ИНТЕЛЛЕКТУАЛЬНЫЕ ТЕХНОЛОГИИ ИДЕНТИФИКАЦИИ ФАКТОГРАФИЧЕСКОЙ ИНФОРМАЦИИ

УДК 004.41:47; 347.77

ДОРОШЕНКО Анастасия Юрьевна
аспирант Национального технического университета «Харьковский политехнический институт».
Научные интересы: интеллектуальные системы, компьютерная лингвистика, автоматизированная обработка текстовой информации.
e-mail: marykate90@mail.ru

ОРОБИНСКАЯ Елена Александровна
аспирант Национального технического университета «Харьковский политехнический институт» и университета им. Люмьер Лион-2 (Лион, Франция).
Научные интересы: информационные системы, математическое моделирование.

Аджит Пратап Сингх Гаутам
аспирант кафедры Интеллектуальных компьютерных систем НТУ «Харьковский политехнический институт».
Научные интересы: интеллектуальная обработка данных, знания в корпоративных информационных системах.

ВВЕДЕНИЕ

Вызов сегодняшнего дня, обусловленный бурным ростом текстовых хранилищ в глобальной сети Интернет и технологическими проблемами их автоматической обработки, заключается в том, чтобы научить сами информационные системы (ИС) обнаруживать и правильно интерпретировать полезную информацию, предоставляемую текстом. В процессе работы над созданием современных компьютерных систем, решающих интеллектуальные задачи (в частности, понимание текстов на естественном языке), на первый план выдвигается проблема идентификации и извлечения знаний. Анализ любой текстовой информации, а особенно фактографической, и извлечение из полнотекстовых документов фактов является актуальной задачей технологии фактографического поиска, предлагаемый подход, основанный на представлении содержания текста в форме семантической сети, позволяет искать факт в семантической сети определенного текста интеллектуальных технологий.

Решение задачи обеспечения пользователей релевантной информацией в системе поиска и обработки определяется в основном правильным подбором инструментов делового анализа. Но немаловажным является и выбор инструментов поддержки процессов извлечения, преобразования, загрузки и хранения данных.

ПОСТАНОВКА ЗАДАЧИ

Построение семантической сети фактографической информации интеллектуальных технологий на основе анализа и обработки текстов, а также решение задачи с помощью алгебры предикатов.

РЕШЕНИЕ ЗАДАЧИ

Знакомство с реальной ситуацией в информационном обслуживании показывает, что фактографическую информацию обычно сознательно или бессознательно трактуют просто как конкретные сведения или данные независимо от того, являются ли они фактическими или прогнозируемыми. Главное, что эти сведения сообщают о какой-то предметной области,

а не о документах, посвященных этой области. Исходя из такого понимания, фактографическую информацию можно классифицировать следующим образом:

1) фактическая и прогнозная (гипотетическая) информация;

2) количественная и качественная фактографическая информация;

3) хорошо структурированная фактографическая информация и плохо структурированная фактографическая информация.

К хорошо структурированным сведениям об ЭВМ относятся, прежде всего, сведения количественного характера, а также качественные (словесно выраженные) сведения, имеющие хорошо регламентированную форму: параметры оборудования и их значения и т. п. К плохо структурированным относятся сведения, представленные разнообразными нерегламентированными словесными инструкциями, т. е. различные описания отдельных фактов, изложение концепций и теорий, сделанных на естественном языке [2,3].

Существует два класса методов представления знаний – естественные и искусственные. Первый класс методов – это методы представления знаний предназначенные для использования их человеком. К ним относятся, прежде всего, метод представления знаний на естественном языке в текстовой (t) и аудиальной (s) форме, графический метод (g) в виде картин, рисунков, чертежей, графиков, диаграмм, а также визуальный метод (v) в виде кино, анимации, балета, пантомимы, жестов и т.д. Второй класс методов представления знаний это методы представления знаний, адресованные искусственным компьютерным системам – так называемым интеллектуальным информационным системам (ИИС).

1. Определение семантической сети

В общепринятом смысле под семантической сетью понимается модель представления знаний посредством сети узлов, связанных дугами, где узлы соответствуют понятиям или объектам, а дуги – отношениям между узлами [5].

