УДК 519.7+681.3

МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ ТЕРМИНОЛОГИЧЕСКИХ СИСТЕМ ДОКУМЕНТОВ В САПР

© 2010 И.В. Арзамасцева

Ульяновский государственный технический университет

Поступила в редакцию 14.05.2010

В данной статье рассмотрены математические модели терминологических систем и их использование в САПР.

Ключевые слова: математическое моделирование, терминологические системы документов, САПР

С развитием информационных технологий и возрастанием их роли в обработке большого объема информации особое значение приобретает математическое моделирование языковых структур, которое расширяет возможности автоматизации обработки информации на естественном языке. Многие задачи автоматизированного проектирования могут быть решены только на основе взаимодействия математических методов с классическими методами языкового анализа.

Статистическое моделирование предполагает формально-структурное членение лингвистического объекта и выделение в нем формальных элементов, которые становятся предметом дальнейшего изучения и использования в системах АПР.

В настоящий момент в системах автоматизированного проектирования не существует методов и методик для создания словарей-тезаурусов. Зато такие методы накоплены в лингвистике. Использование лингвистических методов в системах автоматического проектирования расширяют возможности САПР.

Для точного индексирования текстов на *ЕЯ* необходимо создать модель математической зависимости лексики от статистики. Были построены две математические модели терминологических систем для определения предметной области (ПО) документов для использования их в САПР.

Рассмотрим модель, построенную на основе редукции конечного количества правил, где входные переменные определяются словарями предметной области (в нашем случае Нечеткой логики). В процессе исследования были выделены 6 подсловарей данной предметной области – “Нечеткая логика”, “Логика”, “Математика”, “Компьютер”, “Искусственный интеллект” и “Управляющие системы”. Таким образом в модели нечеткого вывода выделяем 6 входных переменных, которые могут принимать значения, соответствующие относительной частоте встречаемости терминов данного словаря в множестве терминов предметной области. Обозначим их *F, L, M, C, KI* и *LT*:

F – относительная частота встречаемости терминов подсловаря “Нечеткая логика” (X_1);

L – относительная частота встречаемости терминов подсловаря “Логика” (X_2);

M – относительная частота встречаемости терминов подсловаря “Математика” (X_3);

LT – относительная частота встречаемости терминов подсловаря “Управляющие системы” (X_4);

C – относительная частота встречаемости терминов подсловаря “Компьютер” (X_5);

KI – относительная частота встречаемости терминов подсловаря “Искусственный интеллект” (X_6).

Для описания переменных введены три терма {“min”, “med” и “max”}, описывающие значения этой переменной.

Очевидно, что степень принадлежности “0” к “min” = 1, а степень принадлежности “0” к “max” – соответственно 0. В качестве значений степени принадлежности возьмем нормированную относительную частоту, описываемую стандартными треугольными функциями принадлежности. На рис. 1 представлены функции принадлежности первой входной переменной модели:

По данным о 30 группах текстов, полученных статистическим путем, были сформулированы нечеткие правила отнесения текста к определенной предметной области (в нашем случае – к области НЛ) [1].

Выходная переменная отражает принадлежность текста к предметной области НЛ. Для описания переменной использованы два терма {“F”,

Арзамасцева Иветта Вячеславовна, старший преподаватель кафедры “Прикладная лингвистика”, аспирант кафедры “Информационные системы”.
E-mail: lingua@ulstu.ru.

