УДК 81'33(07):378

Л.А. ФЕДОРЧЕНКО[1], Н.Ф. ХАЙРОВА[2], А.И. ДОВНАРЬ[1], С.О. БУЛГАКОВ[3]

[1] ***Международный Славянский университет, Харьков, Украина***
[2] ***Национальный технический университет «ХПИ», Харьков, Украина***
[3] ***ЧП "Мейн-Мейкер"Харьков, Украина***

МЕТОД АВТОМАТИЗИРОВАННОГО ПОСТРОЕНИЯ СЕМАНТИЧЕСКОЙ СЕТИ ТЕРМИНОВ УЧЕБНОЙ ДИСЦИПЛИНЫ

В работе предложен метод автоматизированного построения семантической сети терминов предметной области. В качестве исходной информации для построения сети может быть использован терминологический словарь предметной области. В основу метода положено формализованное описание семантического поля и процедура вычисления силы семантических связей между терминами предметной области. Процедура вычисления силы семантических связей использует результаты компонентного анализа терминов и их дефиниций. Использование метода позволяет создавать матричное представление семантической сети. Метод может быть полезен при разработке онтологий для корпоративных баз знаний.

Ключевые слова: *учебная дисциплина, семантическая сеть, терминологический словарь, онтология.*

Введение

В системе высшего образования накопилось множество противоречий, которые образуют большую комплексную проблему ее реформирования. Одним из них авторы [1] отмечают противоречия между методологическими основами педагогики и языкознания (лингвистики). Методы языкознания на современном этапе развития образования Украины вступают в противоречие с методологической базой педагогики как средства коммуникации между студентами и преподавателями, студентами и учебно-методической литературой, научно-педагогическими работниками и научной, и учебно-методической литературой. В настоящее время особенно остро ощущаются противоречия между терминологическими системами тех или иных предметных областей, которые изучаются в разных ВУЗах. Одним из путей разрешения указанного противоречия авторы предлагают использование технологического подхода в организации и проведении учебного процесса. Реализация технологического подхода требует разработки моделей и методов построения инструментальных средств, позволяющих автоматизировать процесс интеллектуальной обработки научной и учебной текстовой информации.

В работе [2] проведен анализ особенностей текстов учебно-методических материалов, используемых в учебном процессе ВУЗа, обосновывается целесообразность использования онтологического моделирования для построения инструментальных средств поддержки учебного процесса. Использование онтологического подхода предусматривает анализ предметной области, выделение ее понятий, определение отношений между ними, и формулировку правил логического вывода с учетом этих отношений [3, 4]. В приведенных работах отмечается, что указанные процедуры связаны с существенными затратами интеллектуальных и временных ресурсов. В этой связи, решение задач направленных на автоматизацию указанных процедур представляется **актуальным**.

Целью настоящей работы является разработка метода автоматизированного построения семантической сети терминов как основы для создания онтологии учебной дисциплины с использованием результатов анализа текстовых источников информации, представляющих ее содержательную часть.

Результаты исследований

Идея использования самантических сетей в задачах автоматической обработки текстов изложена в работе [5]. Автор отмечает достоинства сетевого представления лексики, к которым относит следующие:

- эксплицитное представления семантических связей между словами;
- возможность определения количественных параметров, характеризующих систему семантических связей и семантическую структуру лексики;
- возможность определения семантической связи между любыми двумя единицами лексико-семантической системы;
- возможность определения силы семантической связи между единицами лексико-семантической системы.

Необходимость в разработке метода обусловлена тем, что в указанной работе построение семанти-

ческих сетей предусмотрено вручную, по результатам анализа текстов толковых или идеографических словарей. Однако это не тривиальная задача и ее решение связано с довольно трудоемкой процедурой выявления семантических связей между терминами, а также отслеживания цепочек связей между ними.

Как отмечается в указанной работе, представление семантических связей между словоформами, описывающими определенную область знаний и, в частности, учебной дисциплины, предполагает задание в явном виде их направления, содержания и силы.

