УДК 004.822

Н.Ф. Хайрова

Национальный технический университет «ХПИ», Харьков

БИНАРНАЯ ЛОГИЧЕСКАЯ СЕТЬ ИЗВЛЕЧЕНИЯ ЗНАНИЙ ИЗ НЕСТРУКТУРИРОВАННЫХ ПОТОКОВ ТЕКСТОВОЙ ИНФОРМАЦИИ

Сформулированы задачи извлечения знаний для пополнения базы знаний корпоративной информационной системы. Предложено использовать математические средства теории интеллекта и наработки компьютерной лингвистики для увеличения семантической силы моделей представления знаний. Разработана математическая модель отношений между локальной областью исследования менеджера и достоверными глубинными знаниями, представленными в документе. Осуществлена графическая реализация предиката персонификации области знаний менеджера в виде логической сети.

Ключевые слова: *бинарная логическая сеть, корпоративные знания, алгебра конечных предикатов, общее информационное пространство, интеллектуальный потенциал компании, предлингвистический анализ текста.*

Введение

Сегодня корпоративные знания — это многообразная информация, которую необходимо иметь для поддержки основных бизнес-процессов организации для быстрого и адекватного реагирования на различные внешние воздействия. Функции информационного обеспечения корпораций уже оформились в самостоятельную, но при этом недостаточно структурированную и слабо интегрированную в систему управления отрасль. Без специальных технологий учета, регистрации, хранения и мобилизации информационных ресурсов, накопленные опыт и знания не могут быть использованы в полной мере для решения насущных задач управления.

Актуальной становится разработка систем, которые способны не только оперативно извлекать информацию из огромного массива разнородных источников для текущих и перспективных задач управления, но и систем, обладающих возможностью адсорбировать, аккумулировать, анализировать документированный опыт профессионалов, для обеспечения компании полной и релевантной информацией [1]. Информационному обеспечению необходимо придать свойства не только фильтра, концентратора, накопителя и регулятора информационных потоков, но и производителя и поставщика необходимой для работы персонала информации.

Целью статьи является изложение подхода к машинному извлечению знаний из естественно-языковых текстов в рамках общего информационного пространства компании на основе логических и символьных сетей.

Постановка задачи исследования

Общее информационное пространство предприятия представляет собой знания, распределенные по всей организации, образующие полный интеллектуальный потенциал компании. При этом современная корпоративная информационная система (КИС) является многопользовательской и области деятельности менеджера далеко не всегда совпадают с принятым разделением по предметным областям. Поэтому сотрудники любого звена все больше времени тратят на поиск необходимой им информации при принятии того или иного управленческого решения.

Под информационными ресурсами конкретного менеджера понимаются некоторые полученные из общего информационного пространства предприятия сведения, данные, которые обеспечивают удобство принятия решений в области целевой деятельности данного менеджера.

Необходимо разрабатывать модели представления знаний, позволяющие, с одной стороны, расширять размер моделируемой ПО, а, с другой, разбивать ее на области информационных ресурсов отдельных менеджеров корпорации. Для решения поставленной проблемы предлагается использовать достижения в области искусственного интеллекта и компьютерной лингвистики, создавая алгоритмы лингвистического и семантического «понимания» машиной естественного человеческого языка, что в свою очередь, приведет к созданию моделей представления знаний на основе логических и символьных сетей.

Для увеличения семантической силы моделей представления знаний, необходимо локализовать области исследования менеджеров организации в конкретных информационных зонах. Для реализации подобной модели предлагается использовать математические средства теории интеллекта, которые успешно используются для моделирования различных направлений интеллектуальной деятельности человека [2].

Описание математической модели

В качестве базовых средств модели используем средства алгебры конечных предикатов. Вводим универсум элементов U, включающий все возможные текстовые документы, поступающие в корпоративную информационную систему менеджеру на обработку (справки, выписки, отчеты, распоряжения, решения и т.д.), а также понятия и объекты анализа рассматриваемой предметной области, специализированные словари, тезаурусы, отображающие специфику данной предметной области.

Из элементов универсума в соответствие с конкретной задачей обработки информации образуются подмножества M_{1i}, M_{2i}, …, M_{mi}, на декартовых произведениях которых $M_{1i} \times M_{2i} \times \ldots \times M_{mi}$ определяются предикаты P_j, характеризующие работу системы.

Предикатом P, заданным на универсуме будем называть любую функцию, отображающую множество элементом универсума в ноль (предикат тождественно ложный) или единицу (предикат тождественно истинный).

Так как множество элементов универсума информационной системы корпорации, конечно, то и предикат P соответственно конечен.

