УДК 347.77

О.И. Король[1], Н.В. Шаронова[2]
[1] ХПИ, г. Харьков, Украина, korolka@bk.ru
[2] ХПИ, г. Харьков, Украина, nvsharonova@mail.ru

ИНТЕЛЛЕКТУАЛЬНАЯ ОБРАБОТКА ДАННЫХ ПРИ ФОРМИРОВАНИИ ПАТЕНТНО-КОНЪЮНКТУРНЫХ БАЗ ЗНАНИЙ

Создание модели интеллектуальной системы, ориентированной на описание закономерностей в текстовой патентно-конъюнктурной информации, представленных на естественном языке, и моделирование всех уровней лингвистической обработки в процессе формирования баз знаний с помощью универсального математического аппарата, основанного на алгебре конечных предикатов.

ИНТЕЛЛЕКТУАЛЬНАЯ ОБРАБОТКА ДАННЫХ, ПАТЕНТНО-КОНЪЮНКТУРНАЯ ИНФОРМАЦИЯ, АЛГЕБРА КОНЕЧНЫХ ПРЕДИКАТОВ

Введение

Современные справочно-информационные поисковые программы и патентные базы данных сохраняют детали того или иного объекта интеллектуальной собственности с довольно высокой точностью. Однако объемы данных растут, они легкодоступны в сети Интернет, но при этом не обладают точной и понятной структурой и не являются знанием в полном смысле слова. При обработке информации нужно получить закономерности, а не потоки и списки данных. Таким образом, стоит задача интеллектуальной обработки патентно-конъюнктурной информации (ПКИ) для извлечения знаний, существующих в хранилищах данных, а также применение лингвистических технологий для более полной обработки.

Рассмотрим основные особенности патентно-конъюнктурных данных, представленных в системах хранения информации, с точки зрения поиска закономерностей, существующих в них:

• Как правило, описание объектов содержит не менее 25-50 характерных признаков или полей базы данных, где каждый признак может быть дискретным (и иметь 5-10 и более значений) или непрерывным.

• Множество значений и признаков не является окончательным и может измениться.

• Критерии, определяющие качество объектов, носят как формальный, так и экспертный характер, т.е. не всегда могут быть выражены явной зависимостью.

• Объем данных достаточно велик и растет, причем некоторые существенные в прошлом по отношению к данной задаче объекты теряют свое качество от времени из-за постоянного развития технологий.

• Опытный эксперт обладает интуитивным знанием закономерности, определяющей качественный объект, и способен отделить существенные по отношению к данной задачи объекты от несущественных, но не может выразить свои знания в явном виде.

• Процесс оценки и ранжирования объектов занимает достаточно существенное время.

Анализ последних исследований показал, что наиболее известной технологией интеллектуальной обработки данных является интеллектуальный анализ данных (Data Mining) – исследовательский анализ данных, имеющий целью отыскание интересных взаимосвязей между данными, скрытых закономерностей, которые могут использоваться при принятии решений [1]. Основная особенность Data Mining – объединение широкого математического инструментария и последних достижений в сфере информационных технологий, разработанных на основе искусственного интеллекта к организации процесса извлечения знаний из потока данных. К наиболее известным подходам относят системы на основе нейронных сетей [2], статистических методов [3, 8], нечеткой логики, методов обобщения по примерам объектов (KAD [4], АТ-ТЕХНОЛОГИЯ [5], INDUCE [6] и др.), которые обеспечивают работу в средах с разными типами данных и могут работать с экспертом, не являющимся программистом. Рассматриваемые системы реализуют процесс обобщения и некоторый уровень обработки входной информации для подготовки исследуемых данных. Например, АТ-ТЕХНОЛОГИЯ позволяет преобразовывать структурированную информацию базы данных в базу знаний экспертной системы на основе применения алгоритмов класса ID3 [5].