Научной основой построения семантических сетей является теория графов. Семантические сети представляют знания в виде графовой структуры, которая является более наглядной и естественной по сравнению с другими структурами знаний. Решаемая задача использует структуру, моделирующую семантические связи, которые мы используем для получения одних фактов на основе других [5]. Построение графа помогает находить противоречия в знаниях, а также выявлять недостающие фрагменты знаний. Представление фрагментов знаний рассматривается как участок семантической сети и базируется на понятиях фреймов Минского [99] и сценариев Шенка [168].

Среди особенностей семантических сетей можем выделить:

- описание объектов ПрО (полной семантической сети) осуществляется средствами естественного языка;
- все факты, включая и вновь поступившие, накапливаются в относительно однородной структуре памяти;
- на сетях определяют ряд унифицированных семантических отношений между объектами и соответственно унифицированные методы вывода;
- структурное представление семантических знаний позволяет определить на них дополнительную семантику, определяющую относительную силу семантических связей, облегчающую процесс вывода в сетях.

2. Семантические и грамматические особенности объектно-признаковых языков

Семантические и грамматические особенности объектно-признаковых языков (ОПЯ) определяются необходимостью фиксировать связь «объект - признак - значение». При таком подходе в качестве алфавита ОПЯ выступает алфавит естественного языка, цифры, специальные символы, а в качестве лексических единиц — слова и словосочетания.

В составе лексики ОПЯ можно выделить три основных лексико-семантических класса названий: объектов, признаков и значений признаков.

Объектами являются основные единицы (изделия, материалы технологические процессы и т. п.) фактографического поиска описываемые с помощью совокупности пар «признак – значение». В каждой области знаний система объектов, естественно, своя.

Все присущие объектам фактографического поиска признаки делятся на количественные и качественные.

Количественные признаки – это именованные и неименованные числа.

Качественные признаки – это признаки, значение которых выражается описательно, словесно.

Разновидностью качественных признаков являются признаки наличия свойства и признаки степени свойства.

В качестве признаков наличия свойства выступают слова да и нет и различные вариации: есть, не был, был, отсутствие, наличие.

В качестве признаков степени свойства выступают слова: слабый, средний, сильный, интенсивный, малоинтенсивный и т. п.

Особенности лексики ОПЯ обусловлены как спецификой фактографического поиска, триадностью фактографической информации, так и источниками отбора лексических единиц: использование массивов вторичных документов для отбора лексики является неприемлемым. В качестве источников для отбора лексики в ОПЯ служат первичные документы, практически все виды научной, технической и производственной литературы. Особое значение при этом отводится техническим каталогам, прейскурантам, научно-технической документации, адресным книгам, деловой переписке, отчетно-статистической документации и другим источникам, содержащим хорошо структурированную фактографическую информацию.

3. Технология фактографического поиска

Технология фактографического поиска основана на представлении содержания текста в форме семантической сети. Семантическая сеть содержит значимые слова и словосочетания, упоминавшиеся в тексте, которые связанны друг с другом различными типами синтактико-семантических связей. Элементарная семантическая сеть представляет результат синтаксического анализа и постсинтаксических трансформаций дерева синтаксических зависимостей между словами в отдельном предложении. Полная семантическая сеть текста есть совокупность отдельных семантических сетей, соответствующих предложениям.

Поиск факта есть поиск в семантической сети текста такой подсети, которая изоморфна одному из шаблонов. Если подсеть найдена факт считается установленным, после чего производится извлечение сущностей и их маркировка ролями, заданными в соответствующих узлах лингвистических описаний.

Таким образом, результатом поиска является имя (типа) факта и набор указателей на сущности семантической сети с указанием соответствующих им ролей в лингвистическом описании.

