф} \wedge x_2^{о} \wedge x_3^{р} \wedge x_4^{м} \wedge x_5^{у} \wedge x_6^{л} \wedge x_7^{а}=1, \quad (1)$$

$$x_1^{у} \wedge x_2^{с} \wedge x_3^{т} \wedge x_4^{р} \wedge x_5^{о} \wedge x_6^{й} \wedge x_7^{с} \wedge x_8^{т} \wedge x_9^{в} \wedge x_{10}^{о}=1, \quad (2)$$

или

$$(x_1^{п} \wedge x_2^{р} \wedge x_3^{о} \wedge x_4^{ц} \wedge x_5^{е} \wedge x_6^{с} \wedge x_7^{с}) \vee$$
$$\vee(x_1^{с} \wedge x_2^{п} \wedge x_3^{о} \wedge x_4^{с} \wedge x_5^{о} \wedge x_6^{б}) \vee \quad (3)$$
$$\vee(x_1^{м} \wedge x_2^{е} \wedge x_3^{т} \wedge x_4^{о} \wedge x_5^{д})=1,$$

$$(x_1^{в} \wedge x_2^{е} \wedge x_3^{ш} \wedge x_4^{е} \wedge x_5^{с} \wedge x_6^{т} \wedge x_7^{в} \wedge x_8^{о}) \vee$$
$$(x_1^{к} \wedge x_2^{о} \wedge x_3^{м} \wedge x_4^{п} \wedge x_5^{о} \wedge x_6^{з} \wedge x_7^{и} \wedge x_8^{ц} \wedge x_9^{и} \wedge x_{10}^{я})=1, \quad (4)$$

или

$$x_1^{п} \wedge x_2^{р} \wedge x_3^{и} \wedge x_4^{м} \wedge x_5^{е} \wedge x_6^{н} \wedge x_7^{е} \wedge x_8^{н} \wedge x_9^{и} \wedge x_{10}^{е}=1 \quad (5)$$

или

$$x_1^{п} \wedge x_2^{р} \wedge x_3^{и} \wedge x_4^{о} \wedge x_5^{р} \wedge x_6^{и} \wedge x_7^{т} \wedge x_8^{е} \wedge x_9^{т}=1. \quad (6)$$

В качестве ограничителя или запретной комбинации пока будем использовать такие слова: «реклама», «продам», «акция»:

$$x_1^{р} \wedge x_2^{е} \wedge x_3^{к} \wedge x_4^{л} \wedge x_5^{а} \wedge x_6^{м} \wedge x_7^{а}=1, \quad (7)$$

$$x_1^{п} \wedge x_2^{р} \wedge x_3^{о} \wedge x_4^{д} \wedge x_5^{а} \wedge x_6^{м}=1, \quad (8)$$

$$x_1^{а} \wedge x_2^{к} \wedge x_3^{ц} \wedge x_4^{и} \wedge x_5^{я}=1. \quad (9)$$

Поиск осуществляем с помощью алфавитных операторов, а обработку методом компараторной идентификации документов [6].

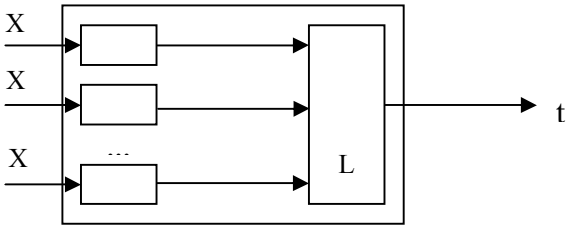

Рис. 1. Сущность метода компараторной идентификации

На вход системы подается множество сигналов (тексты патентных документов, информация о научных конференциях, выставках и т.д.) $x_1, x_2, \ldots, x_n$. Входные сигналы берутся из конечных множеств $X_1, X_2, \ldots, X_n$, причем $x_1 \in X_1, x_2 \in X_2, \ldots, x_n \in X_n$. В результате работы системы, осуществляющей обработку патентно-конъюнктурной информации, на выход поступает определенное множество элементов (ключевые понятия, дескрипторы, основные элементы исследуемого объекта и т.д.) $y_1, y_2, \ldots, y_n$, причем $y_1 \in Y_1, y_2 \in Y_2, \ldots, y_n \in Y_n$. Элементы $y_1, y_2, \ldots, y_n$ однозначно зависят от сигналов $x_1, x_2, \ldots, x_n$, что показывает на существование функций $y_1 = f_1(x_1), y_1 = f_1(x_1), \ldots, y_1 = f_1(x_1)$, каждая из которых представляет собой сюръекцию, отображающую множество X_i на множество Y_i, $i \in \{1,2,\ldots,n\}$ [6].