Из вышесказанного видно, что поиск закономерностей вручную задача трудоемкая и требует применения современных технологий автоматизации обработки данных и интеллектуальных систем, ориентированных на описание закономерностей и моделирование всех уровней лингвистической обработки текстовой ПКИ. На сегодняшний день подобные технологии не применяются, а поиск ПКИ осуществляется путем обращения к государственным патентным базам данных и с помощью универсальных internet поисковиков (Google, Yandex, Yahoo и др.).

1. Интеллектуальные системы обобщения

Рассмотрим интеллектуальную систему «Трейд» [7], обеспечивающую пользователю поддержку при обработке потоков данных, поступающих из различных источников, и попробуем применить ее для нашей прикладной области. Архитектура интеллектуальной системы «Трейд» приведена на рис. 1.

Основой системы является алгоритм обобщения по примерам, подробно рассмотренный в ряде работ [5, 6]. Алгоритм требует представления описания рассматриваемой области как многомерного дискретного пространства D_n, где n – число координат-характеристик объекта $X = \{x_1, x_2, \ldots, x_n\}$, а описания объектов как точки пространства D_n в виде вектора, содержащего значения признаков $\phi = \{a_1, a_2, \ldots, a_n\}$, где a_i – значение i-го параметра. Пространство содержит неизвестные, т.е. не получившие оценку эксперта факты, и известные, составляющие базу фактов. База фактов разделена на две части: T – существенные объекты по отношению к данной задаче и F – несущественные объекты. Разделение осуществляется либо экспертом, либо с помощью набора критериев $C = \{f_1, \ldots, f_k\}$, где C_i определяет i-ю локальную закономерность на данном наборе фактов, а f_k – реализация k-го критерия для C_i.

На множествах T и F строится разделяющая их функция выбора $y(\phi_i)$ такая, что $y(\phi) \geq 0$, если $\phi \in T$ и $y(\phi) < 0$ если $\phi \in F$. Для неизвестной части пространства D_n функция выбора будет разделять объекты в соответствии с законом, полученным на базе фактов. В работе [3] рассмотрен процесс построения многоуровневой функции выбора для работы в больших пространствах.

Система «Трейд», использующая рассмотренный алгоритм обобщения, должна выделить признаки и их описания для каждой предметной области, определить базы фактов для каждой из локальных закономерностей C_i, построить функцию выбора и сохранить ее в базу знаний. Для получения результата необходимо осуществить следующие действия.

• Выделить поля БД, которые являются входными характеристиками (множество X).

• Выделить поля, являющиеся оценками ситуаций, или ввести эти оценки с помощью эксперта.

• Обратиться к блоку выделения подпространств для определения множества целей $\{C_1, C_2, \ldots, C_s\}$ и построить множество Φ_i (множество существенных примеров для i-й цели) и его дополнение 1 – Φ_i (множество несущественных примеров для i-й цели) по каждой цели C_i.

Для каждой цели C_i необходимо породить многоуровневую функцию выбора [7].

Видно, что система «Трейд» легко работает, когда пользователю надо сформировать запрос из одного-двух критериев. Ситуация усложняется, когда цель определена множеством критериев (например, не только поиск существующего технического решения, но и учет степени тождественности или похожести до степени смешения и т.п.). Это свидетельствует о существовании локальных закономерностей внутри общего поля знания. Причем чем сложнее описание объекта, тем больше будет локальных областей (при условии существования большой выборки соответствующих им фактов). Следовательно, нам необходим универсальный математический аппарат, который был бы ориентирован на моделирование всех уровней лингвистической обработки текстовой патентно-конъюнктурной документации (ПКД), а также поддерживал процесс формирования