Пример семантической сети для фактографической информации с введенными признаками:

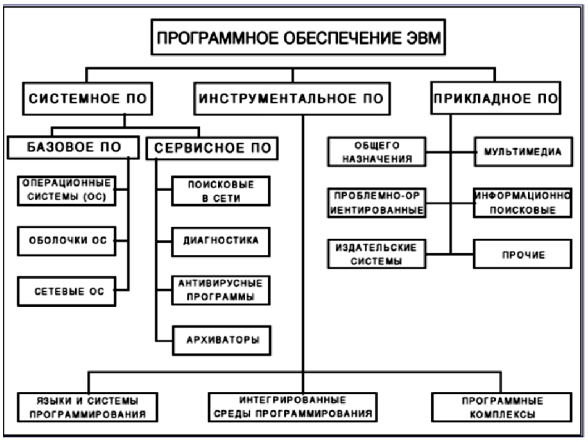

Рисунок 1 – ПО ЭВМ

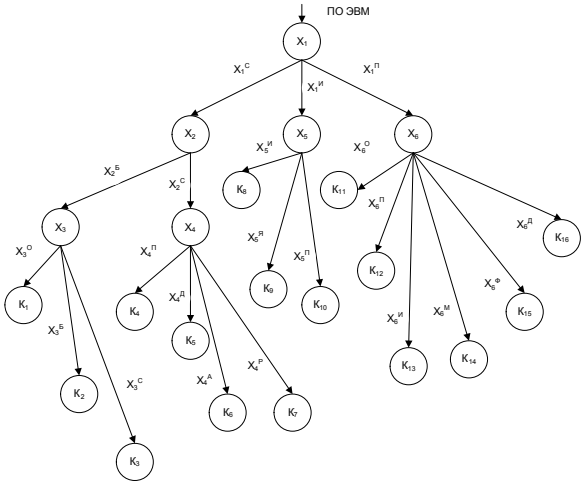

Рисунок 2 – Семантическая сеть ПО ЭВМ

Алгебра предикатов дает возможность описывать функции интеллекта в виде предикатных уравнений.

Введенные признаки:

X_1 – вид ПО, X_2 – вид системного ПО, X_3 – вид базового ПО, X_4 – вид сервисногоПО,

X_5 – вид інструментального ПО, X_6 – вид прикладного ПО, $K_{1\text{-}16}$ – возможные варианты и т.д.

$$
\begin{array}{lll}
K_1 = X_3^{О} \cdot X_2^{Б} \cdot X_1^{С} = 1, & K_6 = X_4^{А} \cdot X_2^{С} \cdot X_1^{С} = 1, & K_{11} = X_6^{О} \cdot X_1^{П} = 1, \\
K_2 = X_3^{Б} \cdot X_2^{Б} \cdot X_1^{С} = 1, & K_7 = X_4^{Р} \cdot X_2^{С} \cdot X_1^{С} = 1, & K_{12} = X_6^{П} \cdot X_1^{П} = 1, \\
K_3 = X_3^{С} \cdot X_2^{Б} \cdot X_1^{С} = 1, & K_8 = X_5^{И} \cdot X_1^{И} = 1, & K_{13} = X_6^{И} \cdot X_1^{П} = 1, \\
K_4 = X_4^{П} \cdot X_2^{С} \cdot X_1^{С} = 1, & K_9 = X_5^{я} \cdot X_1^{И} = 1, & K_{14} = X_6^{М} \cdot X_1^{П} = 1, \\
K_5 = X_4^{Д} \cdot X_2^{С} \cdot X_1^{С} = 1, & K_{10} = X_5^{П} \cdot X_1^{И} = 1, & K_{15} = X_6^{Ф} \cdot X_1^{П} = 1, \\
\multicolumn{3}{c}{K_{16} = X_6^{Д} \cdot X_1^{П} = 1, .}
\end{array}
\tag{1}
$$

$$
X_1^{С} \vee X_1^{И} \vee X_1^{П} = 1,\ X_1^{С} \vee X_1^{И} = 0,\ X_1^{С} \vee X_1^{П} = 0,\ X_1^{И} \vee X_1^{П} = 0,\ X_1^{С} \vee X_1^{\overline{С}} = 0,\ X_1^{И} \vee X_1^{\overline{И}} = 0,\ X_1^{П} \vee X_1^{\overline{П}} = 0. \tag{2}
$$

ОСНОВНЫЕ РЕЗУЛЬТАТЫ И ВЫВОДЫ

Знание не такое определенное понятие, как факт. Оно лишь ограничивает множество возможных состояний мест предметного пространства. Поиск факта есть поиск в семантической сети текста такой подсети, которая изоморфна одному из шаблонов. Если подсеть найдена, факт считается установленным, после чего производится извлечение сущностей и их маркировка ролями, заданными в соответствующих узлах лингвистических описаний [1,4]. Таким образом, результатом поиска является имя факта и набор указателей на сущности семантической сети с указанием соответствующих им ролей в лингвистическом описании. Построена семантическая сеть фактографической информации интеллектуальных технологий на основе анализа и обработки текстов, а также с помощью алгебры предикатов.