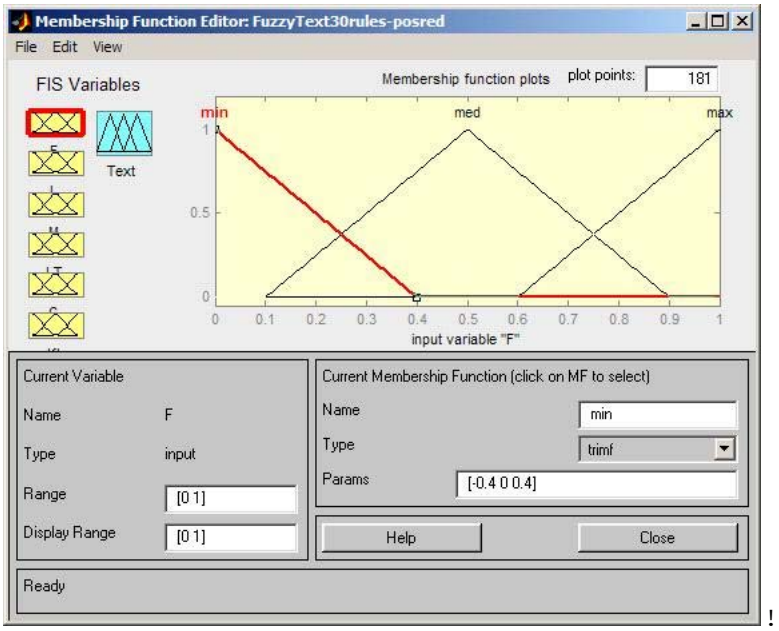

Рис. 1. Функции принадлежности входной переменной X_1

и “nF”}, отражающие принадлежность текста к данной предметной области:

F – степень принадлежности текста к предметной области НЛ.

nF – степень непринадлежности текста к предметной области НЛ.

Для каждого терма использована линейная функция принадлежности (рис. 2)

Таким образом, рассмотрена моделирующая зависимость вида $y = f(x_1, x_2, x_3, x_4, x_5, x_6)$ с использование одной базы знаний.

Введем оператор *Fuzzy*, который будет выполнять набор операций: импликация и агрегация. Результатом выполнения этих операций над фазифицированным вектором входных переменных *X* оператора *F* будет множество:

$$\widetilde{y} = \sum_{i=1}^{n} \frac{\mu_{d_1}(X)}{d_1},$$

где *м* – функция принадлежности,

X – входной вектор,

d – степень принадлежности текста к предметной области.

Для базы продукций получаем:

оператор $fuzzy\ (x_i) = fuzzy = x_i = fuzzy(x_i) = [\mu_{x_i^j}, x_1]$.

Получаем следующую математическую модель определения принадлежности текста пред-

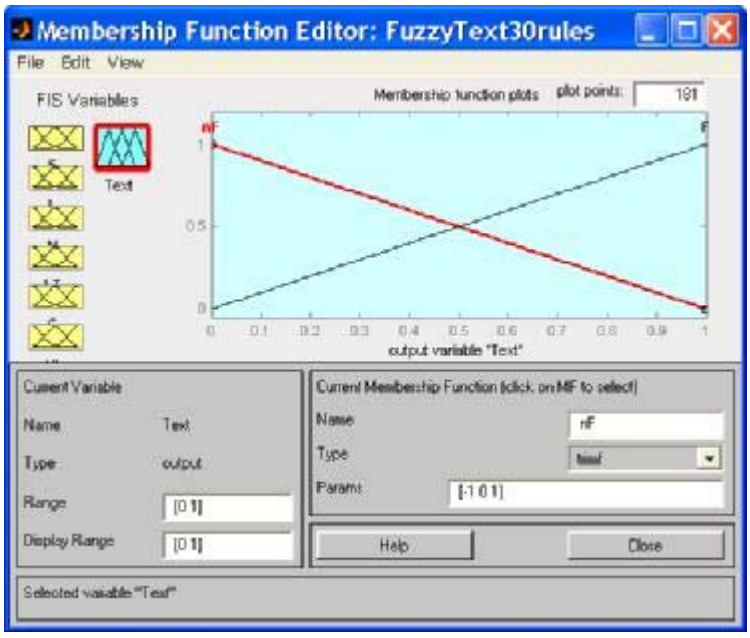

Рис. 2. Функции принадлежности выходной переменной модели

метной области на основе нечеткого вывода по Мамдани:

$$\tilde{y} = F(Fuzzy(\{x_i\}_{i=1,6}))$$

или

$$\tilde{y} = \tilde{A}(\{x_i\}_{i=1,6}) \circ R(\{x_i\}_{i=1,6}, y);$$
$$\tilde{y} = A_{j=1,6}\{x_i\} \circ R_{k=1,30}(\{x_{i=1,6}\}, y),$$

где $\tilde{y}$ – нечеткая выходная переменная;

x – входная переменная;

i – индекс для входа;

j – индекс для выхода;

Γ – функции принадлежности входных переменных $\{A_1, A_2, A_3, A_4, A_5, A_6\}$;

$\circ$ – знак композиции.

