ОБ ОДНОМ ПОДХОДЕ К ФОРМИРОВАНИЮ БАЗЫ ЗНАНИЙ

Парамонов А.И.
Донецкий национальный университет, г. Донецк
кафедра компьютерных технологий
E-mail: anton@paramonov.info

Abstract

***Paramonov A.I. About one approach to knowledge base construction**. The knowledge acquisition and representation problem is considered. In article approach to knowledge extraction is proposed. The knowledge extraction formalization mechanism is offered. Program realization is described.*

Введение.

Создание интеллектуальных систем предполагает этапы построения и формирования базы знаний (БЗ) для описания понятий соответствующей предметной области на языке экспертов. Особенно актуальна задача построения мировой модели для систем интерпретации текстовой информации, в том числе для интеллектуальных поисковых машин. Термины, которые ассоциированы со значимыми понятиями предметной области, играют большую роль в методах обработки и анализа текстов на естественно языке (ЕЯ). Мера адекватности представления знаний очень сильно зависит от сформированной модели «окружения». Понимание проблемно-ориентированного ЕЯ-текста основано не только на знании множества терминов, но и на знании семантической структуры терминов и семантических отношений между ними. Описанная в [1] нечеткая гибридная модель представления и обработки знаний основана на сложной многоуровневой БЗ некоторой предметной области.

Проблема формирования БЗ рассматривается как процесс приобретения знаний, который включает в себя не только извлечение специфических знаний о предметной области, но и интерпретацию извлеченных данных применительно к некоторой концептуальной оболочке и формализацию их таким способом, чтобы программа могла действительно использовать их в процессе работы. На сегодняшний день функция приобретения знаний является одним из главных «узких мест» технологий экспертных систем [2, 3].

Существуют теории, на которых базируется методика приобретения и представления знаний. Методология извлечения знаний разделяется на три основные группы: использование опроса экспертов для извлечения знаний, автоматизация процесса извлечения знаний и приобретение новых знаний на основе существующих. Оптимальным вариантом считается совместное использование методов из различных групп. Что способствует увеличению количественного и качественного наполнения БЗ, и к тому же соответствует природе приобретения знаний. Однако анализ работ показал, что такая организация процесса формирования БЗ требует учета многих аспектов проблемы и согласования методов между собой. Эту задачу можно вынести в отдельную проблему, поэтому зачастую многие экспертные системы разрабатываются на основе отдельных методик.

В работе рассмотрен подход к формированию БЗ гибридной модели на основе метода опроса экспертов. Определены основные принципы построения системы опроса. Представлен механизм формализации извлеченных знаний в терминах гибридной модели для использования их в интеллектуальных системах.

Гибридная модель (ГМ) включает семантическую сеть для представления знаний об объектах, семантическую сеть для представления знаний о действиях, пропозициональную сеть

для представления знаний о событиях. В указанных сетях представлены уникальные знания о прототипах объектов, действий и событий, соответственно. Рабочими элементами модели выступают: объект и действие. В мировой модели рабочие элементы представлены терминами. Термин – слово или словосочетание, обозначающее понятие специальной области знания или деятельности. Термин входит в конкретную лексическую систему языка, но лишь через посредство конкретной терминологической системы. Терминология – совокупность терминов определенной отрасли знания или производства, а также учение об образовании, составе и функционировании терминов [4]. Терминология выступает неким абстрактным образом предметной области, оторванным от реальности, т.е. без учета возможных отношений терминов. Семантические сети выступают хранилищем отношений между терминами в мировой модели. Взаимодействия терминов сохраняются в виде пропозиций и пропозициональных сетей.

Методика опроса экспертов для извлечения знаний.