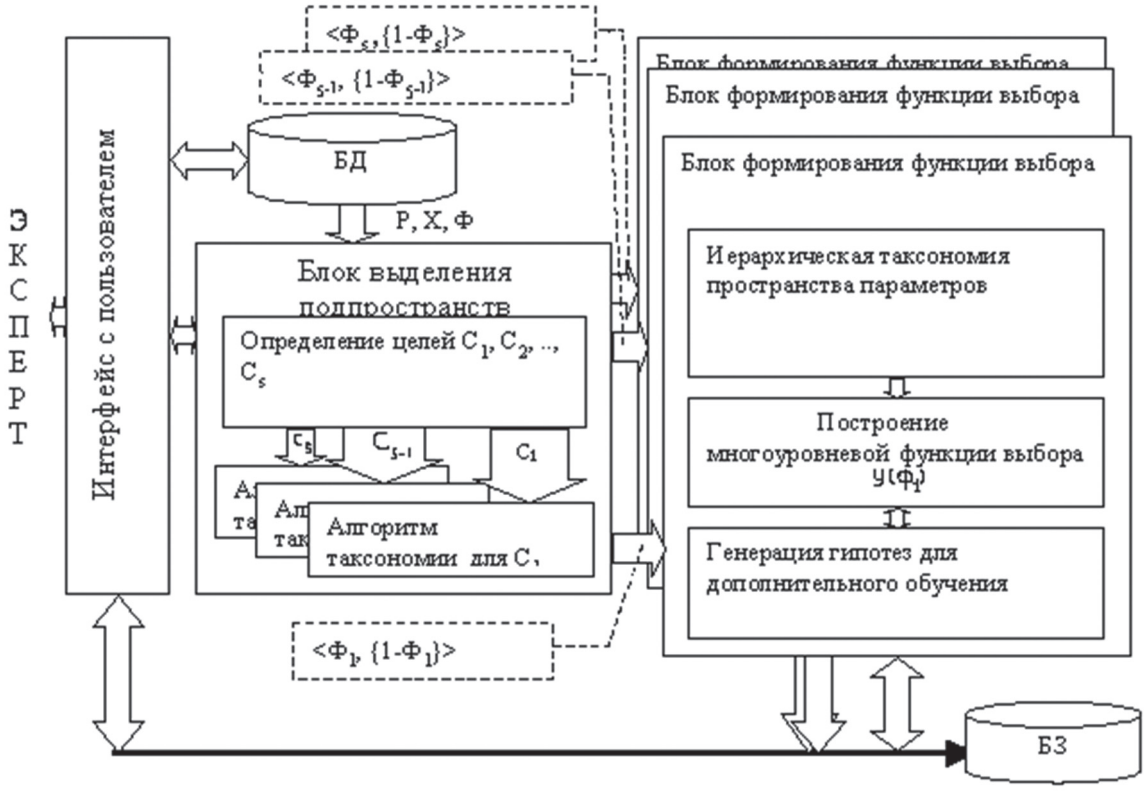

Рис. 1. Архитектура интеллектуальной системы «Трейд»

исходных наборов фактов для работы алгоритмов обобщения с учетом следующих функциональных возможностей [9]:

• Выделение множества целей на основе критериев качества объектов.

• Выделение существенных признаков, коррекция.

• Формализация описания предикатов, которые реализуются при любом виде интеллектуальной обработки ПКИ, для формирования уровней, описывающих свойства этих предикатов.

• Поддержка процесса извлечения закономерностей на основе алгоритма обобщения.

• Поддержка процесса немонотонного развития системы во времени.

2. Алгебра конечных предикатов — универсальный аппарат для описания закономерностей ПКИ

Вышеуказанные функции можно осуществить с помощью алгебры конечных предикатов (АКП) [10]. АКП полна в том смысле, что любой ее предикат можно представить в виде суперпозиции базисных операций, примененных к базисным элементам.