Подход к автоматизированному определению направления и содержания семантических связей между терминами предложен в работе [6], который основан на использовании результатов анализа синтаксических схем и лексических единиц дефиниций в текстах. Задачей метода, предлагаемого в настоящей работе, является определение силы семантических связей между терминами учебной дисциплины. При этом будем понимать, что сила связи характеризует удаленность терминов в некотором гипотетическом семантическом пространстве, которое будем называть семантическим полем учебной дисциплины. В отличие от понятия поля в математике или физике, семантическое поле в лингвистике используется для обозначения совокупности языковых единиц, объединенных каким-то общим (интегральным) семантическим признаком. В настоящей работе интегральным признаком является лексика учебной дисциплины.

Для формализованного описания семантического поля учебной дисциплины обозначим через $U=\{x_k\}$ - множество словоформ, которые употреблены в текстах при описании содержательной части учебной дисциплины, где $k=\overline{1,K}$, K - кардинальное число множества. Тогда под семантическим полем будем понимать кортеж

$$P=(T,S,R),$$

где $T=\{t_j\}$ - множество терминов учебной дисциплины, $j=\overline{1,J}$; $S=\{s_i\}$ множество смысловых содержаний (значений) этих терминов, $i=\overline{1,I}$; J и I - кардинальные числа множеств T и S соответственно ($s_i=\{x_{i_k}\}$, $x_{i_k}\subset U$); $R=\{r_{j_1,j_2}\}$ - множество семантических отношений между элементами множества T ($j_1,j_2\subset J; j_1\neq j_2$), которые выражают силу связи между терминами $t_{j_1}\{x\}$ и $t_{j_2}\{x\}$.

Элементы $t_j\{x\}$ и $s_i\{x\}$ - цепочки словоформ $x_1,x_2,...,x_k$, которые формируются из элементов множества $U=\{x_k\}$, с использованием которых выражаются термины и их смысловые содержания, соответственно.

Следует отметить, что в идеальном случае, одному термину (значению элемента $t_j\{x\}$) всегда должно соответствовать одно смысловое содержание (один элемент $s_i\{x\}$). В реальных ситуациях одному термину может соответствовать несколько смысловых содержаний. Однако в педагогической практике, при изложении содержания учебной дисциплины, авторы, в процессе формулировки смысловых содержаний для терминов, используемых в учебной дисциплине, обычно стремятся давать однозначное определение, при этом устанавливают бинарные отношения между элементами множеств T и S, в результате чего эти пары образуют множество дефиниций $D=\{d_f\}$. Указанные бинарные отношения могут быть представлены конкатинацией: $d_f=t_j+s_i$. Конкатенация - это бинарная операция, которая задана на словах определенного алфавита. В нашем случае таким алфавитом является множество $U=\{x_k\}$, а словами являются цепочки словоформ образованные из этого алфавита. Совокупности словоформ, из которых формируются элементы t_j и s_i. будем называть **компонентами**.

Следует отдельно остановиться на допустимости приведенного формализованного представления.

Многие исследователи в языкознании обращались к проблеме значения слов и методов выявления семантических связей между ними. В отличие от синтаксической связи между словоформами во фразе, семантическая связность принадлежит к тем объектам, которые непосредственно не наблюдаемы. Поэтому для выявления семантических связей необходимо использовать либо неформальные методы, либо формальные, которые основаны на корреляционной зависимости между семантической связью и некоторыми наблюдаемыми признаками языковых явлений.

Неформальные методы по своей сути интуитивные и используют «языковое чутье» носителей языка. Результатом применения неформальных методов является множество существующих толковых, идеографических и терминологических словарей, которые достаточно адекватно описывают систему языка и могут использоваться как инструмент лингвистических исследований в формализованных процедурах обработки естественно-языковых текстов.

В основу формальных методов положена гипотеза о существовании корреляционной зависимости между семантической связью слов и некоторыми наблюдаемыми признаками, которые являются косвенными показателями этой связи. В настоящее время получил широкое распространение ряд формальных методов выявления семантической связи между словами, основанных на анализе их дистрибуций (определений или дефиниций) [5, 7, 8].

По мнению авторов, семантические связи между словоформами обусловлены сходством их лексических значений. На этой основе образуются различные группировки, внутри которых слова характеризуются определенной степенью сходства, подобия или близости.