Из всех существующих методов извлечения терминов в системе формирования БЗ гибридной модели за базовый выбран метод опроса экспертов. Использование этого метода имеет свои преимущества и недостатки, а его реализация затрагивает несколько важных аспектов работы с экспертами. Основным преимуществом опроса является то, что множество терминов будет представлено в БЗ наиболее полно и в тоже время не избыточно. Полноту множества гарантирует и тот факт, что опрос будет проводиться с несколькими экспертами, которые могут дополнить и/или поправить друг друга. Явный недостаток такого метода заключается в трудоемкости процесса. Особенно временные затраты возрастают при формировании полноценной БЗ. В терминах гибридной модели полноценность понимается как наполнение каждого термина (объекта гибридной модели) неким набором признаков, а не только описание его в виде «имени» («знака»).

При использовании опроса следует учесть такой немаловажный момент, как конструирование личностных опросников. Это отдельная задача в когнитивной психологии. Специалисты утверждают, что формировать опрос необходимо так, чтобы максимально избегать таких явлений как установка, предубеждение и дискриминация [5, 6]. Иными словами, первое – нельзя эксперта подталкивать к определенным ответам, и второе – необходимо стараться избавиться от эмоциональной окраски терминов и от стереотипов. Первая часть решается путем постановки корректных и нейтральных вопросов. Вторая сглаживается за счет учета мнений нескольких экспертов, тем самым эмоциональный всплеск одного человека кардинально не изменяет общий характер термина.

Совместные знания экспертов в работе названы «коллективным разумом». Групповая работа имеет очевидное преимущество над индивидуальной. Снижается вероятность ошибок, неизбежных при индивидуальной работе. Психологи утверждают, что важным фактором групповой продуктивности является размер группы. В рамках разрабатываемой системы немаловажным является и «качество» группы, то есть квалификация составляющих ее экспертов. В довершение следует отметить, что опрос экспертов необходимо делать независимо и раздельно, чтобы избежать явлений конформизма и уступчивости [6].

Приведенное понятие «коллективного разума» задействовано на всех этапах формирования БЗ гибридной модели от выделения терминов до построения схем.

Главную опасность для понимания термина, как и любого слова, представляет собой не многозначность, вполне естественная для языка, в том числе и для языка науки, а двусмысленность, то есть неясность того, какое из значений имеется в виду в данном контексте. Подобного рода вопросы появляются не только при автоматическом понимании, но и перед любым экспертом-человеком. Это явление закономерно, но вероятность его появление необходимо минимизировать. Разрешение этого вопроса зависит от установленных отношений

между терминами в рамках рассматриваемого контекста. Данное замечание показывает важность выделения связей между терминами в мировой модели.

Поскольку БЗ формируется на основе знаний экспертов-людей, то возникают ситуации, когда кроме денотативного значения слова применяется и его коннотативное значение [6]. Особенно это актуально для научных и узкоспециализированных областей. Данному аспекту следует уделять внимание на этапах выделения терминов и установления связей между ними. Чаще всего коннотативные термины присутствуют в БЗ отдельно от денотативных терминов с таким же обозначением (одинаково пишущиеся). Однако если присутствуют оба таких термина, то возникают проблемы схожие на отношения омонимии, приводящие к двусмысленности.

Рассмотренные аспекты методики опроса экспертов показывают сильную зависимость данной группы методов от человеческого фактора. Означенные проблемы и принципы решаются в работе на программном уровне, описание которого приведено далее.

Механизм формализации извлеченных знаний.

Полученные от экспертов знания необходимо преобразовать к виду, в котором они будут использованы программой. Процесс опроса сформирован таким образом, чтобы наполнение БЗ было согласовано с элементами и уровнями модели, в терминах которой будут представляться знания.

При формировании БЗ гибридной нечеткой модели выделены следующие этапы:

- определение терминов предметной области;
- установление связей между ними;
- построение схем возможных взаимодействий (событийная модель).

На каждом этапе в свою очередь выделяются подзадачи.

Задача формирования терминологического словаря является, пожалуй, наиболее изученной из всех выделенных этапов [4, 7, 8]. Терминоведение является одним из активно развивающихся направлений современных научных исследований. Оно тесно связано с широким кругом междисциплинарных проблем и в последнее время вызывает все больший интерес у лингвистов и у специалистов в области информационных технологий.