АКП характеризуется алфавитом A, состоящим из k символов $\{a_1, a_2, \ldots, a_k\}$ и алфавитом переменных B, состоящим из n символов $\{x_1, x_2, \ldots, x_n\}$ [11]. В логике высказываний «атомом» является высказывание, которое далее не разделяется – предикат. Предикат P, заданный на множестве U^n, представляет собой любую функцию $\varepsilon = P(x_1, x_2, ..., x_n)$, отображающую данное множество U^n в $\sum = \{0,1\}$.

Алгебра предикатов при любом значении n является разновидностью булевой алгебры, в ней выполняются все основные тождества булевой алгебры [10]. Базисными предикатами для АКП являются предикаты вида:

$$x_i^a = \begin{cases} 1, \text{ если } x_i = a \\ 0, \text{ если } x_i \neq a \end{cases} (1 \le i \le n), \qquad (1)$$

где $i = \{1, 2, \ldots, n\}$, a – любой элемент универсума.

Прежде, чем представить системы предикатов ПКИ в виде, понятном АКП, предлагается создать онтологии прикладной области. Построение онтологии предполагает определение классов объектов и описание их отношений с помощью одного из формальных языков, например, дескриптивной логики, что позволяет отвечать на запросы или так называемые компетентные вопросы (competency questions), которые изначально составляются на естественном языке и затем «переводятся» на используемый формальный язык. Процесс построения онтологии проиллюстрирован на рис. 2 и состоит из следующих этапов [12-13]:

• Извлечение терминов: этот этап состоит в том, чтобы обнаружить и извлечь из входного корпуса термины и их свойства. Для этого используются специализированные информационные ресурсы — глоссарий патентных терминов и терминов по интеллектуальной собственности, синтаксические шаблоны. С помощью синтаксического анализатора извлекаются пары (объект, свойство) и триплеты (объект, свойство, объект), относящиеся к общим смысловым блокам.

• Извлечение внешних отношений: на этом этапе применяется реляционный анализ понятий для извлечения внешних отношений.

• В завершение, результаты выполнения двух предыдущих этапов объединяются для получения более полной онтологии.

Таким образом, часть системы онтологий ПКИ может иметь вид, представленный на рис. 3, а система предикатов описана формулами (2) – (7).

Промышленная собственность $X_1^{\text{ПС}}$:

$$\begin{cases} X_1^{\text{ИПМ}} \vee X_1^{\text{ПО}} \vee X_1^{\text{ТИМС}} = 1, \\ X_1^{\text{ИПМ}} \wedge X_1^{\overline{\text{ИПМ}}} = 0, \\ X_1^{\text{ПО}} \wedge X_1^{\overline{\text{ПО}}} = 0, \\ X_1^{\text{ТИМС}} \wedge X_1^{\overline{\text{ТИМС}}} = 0, \end{cases} \qquad (2)$$

где $X_1^{\text{ИПМ}}$ – изобретения и полезные модели, $X_1^{\text{ПО}}$ – промышленные образцы, $X_1^{\text{ТИМС}}$ – топографии интегральных микросхем.

Объекты изобретения и полезной модели $X_{11}^{\text{ИПМ}}$:

$$\begin{cases} X_{11}^{\text{Н}} \vee X_{11}^{\text{ПН}} = 1, \\ X_{11}^{\text{Н}} \wedge X_{11}^{\overline{\text{Н}}} = 0, \\ X_{11}^{\text{ПН}} \wedge X_{11}^{\overline{\text{ПН}}} = 0, \end{cases} \qquad (3)$$

где $X_{11}^{\text{Н}}$ ($X_{11}^{\overline{\text{Н}}}$) – новое (не новое) изобретение, $X_{11}^{\text{ПН}}$ ($X_{11}^{\overline{\text{ПН}}}$) – применение (не применение) ранее известного изобретения по новому назначению.