Базисным для алгебры предикатов является предикат узнавания предмета a по переменной x_i , равный единице, в том случае, если x_i равен a и нулю в противном случае, где i – это любой элемент универсума $(1 \le i \le n)$ [3]:

$$x_i^a = \begin{cases} 1, & \text{если } x_i = a; \\ 0, & \text{если } x_i \neq a. \end{cases}$$

Сущность и результаты исследования

Вводим предметные переменные, определяющие отношение документа, поступающего на обработку в корпоративную информационную систему, к предметной области деятельности менеджера: l – ключевое слово или словосочетание; u – UDK и r – предметная рубрика, поступающего для обработки в КИС документа. Эти переменные отражают суть документа, назначение и взаимосвязь его составляющих, поэтому они объективно определяют истинные и достоверные глубинные знания документа, предоставляющего информацию, необходимую для принятия решений. Значения соответствующие предметных переменных представлены множествами L, U, и R.

В рассматриваемом примере множество ключевых слов и словосочетаний определяется статистико-позиционными методами на этапах предлингвистического анализа [4]:

$$L=\{l^i\},\ 1 \le i \le 14.$$

Иерархическая классификация UDK представлена множеством значений

$$U=\{u^i\},\ 1 \le i \le 5.$$

Рубрикатор представляет собой универсальную иерархическую классификацию областей знаний, принятую для систематизации всего потока научно-технической информации. В примере множество значений рубрикатора представлено

$$R=\{r^i\},\ 1 \le i \le 4.$$

Вводим также, основное для наших рассмотрений, понятие области интеллектуальных знаний менеджера q.

Под областью интеллектуальных знаний конкретного менеджера организации мы понимаем нечетко определенную часть корпоративных знаний, используемую для стандартных управленческих задач данного менеджера. Эта область формируется в сфере мышления и имеет внеязыковую природу, но поскольку мысль не может существовать вне слова, то под областью интеллектуальных знаний менеджера мы будем подразумевать лексическую единицу, представляемую некоторым словом или словосочетанием, и выражающую определенное множество управленческих ситуаций.

Можно выразить область интеллектуальных знаний менеджера через значения предметных переменных r, l, и u :

$$r^1u^1l^{12}=q^1;\ r^4u^2l^{12}=q^2;\ r^1u^1l^{13}=q^3;$$
$$r^1u^1l^{14}=q^4;\ r^1u^3l^1=q^5;\ r^2u^4l^1=q^6;$$
$$r^1u^3l^2=q^7;\ r^1u^3l^3=q^8;\ r^1u^3l^4=q^9;$$
$$r^2u^4l^4=q^{10};\ r^2u^4l^5=q^{11};\ r^2u^4l^6=q^{12};$$
$$r^2u^5l^7=q^{13};\ r^2u^5l^8=q^{14};\ r^1u^3l^8=q^{15};$$
$$r^4u^2l^9=q^{16};\ r^4u^2l^{10}=q^{17};\ r^4u^2l^{11}=q^{18}.$$

Выполняем операцию почленной дизъюнкции возможно большего числа родственных равенств. Родственными равенствами мы будем называть такие равенства, которые после выполнения над ними операции почленной дизъюнкции приводят к равенствам с левой частью в виде логического произведения, каждый сомножитель которого зависит от одной предметной переменной [1]. Введение почленной дизъюнкции с использованием родственных равенств обусловлено необходимостью получения локальных областей интеллектуальных знаний менеджера, определяемые именем конкретного менеджера.

Такие области могут включать больше чем одно исчисляемое ограниченное количество рубрик и предметных областей исследований.

$$r^1u^1(l^{12} \vee l^{13} \vee l^{14}) = q^1 \vee q^3 \vee q^4;$$
$$r^4u^2(l^{12} \vee l^9 \vee l^{10} \vee l^{11}) = q^2 \vee q^{16} \vee q^{17} \vee q^{18};$$

$$r^2u^5(l^7 \vee l^8) = q^{13} \vee q^{14}r^1u^3(l^1 \vee l^2 \vee l^3 \vee l^4 \vee l^8) =$$
$$= q^5 \vee q^7 \vee q^8 \vee q^9 \vee q^{15};$$

$$r^2u^4(l^1 \vee l^4 \vee l^5 \vee l^6) = q^6 \vee q^{10} \vee q^{11} \vee q^{12}.$$

Формируем функцию перехода от предметной области интеллектуальных знаний q к локальной области исследования менеджера m, в профессиональную деятельность которого входит данная область исследования q.