$\circ \equiv \vee \wedge \ (A_j(x_i))$

j=1; i=1,6

где $\vee$ – любая s-норма;

$\wedge$ – любая t-норма,

в нашем случае

$\vee$ – max – sup;

$\wedge$ – min – inf.

$$\circ = \max_{j=1} \min_{i=1,6} (A_j(x_i))$$

$$\circ = \sup_{y} \inf_{x} (A_j(x_i))$$

R_k – множество правил.

Далее рассмотрим модель, построенную на основе мультисловарей, где в качестве входных параметров системы нечёткого вывода будем рассматривать 6 нечётких лингвистических переменных (см. модель 1). А в качестве выходных параметров – 3 нечетких лингвистических переменных, определяющих принадлежность текста к предметной области: *“Fuzzy” – F, “Logik” – L, “Mathematik” – M.*

В качестве терм-множества всех лингвистических переменных (ЛП) будем использовать множество *T1={“min”, “med”, “max”, “none”}* (рис. 3). При этом каждый из термов ЛП будем оценивать по шкале от 0 до 1, при которой цифре 0 соответствует наименьшая принадлежность терминов текста к определенному подсловарю, а цифре 1 – наибольшая.

После обработки 112 текстов по НЛ, 10 текстов по математике и 10 текстов по логике программой *Fuzzy Base* [4] были получены частотные характеристики, на основе которых по средним значениям относительных частот встречаемости терминов построен частотный портрет (рис. 4) [2].

Затем по данным усредненных частот групп текстов каждой предметной области были найдены минимальные и максимальные значения (табл. 1).

По данным этих усредненных частот были сформулированы 3 нечетких правила отнесения текста к предметной области Нечеткой логики, Математики и Логики (система нечёткого вывода типа Мамдани):

ПРАВИЛО 1: ЕСЛИ уровень относительной частоты терминов *F* в тексте – “средний” И уровень относительной частоты терминов *L* – “средний” И уровень относительной частоты терминов *M* – “минимальный” И уровень относительной частоты терминов *LT* – “минимальный” И термины словарей *C* и *KI* – отсутствуют, ТО степень уверенности, что текст принадлежит к пред-