Наибольшей близостью, например, характеризуются слова входящие в синонимические ряды. В соответствии с наиболее общим определением явления синонимии – это группировка таких слов, которые максимально близки между собой и могут рассматриваться как эквивалентные, тождественные друг другу [9]. Однако в естественном языке такой тождественности не существует, т. к. значения слов не имеют четких границ, они размыты. Поэтому, установить абсолютное тождество значений не представляется возможным, что приводит к необходимости оперировать понятием близости значений, которое несет в себе представление о некоторой относительности.

При использовании формальных методов исходят из того, что семантическая связь между словоформами является функцией связей между предметами окружающего мира. Однако система семантических связей в лексике, отражая систему связей в предметной области учебной дисциплины, не определяется полностью системой лексики. Система семантических связей в лексике не изоморфна системе связей между объектами и явлениями в предметной области. На картину семантических связей между словоформами накладывают отпечаток чисто языковые факторы (синонимия, омонимия, и другие явления многозначности слов). Указанная размытость границ значений слов создает существенные затруднения при моделировании лексико-семантических систем.

Анализ методов и подходов к решению задачи выявления семантических связей между словоформами показал, что наиболее продуктивной является комбинация неформальных и формальных методов, как это показано в работе [5]. При использовании подхода из выше указанной работы, будем исходить из того, что лексическое значение терминов определяется их дефинициями, состоящими из компонентов x_i, которые формируют значение термина. Известно, что между терминами существуют различные виды отношений, такие как родовидовые, общее-частное, включение и другие. Они и задают конфигурацию семантической сети терминов. Однако в словарях такие отношения заданы в неявном виде. Для представления отношений $R = \left\{ r_{j1,j2} \right\}$ семантического поля учебной дисциплины в явном виде, необходимо определить семантически связанные элементы множества $T = \left\{ t_j \right\}$. Семантически связанными будем считать термины t_{j1} и t_{j2} имеющие общие семантические компоненты в элементах s_{j1} и s_{j2}, и которые определяют смысловые содержания указанных терминов. Чем больше совпадающих семантических компонент в элементах s_{j1} и s_{j2}, тем больше сила семантической связи между терминами t_{j1} и t_{j2}.

Для вычисления силы семантической связи между терминами, т.е. численного значения элементов множества $R = \left\{ r_{j1,j2} \right\}$ необходимо произвести попарное сравнение компонент, определяющих каждый термин и определить количество совпадающих компонент.

Величина силы связи между терминами может быть вычислена по формуле:

$$r_{j1,j2} = \frac{\mathrm{card}(s_{j1} \cap s_{j2})}{\mathrm{card}(s_{j1} \cup s_{j2})}, \qquad (1)$$

где $r_{j1,j2}$ - величина силы связи между терминами t_{j1} и t_{j2}; $\mathrm{card}(s_{j1} \cap s_{j2})$ - количество совпадающих компонент, определяющих значения терминов t_{j1} и t_{j2}; $\mathrm{card}(s_{j1} \cup s_{j2})$ - общее количество компонент, определяющих значения терминов t_{j1} и t_{j2}.

Корректность предложенного метода можно продемонстрировать на фрагменте терминологического словаря [9]. На рис. 1 представлены вербальные значения фрагмента семантического поля учебной дисциплины.

На рис. 2 представлена семантическая сеть фрагмента семантического поля из указанного терминологического словаря. Для практического использования приведенной формулы количественного определения силы семантической связи между терминами, тексты дефиниций должны быть предварительно обработаны. Обработка заключается в фильтрации и нормализации. На этапе фильтрации исключаются словоформы, не формирующие значение термина. К ним относятся предлоги, союзы, местоимения и т.п.

Нормализация заключается в приведении имен существительных и прилагательных в именительный падеж единственного числа, глаголов в форму инфинитива.

Проведение нормализации необходимо по следующей причине. Например, если в разных дефинициях будут употреблены словоформы с падежными изменения: «система», «системе» или «системой», которые с точки зрения семантики являются одним и тем же термином, то при вычислении величины силы связи между терминами они будут учитываться как разные компоненты.