В отличие от традиционных систем рассматриваемые понятия x_i – неоднородные объекты, являющиеся элементами модели проблемной области: понятия, свойства, значения этих свойств, составные понятия и т.п. Отношения между ними – произвольные связи. При этом для представления патентно-конъюнктурных данных с разных точек зрения (хранения и обработки, представления и т.п.), а также для реализации различных стратегий или методик поиска может использоваться множественное индексирование. Следовательно, неоднородность понятий патентной базы данных и конъюнктурной информации определяется также и различием механизмов свертывания информации (представления основного содержания документов набором ключевых слов). В целом же множество понятий X описывает в сжатом виде документы множества Y.

Алгебра конечных предикатов полностью характеризуется алфавитом A, состоящим из k символов $a_1, a_2,\ldots,a_k$, и алфавитом переменных B, состоящих из n символов $x_1, x_2, \ldots, x_n$. Средствами алгебры конечных предикатов может быть описан любой n-местный и k-ичный предикат $f(x_1, x_2, \ldots, x_n)$, заданный алфавитом A. Алгебра предикатов позволяет представить любой предикат в виде суперпозиции базисных операций, примененных к базисным элементам, т.е. с ее помощью могут быть описаны любые конечные отношения [7].

Поэтому для каждой пары x и y существует точно определенное значение соответствия (или несоответствия) данного понятия проблемной области и рассматриваемого документа. Это соответствие может быть выражено некоторым предикатом, который назовем предикатом релевантности R. Для каждой пары документа (сигнала) $x_1 \in X_1, x_2 \in X_2, \ldots, x_n \in X_n$ и понятия (элемента) $y_1 \in Y_1, y_2 \in Y_2, \ldots, y_n \in Y_n$. значение предиката равно 1 в случае соответствия данного понятия патентно-конънктурному документу или 0 – в противном случае. При этом важно, чтобы при каждой попытке установить соответствие x и y предикат релевантности R определялся однозначно. Выполнение этого требования (так называемого постулата существования предиката R) означает, таким образом, установление для каждой пары (x, y) при повторном рассмотрении того же значения предиката R(x, y). В случае, когда компарацию осуществляет человек, постулат существования идеально точно выполняться не будет никогда [8]. Идеально точно выполнение постулата существования возможно лишь при компьютерной обработке информации. Если R(x, y) = 1, назовем документ x истинным относительно понятия y.

Для построения фрагментов результирующей гиперструктуры необходима классификация отобранной информации, разбиение на отдельные логические элементарные группы, логические единицы [9]. Назовем каждый такой отдельный элемент представления информации в гиперструктуре архитектурным конструктивом и введем принадлежность Δ отобранных мультимедиа-документов x_a, x_b одному конструктиву следующим образом:

$$x_a \Delta x_b \Leftrightarrow (\forall y \in Y)\,(R(\mathbf{x}_a, y) = R(\mathbf{x}_b, y)).$$

Отношение Δ определяет минимально расчлененное представление цельности. Действительно, если $x_a \Delta x_b$, то $R(x_a, y) = R(x_b, y)$ для любого понятия или отношения, описывающего заданную предметную область. Это означает, что данные понятия содержатся в индексных записях каждого из этих документов, т.е. информация, представленная в документах x_a и x_b, семантически близка. Если же отношение $x_a \Delta x_b$ не имеет места для данных документов x_a и x_b, это означает, что существует такое понятие $y^* \in Y$, которое соответствует только одному из указанных документов. Следовательно, не все свойства этих документов являются адекватными друг другу относительно заданного множества понятий Y [5]. В свою очередь, для понятий проблемной области, лежащих в основе построения гиперструктуры, также можно ввести отношение П принадлежности понятий y_i, y_j понятийной основе конструктива Y [9]:

$$xi \prod yj \Leftrightarrow (x \in X)\,(R(x, yi) = R(x, yj)). \quad (10)$$

Отношение П определяет закономерности структурирования цельности, а именно: если для любого патентно-конъюнктурного документа выполняется $R(x, y_i) = R(x, y_j)$, следовательно, оба понятия проблемной области yi и yj являются функционально эквивалентными для механизма компрессии отобранной информации. В противном случае, если отношение П не выполняется для некоторых y_i и y_j, то найдется такой документ $x^* \in X$, что $R(x^*, y_i) \neq R(x^*, y_j)$, т.е. какое-либо из понятий не соответствует этому документу.

Аналогично, можно показать, что отношение П обладает свойствами рефлексивности, симметричности и транзитивности, т.е. также является отношением эквивалентности [9]. Отношения Δ и П позволяют ввести соответствующие им предикаты ЕΔ (x_a, x_b) и ЕП (y_i, y_j), которые однозначно определяются предикатом релевантности R.

$$\text{Предикат } E\Pi(x_a, x_b) = (y \in Y)\,(R(x_a, y) \sim R(x_b, y)) \quad (11)$$

задан на множестве $X \times X$

$$\text{Предикат } E\prod(y_i, y_j) = (\forall x \in X)\,(R(x, y_i) \sim R(x, y_j)) \quad (12)$$

задан на множестве $Y \times Y$.

Предикат ЕΔ можно использовать для определения семантической близости документов x_a и x_b из множества X: если ЕΔ $(x_a, x_b) = 1$, то, согласно (1), $R(x_a, y) = R(x_b, y)$ для любого понятия y из множества Y. Следовательно, все свойства документов x_a и x_b, выражаемые понятиями из рассматриваемого множества, совпадают. Если же ЕΔ $(x_a, x_b) = 0$, то найдется такое понятие $y \in Y$, для которого $R(x_a, y) \in R(x_b, y)$, что говорит о семантическом различии x_a и x_b. Подобным образом предикат ЕП (y_i, y_j) может быть использован для определения функциональной эквивалентности понятий y_i и y_j из множества Y: если ЕП $(y_i, y_j) = 1$, то, согласно (12), $R(x, y_i) = R(x, y_i)$ для любого документа Z из множества X, т.е. либо оба понятия соответствуют документу $x \in X$, либо одновременно не соответствуют.

Оба введенных предиката ЕΔ и ЕП являются эквивалентностями, следовательно, факторизуют рассматриваемые множества X и Y [10, 11]. Предикат ЕΔ определяет разбиение множества X на слои S семантически близких документов; документы, принадлежащие различным слоям S, таковыми не являются. При этом классу Vx всех документов $x \in X$, семантически близких документу $s \in X$, соответствует предикат $Vx(x) = E\Delta(x, s)$. Предикат ЕП определяет разбиение множества Y на слои C функционально эквивалентных понятий; при этом поня-

тия из различных слоев С функционально эквивалентными не являются. Классу Vr всех понятий $y \in Y$, функционально эквивалентных понятию $r \in y$, соответствует предикат $Vr(y) = EП(y, r)$.

Учитывая (11) и (12), получаем:

$$Vx(x) = (y \in Y)(R(x, y) \sim R(x, y)), \quad (13)$$

$$Vr(y) = (x \in X)(R(x, y) \sim R(x, r)). \quad (14)$$

Формулы (13) и (14) выражают закономерности структурирования информации в процессе построения патентных структур через предикат R, объективно определяемый классификатором при проведении поиска.

Выводы

Сфера интеллектуальной собственности развивается быстро и динамично, что влечет за собой необходимость создания эффективных поисковых систем, позволяющих отсеивать нерелевантные результаты или автоматизировать формулирование запроса, а также применять лингвистические методы для комплексной обработки информации. Это объясняется тем, что в современных условиях информация становится реальным производственным ресурсом.

Системы интеллектуальной обработки патентно-конъюнктурной информации являются тем классом информационных систем, который позволяет превратить данные из специализированных баз данных, из больших потоков неструктурированной интернет-информации и данные из внешних источников в полезную для юридических органов или для бизнеса информацию, на основе которой можно принимать решения.