Отличие рабочих элементов гибридной модели от значений терминологического словаря в том, что узел сети, представляющий некоторый термин, описан как нечеткое подмножество признаков [1]:

$$\underset{\sim}{P} = \{(x_i \mid \mu_{\underset{\sim}{P}}(x_i))\}, \quad (1)$$

где x_i – признак узла сети ГМ, $\mu_{\underset{\sim}{P}}(x_i)$ – функция принадлежности признака узлу $\underset{\sim}{P}$.

Таким образом, выделение терминов предметной области разбивается на две подзадачи: определение терминологического словаря и описание каждого термина в виде набора признаков. Реализация второй подзадачи не является обязательной. Однако следует учесть, что в данном случае знания будут представлены в модели не полностью. А это, соответственно, скажется на качестве работы системы, построенной на основе описываемой модели.

Все термины в БЗ имеют связи с остальными терминами. Все связи между терминами именованы. Связи делятся на две группы по именам: «часть» и «это».

Сети в гибридной модели представлены подграфами вида:

$$N(V, A), \quad (2)$$

где V – множество узлов сетей (множество подмножеств $\underset{\sim}{P}$), A – множество дуг сетей (именованные связи между терминами).

Помимо имен связи имеют значимости. Значимость (или вес) связи определена как поток. Потоком в сети N названо нечеткое множество $\underset{\sim}{\Gamma}$.

$$\underset{\sim}{\Gamma} = \{a_{ij} \mid \varphi(a_{ij})\}; \quad (3)$$

Число $\varphi(a_{ij}) \in [0,1]$ называется потоком по дуге из узла i в узел j.

Связи «часть» соединяют термины, один из которых является логически составной частью другого, в тоже время, являясь самостоятельным термином. На рисунке 1 данными отношениями связаны пары терминов 8-9 и 8-10. Вес дуги именованной как «часть» показывает важность наличия в составном термине связанного этой дугой термина. Природа этих отношений наглядна и проста в восприятии.

Более интересным представляется тип связей с наименованием «это». Связи подобного рода описывают такие аспекты отношений между терминами (объектами гибридной модели) как: обобщение, омонимия, синонимия, эмоционально близкие слова. Связи представляющие отношение обобщения формируют некоторую виртуальную иерархию терминов. Виртуальную в том смысле, что в семантической сети, которой описаны термины предметной области, нет иерархии, и каждый узел имеет связь с любым другим. В качестве примера обобщения можно использовать представление отношений между организмами в эволюционной теории Ч. Дарвина. На рисунке 1 термин 3 является обобщением терминов 5 и 6, а термин 1 в свою очередь является обобщением терминов 2, 3 и 4. Поток между данными отношениями указывает меру близости дочерних терминов с обобщающим, то есть насколько достоверна будет замена одного термина другим.

Связи представляющие омонимию порождают и одновременно решают описанную выше ситуацию двусмысленности. Делается предположение, что если значимость всех связей «это» (вес всех соответствующих дуг) исходящих из узла, описывающего термин-омоним, будет одинаковой, то отношение данного термина к тому или иному контексту будет равнозначным для всех возможных контекстов. В реальности термины никогда не имеют нейтрального характера. Значимость связей устанавливается исходя из предубеждений и установок каждого из экспертов. Таким образом, в одном контексте данный термин будет восприниматься однозначно, а в другом может привести к непониманию информации или даже к изменению контекста. В приведенном примере (см. рис.1) отображены две группы терминов одной предметной области, и представлены два термина (4 и 6) относящихся к обеим группам. Отношения термина 4 в данном примере демонстрируют результат сделанного предположения. Термин 6 со своими связями показывает один из возможных вариантов представления омонимии в реальности.