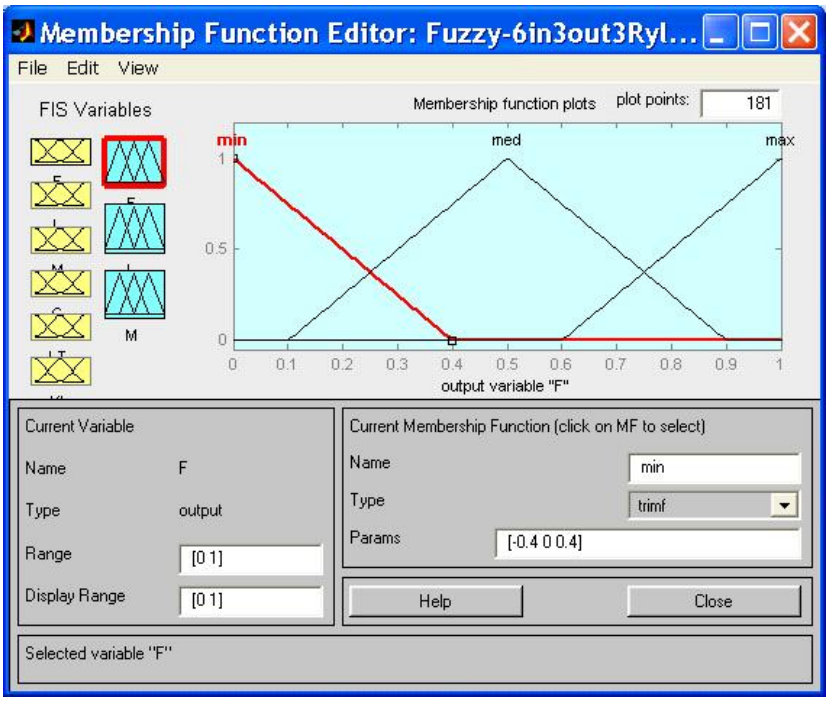

Рис. 3. Функции принадлежности выходной переменной F

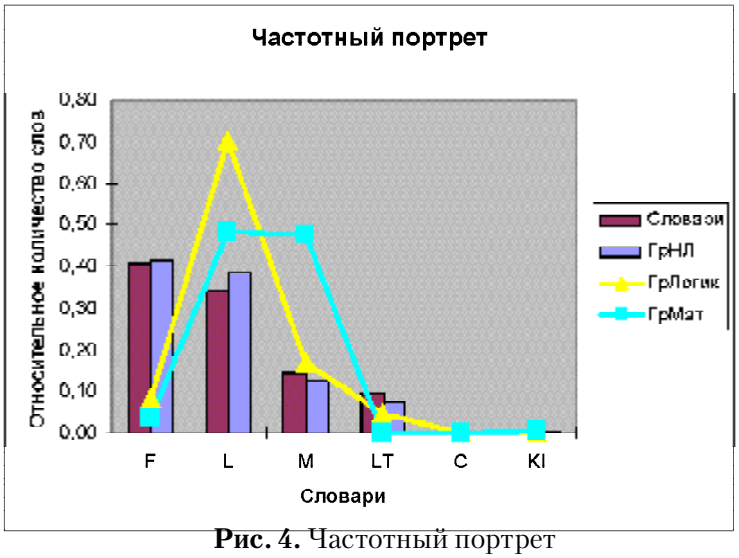

Рис. 4. Частотный портрет

Таблица 1. Усредненные частоты терминов разных предметных областей

	F	L	M	C	LT	KI
F						
Среднее	0,396278	0,407538	0,135755	0,000519	0,056154	0,003756
min	0,031	0,086	0,013	0,000	0,000	0,000
max	0,763	0,872	0,771	0,022	0,571	0,044
L						
Среднее	0,083	0,739	0,168	0,000	0,009	0,001
min	0,000	0,500	0,049	0,000	0,000	0,000
max	0,337	0,937	0,439	0,000	0,038	0,010
M						
Среднее	0,036	0,484	0,477	0,000	0,000	0,003
min	0,000	0,222	0,000	0,000	0,000	0,000
max	0,105	1,000	0,741	0,000	0,000	0,033

метной области *F* – максимальная.

ПРАВИЛО 2: ЕСЛИ уровень относительной частоты терминов *F* в тексте – “минимальный” И уровень относительной частоты терминов *L* – “максимальный” И уровень относительной частоты терминов *M* – “минимальный” И термины словарей *C, LT* и *KI* – отсутствуют, ТО степень уверенности, что текст принадлежит к предметной области *L* – максимальная.

ПРАВИЛО 3: ЕСЛИ уровень относительной частоты терминов *F* в тексте – “минимальный” И уровень относительной частоты терминов *L* – “средний” И уровень относительной частоты терминов *M* – “максимальный” И термины словарей *C, LT* и *KI* – отсутствуют, ТО степень уверенности, что текст принадлежит к предметной области *M* – максимальная.

Формализованное множество данных правил выглядит следующим образом:

R_1 – IF X_1 is “med” AND X_2 is “med” AND X_3 is “min” AND LT is “min” then Text is F.

R_2 – IF X_1 is “min” AND X_2 is “max” AND X_3 is “min” then Text is L.

R_2 – IF X_1 is “min” AND X_2 is “med” AND X_3 is “max” then Text is M.

В табл. 2 приведены эти 3 правила базы знаний, сформулированные на основе частотных портретов текстов.