Топографии интегральных микросхем $X_{12}^{\text{ТИМС}}$:

$$\begin{cases} X_{12}^{\text{ТИМС}} = X_1^{\text{ПС}} \wedge X_0^{\text{ИС}}, \\ X_{12}^{\text{ТИМС}} \wedge X_{12}^{\overline{\text{ТИМС}}} = 0, \end{cases} \qquad (4)$$

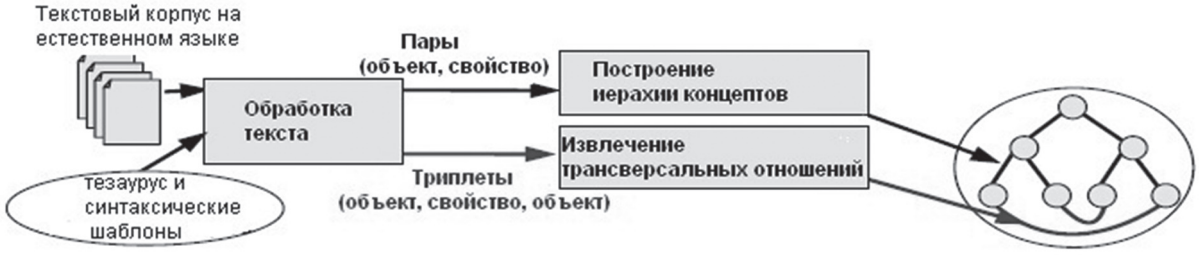

Рис. 2. Методика построения онтологии

где $X_0^{\text{ИС}}$ – интеллектуальная собственность, $X_1^{\text{ПС}}$ – промышленная собственность, $X_{12}^{\text{ТИМС}}$ – топографии интегральных микросхем, $X_{12}^{\overline{\text{ТИМС}}}$ – не топографии интегральных микросхем.

Объекты промышленного образца (ПО) $X_{13}^{\text{ПО}}$:

$$\begin{cases} X_{13}^{\text{К1}} \vee X_{13}^{\text{К2}} \vee .... \vee X_{13}^{\text{К32}} = 1, \\ X_{13}^{\text{К1}} \wedge X_{13}^{\overline{\text{К1}}} = 0, \\ X_{13}^{\text{К2}} \wedge X_{13}^{\overline{\text{К2}}} = 0, \\ ..... \\ X_{13}^{\text{К32}} \wedge X_{13}^{\overline{\text{К32}}} = 0, \end{cases} \quad (5)$$

где $X_{13}{}^{\text{К1}}$, $X_{13}{}^{\text{К2}}$, ... $X_{13}{}^{\text{К32}}$ – классы ПО, $X_{13}^{\overline{\text{К1}}}$, $X_{13}^{\overline{\text{К2}}}$, ... , $X_{13}^{\overline{\text{К32}}}$ – не классы ПО.

Объекты нового изобретения $X_{111}^{\text{Н}}$:

$$\begin{cases} X_{111}^{\text{ПЦ}} \vee X_{111}^{\text{У}} \vee X_{111}^{\text{Ве}} \vee X_{111}^{\text{ШТ}} = 1, \\ X_{111}^{\text{ПЦ}} \wedge X_{111}^{\overline{\text{ПЦ}}} = 0, \\ X_{111}^{\text{У}} \wedge X_{111}^{\overline{\text{У}}} = 0, \\ X_{111}^{\text{Ве}} \wedge X_{111}^{\overline{\text{Ве}}} = 0, \\ X_{111}^{\text{ШТ}} \wedge X_{111}^{\overline{\text{ШТ}}} = 0, \end{cases} \quad (6)$$

где $X_{111}^{\text{ПЦ}}$ – процесс, $X_{111}^{\text{У}}$ – устройство, $X_{111}^{\text{Ве}}$ – вещество, $X_{111}^{\text{ШТ}}$ – штамм микроорганизма, $X_{111}^{\overline{\text{ПЦ}}}$ – не процесс, $X_{111}^{\overline{\text{У}}}$ – не устройство, $X_{111}^{\overline{\text{Ве}}}$ – не вещество, $X_{111}^{\overline{\text{ШТ}}}$ – не штамм микроорганизма.