И переопределяем зависимость локальной области исследования менеджера m от предметных переменных r, *l*, u:

$$m^1 = r^1u^1l^{12} \vee r^1u^1l^{13} \vee r^1u^1l^{14} \vee r^2u^5l^7 \vee r^2u^5l^8 =$$
$$= r^1u^1(l^{12} \vee l^{13} \vee l^{14}) \vee r^2u^5(l^7 \vee l^8);$$

$$m^2 = r^4u^2l^{12} \vee r^4u^2l^9 \vee r^4u^2l^{10} \vee r^4u^2l^{11} =$$
$$= r^4u^2(l^{12} \vee l^9 \vee l^{10} \vee l^{11});$$

$$m^3 = r^1u^3l^1 \vee r^1u^3l^2 \vee r^1u^3l^3 \vee r^1u^3l^4 \vee r^1u^3l^8 =$$
$$= r^1u^3(l^1 \vee l^2 \vee l^3 \vee l^4 \vee l^8);$$

$$m^4 = r^2u^4l^1 \vee r^2u^4l^4 \vee r^2u^4l^5 \vee r^2u^4l^6 =$$
$$r^2u^4(l^1 \vee l^4 \vee l^5 \vee l^6).$$

Предикат локализации области знаний менеджера P описывает связь локальной области исследования менеджера и переменных, объективно определяющих, глубинные знания документа:

$$P(r, l, a, u, m) =$$
$$= m^1r^1u^1(l^{12} \vee l^{13} \vee l^{14}) \vee m^1r^2u^5(l^7 \vee l^8) \vee$$
$$\vee m^2r^4u^2(l^{12} \vee l^9 \vee l^{10} \vee l^{11}) \vee$$
$$\vee m^3r^1u^3(l^1 l^2 \vee l^3 \vee l^4 \vee l^8) \vee m^4r^2u^4(l^1 \vee l^4 \vee l^5 \vee l^6).$$

Можно определить бинарные отношения — предикаты, связывающие переменную локальной области исследования менеджера и предметные переменные r, *l*, u , и отобразить их в виде двудольных графов.

Например, бинарный предикат P_l, определяющий отношения переменной m и предметной переменной *l*:

$$P_l(l,m) =$$
$$= (l^{12} \vee l^{13} \vee l^{14} \vee l^7 \vee l^8)\, m^1 \vee (l^{12} \vee l^9 \vee l^{10} \vee l^{11})\, m^2 \vee$$
$$\vee (l^1 \vee l^2 \vee l^3 \vee l^4 \vee l^8)\, m^3 \vee (l^1 \vee l^4 \vee l^5 \vee l^6)\, m^4$$

Данное отношение $P_l(l,m)$ можно отобразить в виде двудольных графов (рис. 1):

Таким образом, построена математическая модель локализации области исследования менеджера, характеризующаяся системой бинарных отношений и отображаемая двудольными графами. Образуя конъюнкцию предикатов, получим предикат модели P, связывающий между собой предметные переменные r, *l*, u, m:

$$P(r, l, a, u, m) = P_r(r,m) \wedge P_l(l,m) \wedge P_u(u,m).$$

Предикат P можно наглядно изобразить в виде логической сети (рис. 2), которая является графическим представлением результата бинарной декомпозиции многоместного предиката. Логическая сеть состоит из полюсов и ветвей. Каждому полюсу логической сети ставится в соответствие своя предметная переменная модели, которая называется атрибутом этого полюса. С каждым полюсом связана область изменения атрибута полюса, т.е. его домен. Любой полюс логической сети в каждый момент времени несет какое-то знание о значении своего атрибута. Оно представляет собой одно из подмножеств домена полюса. Указывая состояние всех полюсов в данный момент времени, определяем состояние сети в этот момент времени.

Выводы и перспективы дальнейших исследований

Таким образом, разработана математическая модель отношений между областью интеллектуальных знаний менеджера и достоверными глубинными знаниями, представленными в документе. Использование данной модели позволяет извлекать новые понятия и отношения, связывающие данные понятия, т.е. дополнять новой информацией базы знаний корпоративной информационной системы, одновременно увеличивая семантическую силу модели представления знаний КИС.

Графическая интерпретация модели показывает возможность в дальнейшем реализовать ее аппаратно. Логическую сеть можно превратить в электронную схему для автоматического решения неко-