Задача идентификации предметной области состоит в определении степени принадлежности определенного текста к предметной области НЛ на основе нечеткого вывода на базе построенной модели. Точность модели будем оценивать с помощью значения среднеквадратической невязки [5]:

$$R = \frac{1}{M}\sum_{j=1}^{M}(y_r - F(X_r))^2,$$

где *F(X)* – значение выхода нечеткой модели при значении входов, заданных вектором $X = [F, L, M, LT, C, KI]$, M – количество текстов, $\tilde{y} = 1$ – степень уверенности принадлежности текста к предметной области НЛ.

На вход модели в качестве степени уверенности принадлежности терминов текста к соответствующему словарю подаются относитель-

Таблица 2. Нечеткая база знаний модели типа Мамдани

Правила	F	L	M	C	KI	LT	Text F	Text L	Text M
1	med	med	min	none	none	min	max	none	none
2	min	max	min	none	none	none	none	max	none
3	min	med	max	none	none	none	none	none	max

ные частоты терминов всех подсловарей в каждом тексте.

Расчет выбранной оценки по первым десяти текстам приведен в табл. 3.

Математическая модель определения принадлежности текста предметной области на основе нечеткого вывода по Мамдани:

$$\tilde{y}_j = \tilde{A}(\{x_i\}_{i=1,6}) \circ R_j(\{x_i\}_{i=1,6}, \{y_j\}_{j=1,3}),$$

где $\tilde{y}$ – нечеткая выходная переменная;

x – входная переменная;

i – индекс для входов;

j – индекс для выходов;

Γ – функция принадлежности входных переменных $\{A_1, A_2, A_3, A_4, A_5, A_6\}$;

$\circ$ – знак композиции.

$\circ \equiv \vee \wedge \ (A_j(x_i))$

j=1,3; i=1,6

где $\vee$ – любая s-норма;

$\wedge$ – любая t-норма,

в нашем случае

$\vee$ – max – sup;

$\wedge$ – min – inf;

$\circ = \max\limits_{j=1,3} \min\limits_{i=1,6} (A_j(x_i));$

$\circ = \sup\limits_{y} \inf\limits_{x} (A_j(x_i));$

R_k – множество правил.

Использование математических моделе в САПР на промышленных предприятиях обеспечивается в ФНПЦ ОАО “НПО “МАРС” (г. Ульяновск) уже используется программное средство собственной разработки для автоматизации деятельности архивной служ-бы электронных информационных ресурсов (ЭИР). Однако функционал этого средства недостаточно широк. Требуется доработка данной системы с целью автоматизации части функций архивариусов и интеллектуализа-ции части процессов по управлению информацией. Расширением функци-онала данной системы является разработанный интеллектуальный сетевой архив электронных информационных ресурсов (ИСА ЭИР).

Ранее в подсистемах индексации применялись следующие модели:

взвешивание терминов;

“стоп-листы” – механизм уменьшения размерности индекса и шума вносимого в индекс документа за счет удаления наиболее часто употребляемыми терминами, предлогами;

“stemming” – приведение термов к основной форме;

“soundex” – механизмы, учитывающие опечатки и орфографические ошибки;

устранение проблем синонимии и омонимии [3].