Классификация новых устройств $X_{111}^{\text{У}}$:

$$\begin{cases} X_{1112}^{A} \vee X_{1112}^{B} \vee X_{1113}^{C} \vee X_{1112}^{D} \vee X_{1112}^{E} \vee X_{1112}^{F} \vee \\ \vee X_{1112}^{G} \vee X_{1112}^{H} = 1, \\ X_{1112}^{A} \wedge X_{1112}^{\overline{A}} = 0, \\ X_{1112}^{B} \wedge X_{1112}^{\overline{B}} = 0, \\ X_{1112}^{C} \wedge X_{1112}^{\overline{C}} = 0, \\ X_{1112}^{D} \wedge X_{1112}^{\overline{D}} = 0, \\ X_{1112}^{E} \wedge X_{1112}^{\overline{E}} = 0, \\ X_{1112}^{F} \wedge X_{1112}^{\overline{F}} = 0, \\ X_{1112}^{G} \wedge X_{1112}^{\overline{G}} = 0, \\ X_{1112}^{H} \wedge X_{1112}^{\overline{H}} = 0, \end{cases} \quad (7)$$

где X_{1112}^{A}, X_{1112}^{B}, X_{1112}^{C}, X_{1112}^{D}, X_{1112}^{E}, X_{1112}^{F}, X_{1112}^{G}, X_{1112}^{H} – классы новых изобретений в зависимости от сферы применения, $X_{1112}^{\overline{A}}$, $X_{1112}^{\overline{B}}$, $X_{1112}^{\overline{C}}$, $X_{1112}^{\overline{D}}$,