Синонимы, эмоционально близкие термины и соответствующие связи между ними наглядно демонстрируют природу понятия контекста. Это означает, что при наличии установленных связей между синонимами, соответствующий термин в тексте может быть равнозначно заменен другими. В общем случае любая связь именованная «это» может выступать в роли синонимии, однако для наглядности принято называть синонимами те термины, поток между которыми превышает установленный уровень. Уровень выбирается индивидуально для каждой формируемой модели. Понятие эмоционально близких терминов

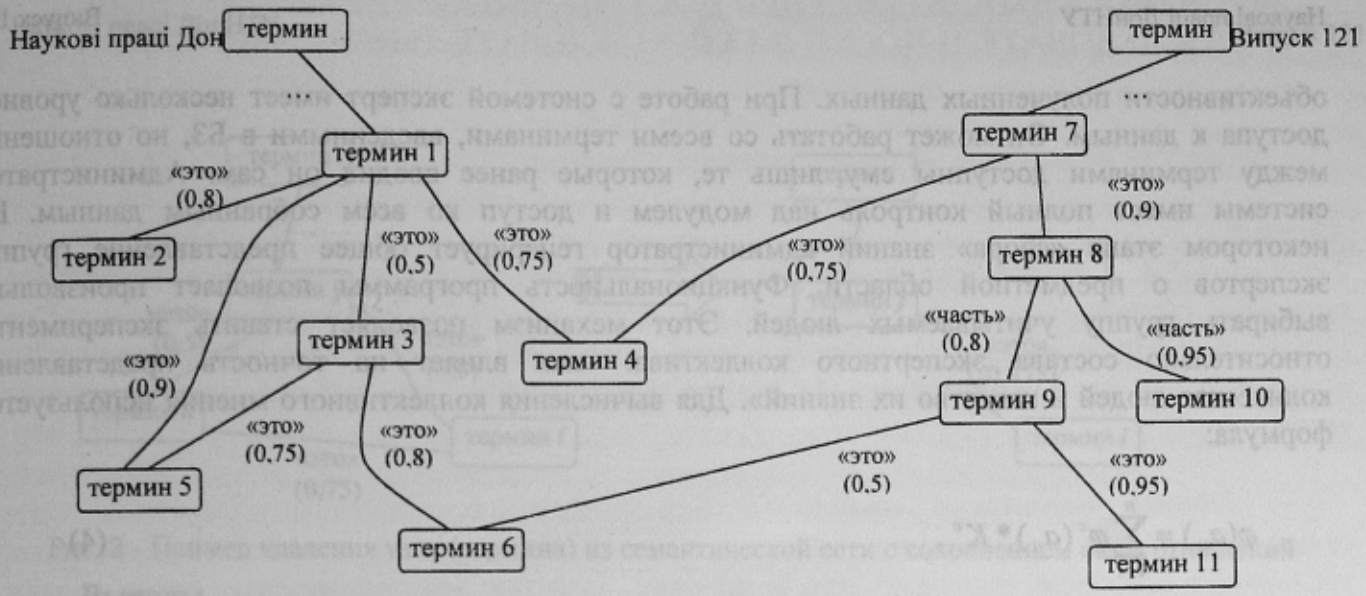

Рис. 1 - Фрагмент семантической сети – пример отношения между объектами в мировой модели

раскрывает суть одного из постулатов модели сравнительных семантических признаков [9], которая выбрана за базовую при организации семантических сетей гибридной модели [1]. Согласно проведенным психологами опытам на то, как человек сопоставляет один термин с другим, влияет частота обращения к этим терминам [10]. Так, например «снегирь» ассоциируется у нас больше с терминов «птица» нежели с термином «отряд воробьиных». На рисунке 1 описанное отношение представлено связями термина 5. Связь между термином 5 и термином 1 более весомая и чаще используется, нежели путь через термин 3. Подобные отношения встречаются в жизни повсеместно и являются практической базой человека, полученной из возникавших ранее ситуаций.

Следует отметить, что связи «это» в свою очередь могут иметь и составляющие термины (термины-части). Таким образом, наличие одного из таких терминов гарантирует существование составного термина. Изображенный на рисунке 1 термин 11 хоть и не имеет связи «часть» все же является составляющим для термина 8.