Для реализации метода была создана лексическая база данных (ЛБД), структура которой представлена на рис. 3.

ЛБД состоит из трех таблиц. Первая таблица – Words, в ней содержатся все термины, описывающие предметную область.

Вторая таблица – Comps, в ней содержатся все множество компонент, которые участвуют в определении терминов из таблицы Words. Третья таблица Genered – таблица связей. В ней ставятся в соответствие каждому термину из таблицы Words только те компоненты из таблицы Comps, которые его определяют. Наличие встроенных в систему управления базой данных функций формирования запросов позволяет задавать различные режимы обработки данных в ЛБД. В частности, в предложенной ЛБД с помощью комплекса запросов производится попарное сравнение компонентов двух терминов и вычисление величины их семантической связи по формуле (1).

Для вычисления силы семантической связи между терминами была разработана программа. В качестве исходных данных она использовала лексическую базу данных, построенную на основе терминологического словаря [9]. Результатом работы программы является табличное представление семантической сети (рис. 4). Как видно из рисунка таблица состоит из шести столбцов. В первом и третьем столбцах указаны семантически связанные термины. Во втором и четвертом столбцах соответственно, указаны общие количества компонент для каждого термина, которые формируют значения сравниваемых терминов. В пятом столбце приведено количество совпадающих компонент у сравниваемых терминов. В шестом столбце – значение величины силы семантической связи, которое вычислено по формуле (1).

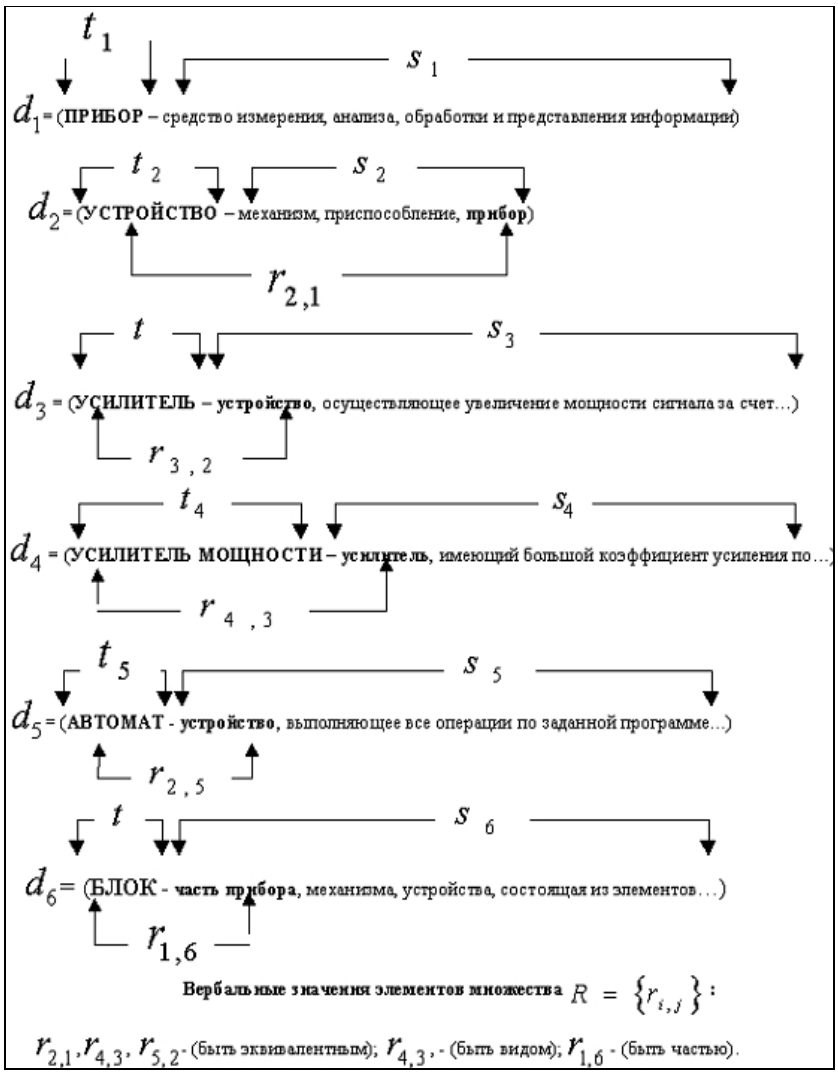

Рис. 1. Вербальные значения элементов фрагмента семантического поля учебной дисциплины