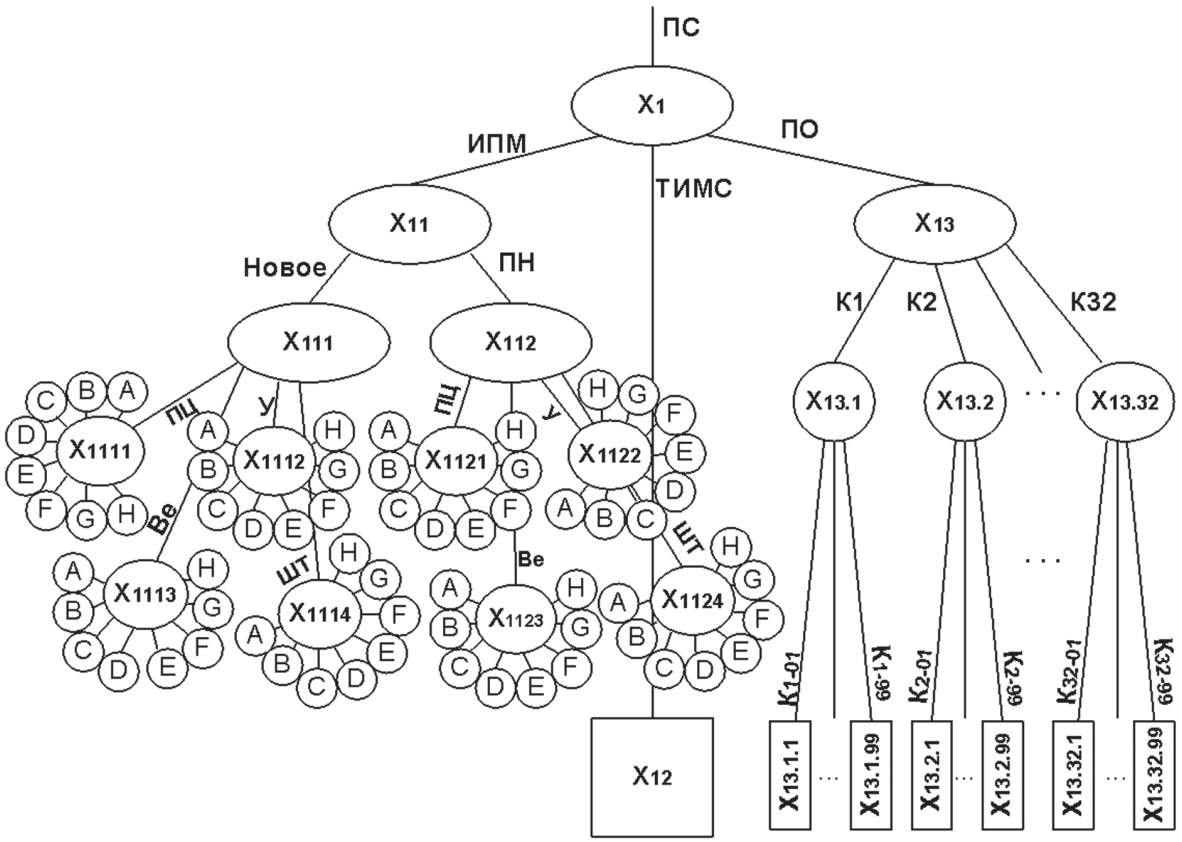

Рис. 3. Часть системы онтологий патентно-конъюнктурной информации.
Ветвь X_1 «Промышленная собственность», где ПС – промышленная собственность, ИПМ – изобретения и полезные модели, ТИМС – топографии интегральных микросхем, ПО – промышленные образцы, Новое – новое изобретение, ПН – применение ранее известного изобретения в новом качестве или новым способом, ПЦ – процесс, У – устройство, Ве – вещество, ШТ – штамм микроорганизма, A, B, C, D, E, F, G, H – классы изобретений в зависимости от сферы применения, К1, К2, ... К32 – классы промышленных образцов в зависимости от сферы применения, К1-01, ..., К1-99, К2-01, ..., К2-99, ..., К32-01, ..., К32-99 – подклассы промышленных образцов

$X^{\overline{E}}_{1112}$, $X^{\overline{F}}_{1112}$, $X^{\overline{G}}_{1112}$, $X^{\overline{H}}_{1112}$ – данные не относятся к заданным классам изобретений.

Выводы

В работе были рассмотрены существующие на сегодняшний момент технологии интеллектуальной обработки данных, и выбран для применения в нашей предметной области универсальный математический аппарат, основанный на алгебре конечных предикатов.

На языке АКП могут быть описаны любые конечные отношения, поэтому для нас это наиболее подходящий математический аппарат, т.к. позволяет легко обнаружить и извлечь из входного корпуса термины и их свойства, отображать многоместные отношения, связывающее текстовую ПКИ.

Была построена модель, представляющая ПКД на языке АКП, которая каждый новый признак разбивает на непересекаемые классы эквивалентности. Она является полной, несократимой и непротиворечивой. Часть данной модели была приведена в работе.

Путь к каждому предикату можно описать формулами, примеры которых приведены в работе.