ЛИТЕРАТУРА:

1. Orobinskaya E.A. Yazykovaya kompetentsiya infopmatsionnyh sistem /E.A. Orobinskaya, O.I. Kopol', N.V. Sharonova //Visnik Natsional'nogo tehnichnogo ynivepcitety «Harkivc'kiy politehnichniy instityt». Problemi informatiki i modelyuvannya. – H.: NTU «HPI», 2012.
2. Fedotov N.N. Credctva informatsionnogo obespecheniya avtomatizirovannyh sistem ypravleniya /N.N. Fedotov, L.B. Venchkovckiy. – M.: Izd-vo standartov, 1989. – 192 s.
3. Bondarenko M.F. Teoriya intellekta: ycheb. /M.F. Bondarenko, Yu.P. Shabanov-Kyshnarenko. – Har'kov: Kompaniya SMIT, 2006. – 576 s.
4. Sharonova N.V. Avtomatizirovannye informatsionnye bibliotechnye sistemy: zadachi obrabotki informatsii: monografiya. /N.V. Sharonova, N.F. Hayrova. – Har'kov, 2003. – 120 s.
5. Zybov A. V. Osnovy iskysstvennogo intellekta dlya lingvistov /A.V. Zybov, I.I. Zybova. – M.: Univercitetckaya kniga; Logos, 2007. – 320 s.
6. Ermakov A.E. Avtomatizatsiya ontologicheskogo inzhiringa v sistemah izvlecheniya znaniy iz teksta /A.E. Ermakov //Trudy Mezhdynapodnoy konferentsii Dialog'2008. – Moskva: Nayka, 2008. – S.136-140.
7. Kanischeva O. V. Ispol'zovanie algebry predikatnyh operatsiy dlya opisaniya estestvenno-yazykovyh otnosheniy /O.V. Kanischeva //Informatsiyni tehnologiï: nayka, tehnika, tehnologiya, osvita, zdopov'ya : materiali XVII mizhnar. nayk.-prakt. konf. – Harkiv: NTU «HPI», 2009. – S.16.
8. Aliseyko Z.A. Ispol'zovanie algebry predikatov i predikatnyh operatsiy dlya formalizatsii deklarativnoy i protsedyrnoy sostavlyayuschih znaniy /Z.A. Aliseyko, V.I. Bylkin, O.V. Kanischeva, N.V. Sharonova //Bionika intelekta. – Harkiv: HNURE, 2006. – №1 (64). – S.59-63.
9. Amamiya M. Arhitektypa EVM i iskysstvennyj intellekt /M. Amamiya, Yu. Tanaka. – M.: Mir, 1993. – 400 s.
10. Bondarenko M.F. O mozgopodobnyh EVM /M.F. Bondarenko, Z.V. Dydar', I.A. Efimova, V.A. Leschinskiy, S.Yu. Shabanov-Kyshnarenko //Radioelektronika i informatika. – Har'kov: HNYRE, 2004. – №2. – S.89-105.
11. Bylkin V.I. Matematicheckie modeli znaniy i ih realizatsiya c pomosch'yu algebropredikatnyh stryktyr /V.I. Bylkin, N.V. Sharonova: monogpafiya. – NTU «HPI», MEGI.: Donetsk, 2010. – 304 s.
12. Sussna M. Word sense disambiguation for free-text indexing using a massive semantic network //The Second International Conference on Information and Knowledge Management. – 2008. – P.67-74.

Рецензент: *д.т.н., проф. Шаронова Н.В., НТУ «Харьковский политехнический институт».*