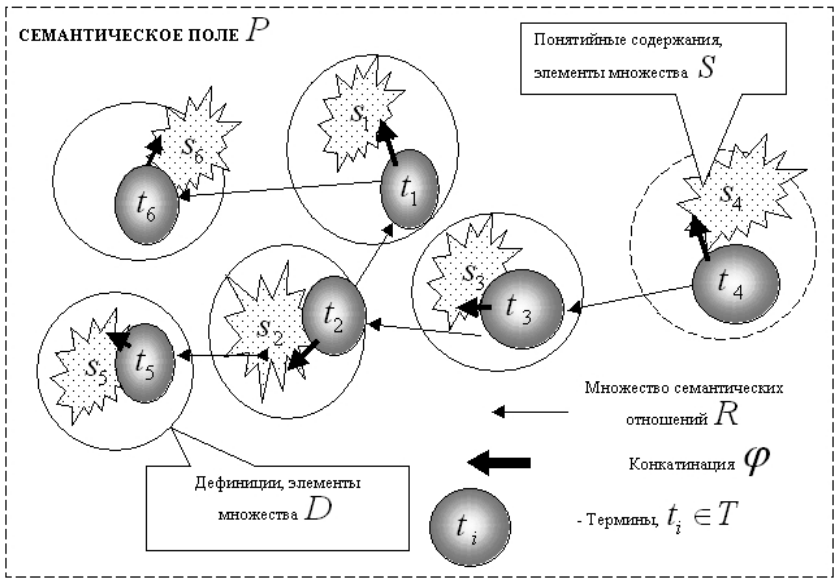

Рис. 2. Фрагмент семантической сети терминологического словаря

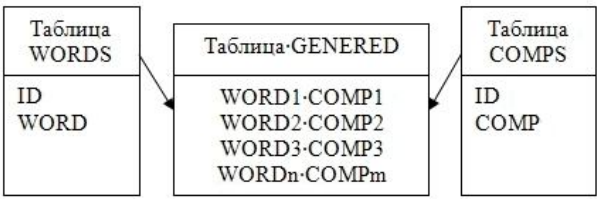

Рис. 3. Структурная схема ЛБД

Первый термин	M	Второй термин	N	C	V
АВТОМАТ	8	БЛОК	10	2	0,22222222
АВТОМАТ	8	УСИЛИТЕЛЬ	10	1	0,11111111
БЛОК	10	АВТОМАТ	8	2	0,22222222
БЛОК	10	УСИЛИТЕЛЬ	10	1	0,1
БЛОК	10	УСТРОЙСТВО	3	2	0,30769231
БЛОК	10	ЭЛЕМЕНТ	10	2	0,2
УСИЛИТЕЛЬ	10	АВТОМАТ	8	1	0,11111111
УСИЛИТЕЛЬ	10	БЛОК	10	1	0,1
УСИЛИТЕЛЬ	9	УСИЛИТЕЛЬ МОЩНОСТИ	13	9	0,81
УСИЛИТЕЛЬ МОЩНОСТИ	13	УСИЛИТЕЛЬ	9	9	0,81
УСТРОЙСТВО	3	БЛОК	10	2	0,30769231
ЭЛЕМЕНТ	10	БЛОК	10	2	0,2

Рис. 4. Табличное представление результатов вычисления силы семантической связи между терминами

Выводы

Таким образом, подводя итог изложенному в работе, сделаем следующие выводы.

Предложенный метод позволяет автоматизировать структуризацию семантических и терминологических связей при формировании понятий и терминов на основе естественно-языковых закономерностей, что позволяет представлять структуру понятийного аппарата учебной дисциплины в явном виде. Представление понятийного аппарата учебных дисциплин в виде семантических сетей придает учебному материалу новое качество, что открывает возможности для задач, решение которых не представлялось возможным при традиционном представлении учебного материала (рис. 5).