Выделив терминологический словарь мировой модели, и установив соответствующие связи, мы все еще не сможем полноценно воспринимать и обрабатывать поступающую информацию. Неотъемлемой частью знаний эксперта является событийная модель мира [11]. Схемы возможных взаимодействий объектов устанавливают соответствующий регламент на отношения объектов предметной области и обеспечивают логическую целостность модели. Отсутствие такой информации не позволит определить корректность обрабатываемой информации и может привести к коллизиям. Выделение событий зависит от квалификации эксперта и его знаний о предметной области. Чем опытнее эксперт, тем больше будет возможных событий, а среди них будет больше макрособытий (абстрактных или обобщенных).

Уникальность знаний каждого эксперта представлена в модели прототипами.

Программный модуль формирования БЗ.

В ходе реализации системы интеллектуального поиска на основе нечеткой гибридной модели был разработан модуль для формирования БЗ модели. При его разработке были учтены основополагающие принципы и возможные проблемы, присущие методам извлечения знаний путем опроса экспертов.

Модуль реализован в виде интерактивного средства общения с экспертом. Раздельный доступ для каждого пользователя на основе аутентификации решает вопросы взаимного психологического воздействия специалистов (конформизм, уступчивость) и способствует

объективности полученных данных. При работе с системой эксперт имеет несколько уровней доступа к данным. Он может работать со всеми терминами, введенными в БЗ, но отношения между терминами доступны ему лишь те, которые ранее вводил он сам. Администратор системы имеет полный контроль над модулем и доступ ко всем собранным данным. На некотором этапе «сбора» знаний администратор генерирует общее представление группы экспертов о предметной области. Функциональность программы позволяет произвольно выбирать группу учитываемых людей. Этот механизм позволяет ставить эксперименты относительно состава экспертного коллектива: «как влияет на точность представления количество людей и качество их знаний». Для вычисления коллективного мнения используется формула:

$$\varphi(a_{ij}) = \sum_{c=1}^{n} \varphi^c(a_{ij}) * K^c, \qquad (4)$$

где $\varphi(a_{ij})$ - коллективный (суммарный) поток по дуге из узла i в узел j (3), $\varphi^c(a_{ij})$ - поток по дуге, полученный на основе знаний c-го эксперта, K^c - уровень квалификации c-го эксперта, n – размер группы экспертов (количество человек).

Данные эксперимента подтвердили выдвигаемые предположения, что для формирования БЗ оптимальным считается учет знаний небольшой группы экспертов (до 10 человек), при этом максимально однородной по качеству знаний. Таким образом, формируется наиболее целостное представление о мировой модели, в котором не содержаться значимые «всплески» и разногласия.

Процесс опроса разбит на этапы интерфейсной частью. Навигация на сайте дает возможность осуществлять переход между этапами в произвольном порядке. При входе в систему пользователь получает информацию об объеме терминологического словаря и об установленных им (пользователем) связях между существующими терминами. Первый раздел модуля предоставляет интерфейс для работы с терминами: добавление новых, редактирование и удаление существующих. Соответственно на действия с общими данными наложен некоторый регламент. Каждый термин помимо имени имеет набор признаков, характеризующих его. Все признаки хранятся в едином хранилище, а с термином связаны посредством ссылок на них. Это дает возможность позволить эксперту не только ввести новый признак объекта, но и выбрать для него признак из уже существующих.

Поскольку доступ к множеству терминов имеет все пользователи системы, то необходимо учесть возможные коллизии с удалением терминов. Например, при удалении термина, у которого другие пользователи уже внесли связи, необходимо их сохранить. Для этого предложен механизм «копирования отношений», представленный на рисунке 2. Расчет производится по формуле 5. При этом пересчет происходит и в общей БЗ, и в БЗ каждого пользователя отдельно.

$$\varphi(a_{ij}) = \varphi(a_{ij}) \oplus (\varphi(a_{ik}) * \varphi(a_{kj})), \qquad (5)$$

где $\varphi(a_{ij})$ - поток по дуге из узла i в узел j, $\varphi(a_{ik})$ - поток по дуге из узла i в узел k, $\varphi(a_{kj})$ - поток по дуге из узла k в узел j, k – удаляемый узел (термин).