Список литературы: 1. Оперативная аналитическая обработка данных: концепции и технологии [Электронный ресурс] / Ивановский гос. энергетический ун-т. – Режим доступа URL: http://citforum.ru/seminars/cis99/sch_03.shtml – 2009.– Загл. с экрана. **2.** *Шапот, М.* Интеллектуальный анализ данных в системах поддержки принятия решений [Текст] / М. Шапот // Журн. открытые системы. – 2008. – №1. С. 30-35. **3.** *Гаврилова, Т.А.* Базы знаний интеллектуальных систем [Текст]: учеб. / Т.А. Гаврилова, В.Ф. Хорошевский. – СПб: Питер, 2000.– 384 с. **4.** *Загоровский, И.М.* Выбор алгоритма обучения в системах приобретения знаний из данных [Текст]: материалы 12-ой национал. конф. по искусственному интеллекту с междунар. участием (КИИ 2010), – М.: Физматлит, 2005. – Т. 1. – С. 131-135. **5.** *Калинина, Е.А.* Применение технологии Data Mining для автоматизированного построения баз знаний интегрированных экспертных систем [Текст] / Е.А. Калинина, Г.В. Рыбина.: материалы 8-ой национал. конф. по искусственному интеллекту с междунар. участием (КИИ 2002), – М.: Физматлит, 2002. – Т. 1. – С. 119-127. **6.** *Чубукова, И.А.* Data Mining [Текст] / И.А. Чубукова.– М.: БИНОМ. Лаборатория знаний, Интернет-университет информационных технологий – ИНТУИТ.ру, 2008.– 384 с. **7.** *Корлякова, М.О.* Многоуровневая экспертная система на основе обобщения примеров по признакам [Текст]: материалы 7-ой национал. конф. по искусственному интеллекту с междунар. участием (КИИ 2000), – М.: Физматлит, 2000. – Т. 1. – С.103-112. **8.** *Туманов, В.Е.* Хранилища данных: Жизненный цикл разработки. [Текст] / В.Е. Туманов // Машиностроитель.– 2005.– № 8.– С. 22-30. **9.** *Король, О.І.* Технології побудови систем інтелектуальної обробки патентно-кон'юнктурної інформації [Текст] / О.І. Король // Вісник Херсонського нац. тех. ун-ту.– Херсон.– 2011.– №2 (41).– С. 163-165. **10.** *Бондаренко, М.Ф.* Теория интеллекта [Текст] : учеб. / М.Ф. Бондаренко, Ю.П. Шабанов-Кушнаренко. – Харьков: Компания СМИТ, 2006. – 576 с. **11.** *Шаронова, Н.В.* Автоматизированные информационные библиотечные системы: задачи обработки информации [Текст]: монография, Нар. Укр. Акад. / Н.В. Шаронова, Н.Ф. Хайрова; [Каф. информац. технологий и документоведения]. – Х., 2003. – 120 с. **12.** *Gomez-Perez, A.* Ontological Engineering: what are ontologies and how can we build them [Текст] / O. Corcho, M. Fernandez-Lopez, A. Gomez-Perez, // In Cardoso (ed) Semantic Web: Theory, Tools and Applications.– IDEA Group.– 2007.– Pages 44-70. **13.** *Suárez-Figueroa,* How to write and use the Ontology Requirements Specification Document [Текст] / M.C. Su11 ptrez-Figueroa, A. Gó mez-Pérez, Boris Villazón-Terrazas // Proceedings of the 8th International Conference on Ontologies, DataBases, and Applications of Semantics (ODBASE 2009). – ISBN: 978-3-642-05150-0. LNCS 5871.– Volume: Part II.– 2009.– Pages: 966-982.

Поступила в редколлегию 17.01.2012

УДК 347.77

Інтелектуальна обробка даних при формуванні патентно-кон'юнктурних баз знань / О.І. Король, Н.В. Шаронова // Біоніка інтелекту : наук.-техн. журнал. – 2012. – № 1 (78). – С. 12-16.

В статті розглянуті існуючі методи інтелектуальної обробки інформації. Підтверджено, що найкращим є математичний апарат, заснований на алгебрі кінцевих предикатів. Спроба створити модель інтелектуальної системи, що описує закономірності текстової патентно-кон'юнктурній інформації, представленої на природній мові.

Іл. 3. Бібліогр.: 13 найм.

UDK 347.77

Intellectual data treatment in the formation of patent conjunctural knowledge bases / O.I. Korol, N.V. Sharonova // Bionics of Intelligense: Sci. Mag. – 2012. – № 1 (78). – P. 12-16.

The article reviewed existing methods of intellectual treatment. The best is the mathematical instrument based on the algebra of final predicates algebra. Trying to create a model of intellectual system oriented to description of regularities in textral patent conjunctural information presented in its original language.

Fig. 3. Ref.: 13 items.