Метод автоматизированного построения семантической сети терминов может использоваться как отдельный этап в проектировании онтологий различного назначения, в частности при создании корпоративных баз знаний.

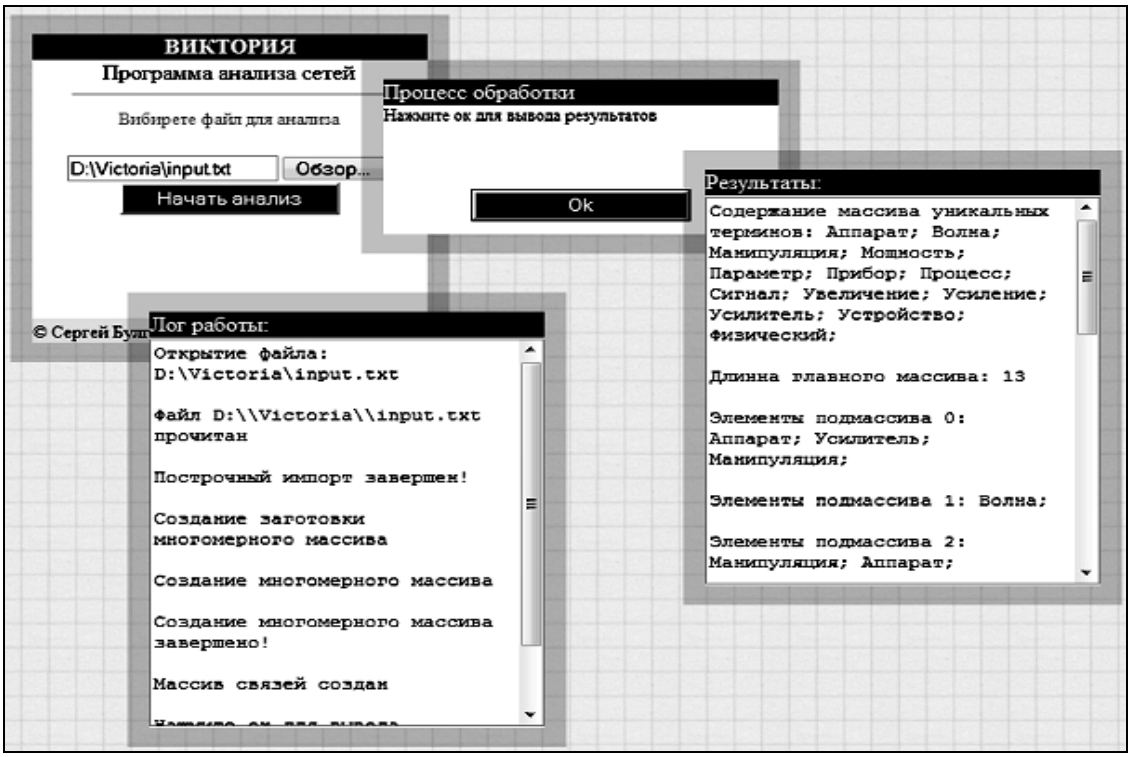

Рис. 5. Программная реализация вычисления силы семантической связи и анализа сети

Литература

1. Метешкін, К.О. Трансферт освітніх технологій як інноваційна складова розвітку України [Текст] / К.О. Метешкін, Х.В. Раковський, Н.Х. Раковська // Вища освіта України. – 2009. – Т. VI (16). Додаток 4. Тематичний випуск «Вища освіта України у контексті інтеграції до Європейського освітнього простору. – 576 с.

2. Федорченко, Л.А. Особенности построение лингвистической онтологии учебно-методического материала [Текст] / Л.А. Федорченко, К.А. Метешкин // Вестник Международного Славянского университета. Серия «Технические науки». –2008. – Т. XI, № 1. – С. 34 – 43.

3. Гаврилова, Т.А. Базы знаний интеллектуальных сител [Текст] / Т.А. Гаврилова, В.Ф. Хорошевский. – СПб.: Питер, 2001. – 384 с.