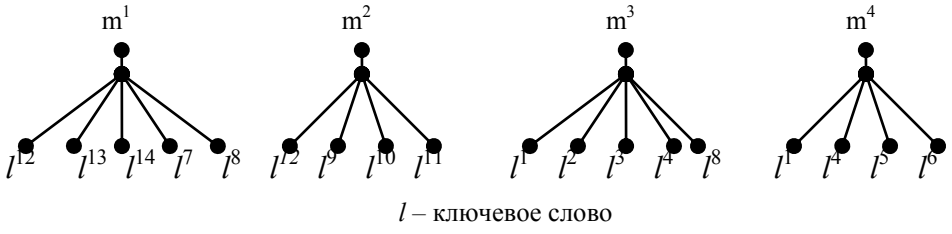

Рис. 1. Двудольный граф предиката $P_l(l,m)$

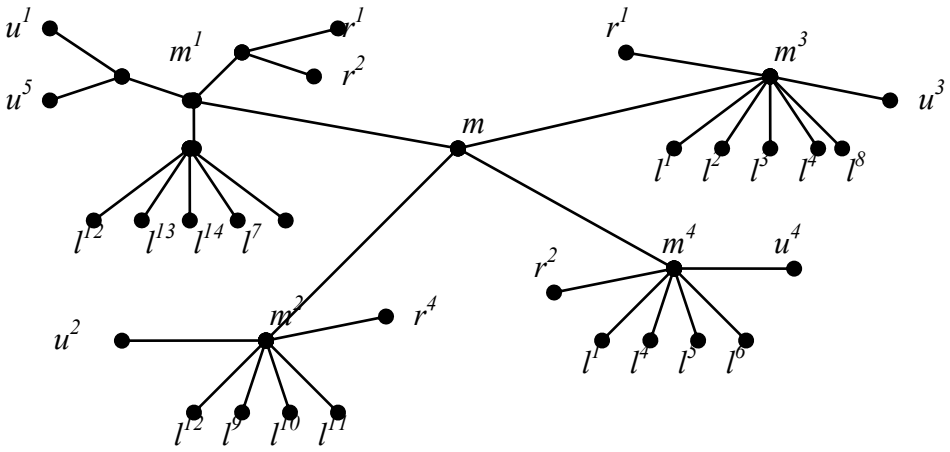

Рис. 2. Логическая сеть, связывающая предметные переменные r, u, *l,* m

торого класса задач, определяемого той моделью, для которой была построена данная сеть [5]. Ее можно изготовить в виде сверхбыстродействующей карты, осуществляющей параллельную обработку информации, и устанавливаемую на материнской плате персонального компьютера.

По мере надобности программа, управляющая работой компьютера, обращается к той или иной карте, которая за доли микросекунды формирует ответ на тот или иной запрос, персональный компьютер при этом из машины последовательного действия превращается в машину последовательно-параллельного действия, значительно повышая производительность своей работы.

Список литературы

1. Батюк А.Є. Інформаційні системи в менеджменті / А.Є. Батюк, З.П. Двуліт, К.М. Обельовська, І.М. Огородник, Л.П. Фарбі. – Львів: Національний університет „Львівська політехніка”, „Інтелект-Захід”, 2004. – 520 с.

2. Бондаренко М.Ф. Теория интеллекта / М.Ф. Бондаренко, Ю.П. Шабанов-Кушнаренко – Х.: Изд-во «СМИТ», 2007. – 576 с.

3. Шабанов-Кушнаренко Ю.П. Теория интеллекта: Технические средства / Ю.П. Шабанов-Кушнаренко. – Х.: Вища школа, 1986. – 134 с.

4. Nina Khairova, Natalia Sharonova. Use of Predicate Categories for Modelling of Operation of the Semantic Analyzer of the Linguistic Processor./Proceedinga of IEEE EAST-West Design & Test Symposium (EWDTS'09). // Moscow, Russia, September 18-21, 2009 – P. 204- 207.

5. Хаханов В.И. Проектирование и тестирование цифровых систем на кристаллах./ В.И. Хаханов, Литвинова Е. И., Гузь О. А. – Харьков: ХНУРЭ. – 2009. – 484 с.

Поступила в редколлегию 21.11.2012

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

БІНАРНА ЛОГІЧНА СІТКА ВИТЯГАННЯ ЗНАНЬ З НЕСТРУКТУРОВАНИХ ПОТОКІВ ТЕКСТОВОЇ ІНФОРМАЦІЇ

Н.Ф. Хайрова

Сформульовано задачі витягання знань для поповнення бази корпоративної інформаційної системи. Запропоновано використовувати математичні засоби теорії інтелекту та здобутки комп'ютерної лінгвістики для збільшення семантичної сили моделей подання знань. Розроблено математичну модель відношень між локальною сіткою областю досліджень менеджера та вірогідними глибинними знаннями, представленими у документі.

Ключові слова: *бінарна логічна сітка, корпоративні знання, алгебра кінцевих предикатів, загальний інформаційний простір, інтелектуальний потенціал компанії, предлінгвістичний аналіз тексту.*

BINARY LOGICAL NET FOR KNOWLEDGE EXTRACTION FROM UNSTRUCTURED TEXT INFORMATIONFLUSSES

N.F. Khairova

It is formulated tasks of knowledge retrieval for corporative information system knowledge base completion. It is suggested to use mathematical techniques of intelligence theory and computer linguistics to increase semantic power of knowledge representation models. It is developed mathematical model for relations between manager local research area true deep knowledge given in the document. It is carried out graphical realization for predicate of manager knowledge area personification as logic network.

Keywords: *binary logic network, corporative knowledge, finite predicate algebra, general information area, company intelligence potential, pre-linguistic text analysis.*