Этап установки признаков, связей и возможных схем взаимодействия терминов реализован в виде опросника. Пользователю предоставляется термин и предлагается выбрать связи между ними. Начать или прекратить опрос можно в любой момент.

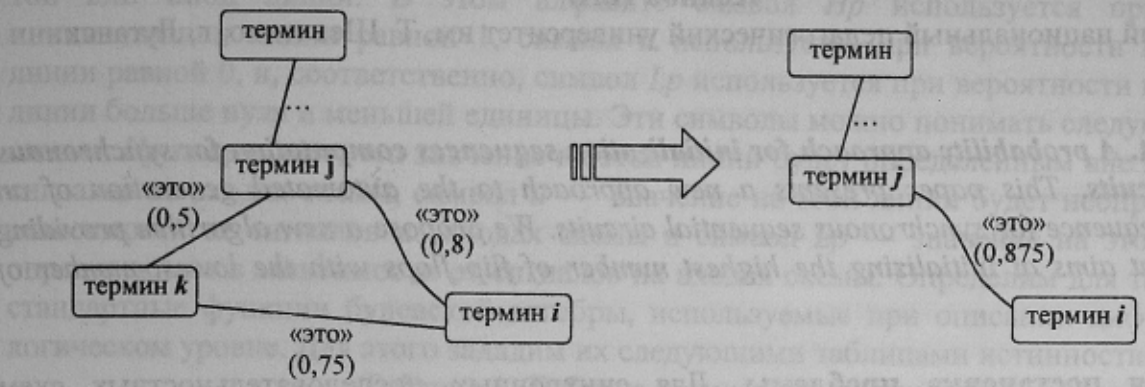

Рис. 2 - Пример удаления узла (термина) из семантической сети с сохранением силы отношений

Выводы.

Задача приобретения знаний до сих пор представляет собой одну из основных проблем в области построения интеллектуальных систем. В работе раскрыта суть данной проблемы и основные аспекты ее решения на примере формирования БЗ гибридной модели. Реализован и описан программный модуль извлечения и представления знаний экспертов.

Литература

1. Парамонов А.И., Каргин А.А. «Гибридная нечеткая модель обработки концептуальной информации» - Труды международной конференции " ИАИ-2005".
2. Buchanan B. G., Barstow D., Bechtel R., Bennet J., Clancey W., Kulikowski C., Mitchell T. M. and Waterman D. A. «Constructing an expert system.» In Building Expert Systems (Hayes-Roth F, Waterman D.A. and Levat D., eds.), MA: Addison-Wesley, 1983.
3. Feigenbaum E. A. «The art of artificial intelligence: themes and case studies of knowledge engineering.» In Proc. 5th International Joint Conference on Artificial Intelligence, 1977.
4. Шелов С.Д., «Определение терминов и понятийная структура терминологии.» – СПб.: Изд-во СпбГУ – 1998.
5. Пол Клайн «Справочное руководство по конструированию тестов» - Личностные опросники. Формулировка вопросов, Киев, 1994.
6. Квинн Вирджиния «Прикладная психология. 4-е международное издание» - СПб: издательство «Питер», 2000. – 560 с.: ил. – (Серия «Учебник нового века»).
7. Сидорова Е.А. «Технология разработки тематических словарей на основе сочетания лингвистических и статистических методов» - Труды международной конференции "Диалог'2005" - М.: Наука, 2005.
8. Добров Б.В., Лукашевич Н.В., Сыромятников С.В. «Формирование базы терминологических словосочетаний по текстам предметной области» - Труды 5-ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL2003.
9. Андерсон Дж. «Когнитивная психология. 5-е изд.» – СПб.: Питер, 2002. – 496с.: ил. – Серия «Мастера психологии»
10. Collins A. M. and Quillian M. R. «Retrieval time from semantic memory.» Journal of Verbal Learning and Verbal Behavior, №8, 1969
11. Солсо Р. «Когнитивная психология» – СПб.: Питер, 2002. – 592с.: ил. – (Серия «Мастера психологии»).