4. Гладун, А.Я. Онтологии в корпоративных системах. Часть 2. [Электронный ресурс] / А.Я. Гладун, Ю.В. Рогушина // Корпоративные системы. – 2006. – № 1. – Режим доступа: http://www.management.com.ua/ims/ims116.html. – 23.03.2007 г.

5. Скороходько, Э.Ф. Семантические сети и автоматическая обработка текста. [Текст] / Э.Ф. Скороходько. – Киев: Наук. Думка, 1983. – 212 с.

6. Федорченко, Л.А. Формализованное представление фрагментов текста учебно-методического материала [Текст] / Л.А. Федорченко // Вестник Международного Славянского университета. Серия «Технические науки». – 2007. – Т. X, № 1. – С. 44 – 52.

7. Апресян, Ю.Д. Лексическая семантика. Синонимические средства языка [Текст]: учеб. пос ./ Ю.Д. Апресян. - М.: Наука, 1974. – 366 с.

8. Новиков, А.И. Семантические расстояния в языке и тексте [Текст] / А.И. Новиков, Е.И. Ярославцева. – М.: Наука, 1990. – 136 с.

9. Словарь терминов по системам управления летательных аппаратов (СУЛА). [Текст] / А.С. Кулик, А.Г. Гордон, В.Н. Кортунов, В.Ф. Симонов, Ю.Н. Соколов. – Х.: Нац. аэрокосмический ун-т. "ХАИ", 2001. – 224 с.

Поступила в редакцию 2.12.2011

Рецезент: д-р техн. наук, проф., зав. кафедрой интеллектуальных компьютерных систем Н.В. Шаронова, Национальный политехнический университет «Харьковский политехнический институт».

МЕТОД АВТОМАТИЗОВАНОЇ ПОБУДОВИ СЕМАНТИЧНОЇ МЕРЕЖІ ТЕРМІНІВ НАВЧАЛЬНОЇ ДИСЦИПЛІНИ

Л.А. Федорченко, Н.Ф. Хайрова, О.І. Довнар, С.О. Булгаков

В роботі запропоновано метод автоматичної побудови семантичної мережі термінів предметної галузі. Як вихідна інформація для побудови семантичної мережі може використовуватись термінологічний словник предметної галузі В основу метода покладено формалізований опис семантичного поля та процедура обчислення сили семантичних зв'язків між термінами предметної галузі. Процедура обчислення сили семантичних зв'язків використовує результати компонентного аналізу термінів та їх дефініцій. Використання метода дозволить створювати матричне представлення семантичної мережі. Метод може бути використано при побудові онтологій для корпоративних баз знань.

Ключові слова: навчальна дисципліна, семантична мережа, термінологічний словник, онтологія.

A METHOD FOR AUTOMATED CONSTRUCTION OF A SEMANTIC NETWORK TERMS IN THE ACADEMIC DISCIPLINE

L.A. Fedirchenko, N.F. Khairova, A.I. Dovnar, S.O. Bulgakov

In this paper we propose a method for automated construction of a semantic network terms in the domain. As initial information for constructing the network can be used terminology dictionary domain. The method is based laid formal description of the semantic field and the procedure for calculating the strength of semantic relations between domain terms. The procedure for calculating the strength of semantic relations uses the results of component analysis of terms and their definitions. Using the method allows you to create a matrix representation of a semantic network. The method may be useful in the development of ontologies for corporate knowledge bases.

Key words: academic discipline, the semantic network terminology dictionary, ontology.

Федорченко Леонид Аксентьевич – ст. преподаватель кафедры информационных технологий и высшей математики Международного Славянского университета Харьков, e-mail: leo22091946@yandex.ru.

Хайрова Нина Феликсовна – канд. техн. наук, доцент, доцент кафедры интеллектуальных компьютерных систем Национального технического университета «Харьковский политехнический институт», e-mail: nina_kajrova@yahoo.com.

Довнарь Александр Иосифович – канд. техн. наук, доцент, заведующий кафедрой информационных технологий и высшей математики Международного Славянского университета, Харьков, e-mail: dov-alexandr@yandex.ua.

Булгаков Сергей Олегович – специалист по информационным технологиям, ЧП "Мейн-Мейкер", e-mail: bulgakoff08@gmail.com.