При организации поиска ПКИ и ПАИ, на вход поступает информация об объекте исследования из разных ресурсов, распределенных в сети. На выходе получаем множество материалов, объединенных в логическую последовательность в индивидуальном подходе поиска и обработки. Механизмом данной функции является адаптация, организованная в виде навигационных правил. Предложенные в статье средства представления и классификации неструктурированных патентно-конъюнктурных а ассоциативных данных основаны на использовании метода компараторной идентификации для разбиения на классы эквивалентности и связывания в гиперструктуру документов, отобранных в результате запроса к базе патентных данных.

Список литературы

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

2. Петренко О.І. Grid та інтелектуальна обробка даних Data Mining / О.І. Петренко // Системні дослідження та інформаційні технології. – 2008. – №4.

3. Маклаков С.В. Анализ данных // С.В. Маклаков, Д.В. Матвеев. – СПб.: БХВ-Петербург, 2003. – 496 с.

4. Туманов В.Е. Проектирование хранилищ данных для систем бизнес-аналитики [Электронный ресурс]. – Режим доступа к курсу: www.intuit.ru/.

5. Зайцев И.Е. Адаптивные технологии в современных автоматизированных системах / И.Е. Зайцев // Известия РГПУ им. А.И. Герцена. Аспирантские тетради. – 2007. – №21(51). – С. 214-217.

6. Шаронова Н.В. Автоматизированные информационные библиотечные системы: задачи обработки информации: монография / Н.В. Шаронова, Н.Ф. Хайрова. – Х.: Нар. Укр. акад. [Каф. информац. Технологии и документоведения], 2003. – 120 с.

7. Бондаренко М.Ф. Теория интеллекта / М.Ф. Бондаренко, Ю.П. Шабанов-Кушнаренко. – Х.,2006. – 576 с.

8. Шабанов-Кушнаренко Ю.П. Компараторная идентификация лингвистических объектов / Ю.П. Шабанов-Кушнаренко, Н.В. Шаронова. – К.: Институт системных исследований образования Украины, 1993. – 115 с.

9. Горбач Т.В. Представление и классификация неструктурированных данных в адаптивных обучающих мультимедиа-системах на основе метода компараторной идентификации / Т.В. Горбач, Я.В. Святкин, И.Ю. Шубин // Проблемы информационных. технологий. – 2009. – №1 (005).

10. O'Connor, Wade V. Informing Context to Support Adaptive Services. Proc. Conf. Hypermedia and Adaptive Web-Based Systems // LNCS 4018, Springer-Verlag, 2006.

11. Buitelaar P. Ontology Learning from Texts: An Overview / P. Buitelaar, P. Cimiano, B. Magnini // Frontiers in Artificial Intelligence and Applications. – 2005. – 123.

Поступила в редколлегию 3.10.2011

Рецензент: д-р техн. наук, проф. Е.А. Дружинин, Национальный аэрокосмический университет им. Н.Е. Жуковского «ХАИ», Харьков.

ПРЕДСТАВЛЕННЯ ТА КЛАСИФІКАЦІЯ НЕСТРУКТУРОВАНИХ ПАТЕНТНО-КОН'ЮНКТУРНИХ ДАНИХ НА ОСНОВІ ВИКОРИСТАННЯ АЛГЕБРИ КІНЦЕВИХ ПРЕДИКАТІВ

О.І. Король

Розглянуто питання представлення та класифікації неструктурованої патентно-кон'юнктурної та патентно-асоціативної інформації), на основі методу компараторної ідентифікації теорії Булевої алгебри.

Ключові слова: *інтелектуальна обробка даних, патентно-кон'юнктурна та патентно-асоціативна інформація, метод компараторної ідентифікації.*

PRESENTATION AND CLASSIFICATION OF NON-STRUCTURED PATENT INFORMATION ON BASE OF THE FINITE-PREDICATE ALGEBRA

O.I. Korol

In the article made an analyze of exciting systems of intellectual treatment of information. Also deals with the matter of non-structured patent and associated information. It is based on the method of comparator's identification of the method of Boolean algebra.

Keywords: *intellectual data treatment, patent and market information, associated information, method of comparator's identification.*