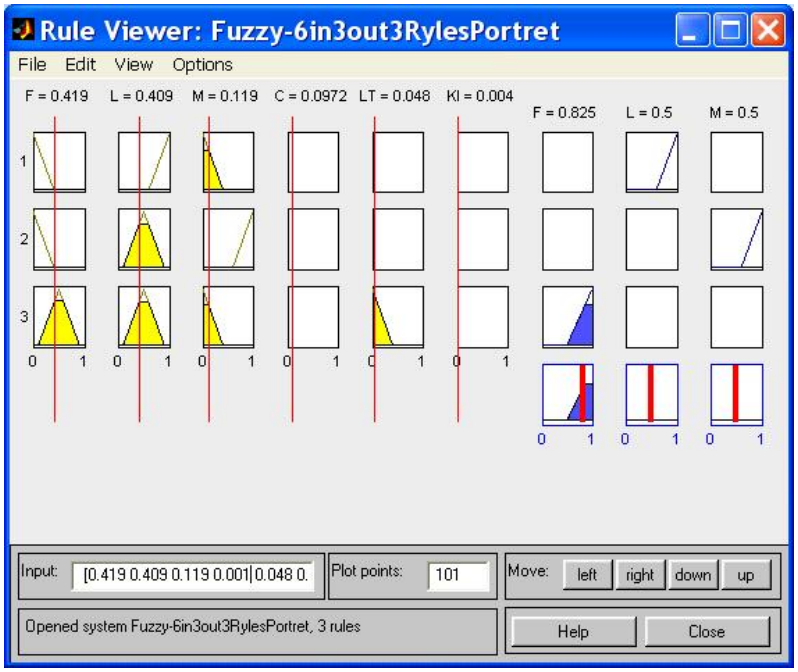

Рис. 5. Правила нечеткого вывода для вектора данных средних значений текстов по НЛ

Таблица 3. Определение среднеквадратической невязки текстов по НЛ

№	F	L	M	C	LT	KI	Text F	Text L	Text M
1993-1	0,256	0,395	0,326	0,000	0,023	0,000	0,775	0,5	0,5
1993-2	0,244	0,415	0,171	0,000	0,171	0,000	0,794	0,5	0,5
1993-3	0,206	0,235	0,382	0,000	0,176	0,000	0,758	0,5	0,5
1993-4	0,453	0,333	0,189	0,000	0,025	0,000	0,811	0,5	0,5
1993-5	0,293	0,414	0,150	0,000	0,143	0,000	0,807	0,5	0,5
1993-6	0,497	0,293	0,156	0,000	0,054	0,000	0,807	0,5	0,5
1993-7	0,053	0,342	0,289	0,000	0,316	0,000	0,5	0,5	0,5
1993-8	0,137	0,402	0,206	0,000	0,255	0,000	0,764	0,5	0,5
1993-9	0,155	0,397	0,207	0,000	0,241	0,000	0,769	0,5	0,5
1993-10	0,139	0,417	0,111	0,000	0,333	0,000	0,765	0,5	0,5
Среднее	*0,419*	*0,409*	*0,119*	*0,001*	*0,048*	*0,004*	*0,825*	*0,5*	*0,5*
Невязка							0,0676		

Одной из подсистем интеллектуального проектного репозитария является индексатор. Он отбирает из текста стоп-слова и на основе оставшихся терминов частично определяет предметную область документа.

Мы заменили в индексаторе словарь со стоп-словами на словарь-тезаурус, сформированный на основе анализа ТС. Тезаурус – это терминологический ресурс, реализованный в виде словаря понятий и терминов со связями между ними. Основное назначение тезауруса в нашей системе – определение предметной области: на основе связей тезауруса можно построить терминосистему, а навигация по связям тезауруса помогает получать на базе ТС точную идентификацию предметной области документа.

На первой стадии анализа в тексте ищутся термины, описанные в Тезаурусе (как слова, так и словосочетания). На основе связей Тезауруса термины группируются по смысловой близости во фреймы и подфреймы.

Каждый термин в тексте получает свою оценку релевантности относительно содержания документа, в зависимости от того, элементом какой ТС он является. Максимальный вес получают термины той ТС, которые встречались чаще, минимальный – упоминавшиеся термины. Иногда в тексте встречается минимальное количество терминов, но они настолько значимы, что текст необходимо отнести их именно к данной области. В этих случаях в программе используется коэффициент значимости термина, который можно настраивать.

Понятия с определенной таким образом оценкой релевантности образуют терминологический поисковый образ документа или тематическое представление содержания документа. Тематическое представление является основой для рубрицирования и аннотирования.

Для более точного определения предметной области документов необходимо было расширить словари ПО “Нечеткая логика”, распределить данные подсловаря “НЛ” по фреймам и построить иерархический словарь тезаурус.

Необходимо настроить словарь внутри одного подсловаря, разделив термины по узким темам для более точного определения ПО.

Иерархическая структура тезауруса подсистемы:

.....Нечеткие системы
.....I. Теория
.....Теория нечетких множеств
.....1.Определение множеств
.....1а. Визуальные графики
.....2. Виды алгебр
.....2а. Операции
.....Теория нечетких систем
.....3. Нечеткие правила (базы)
.....4. Схемы вывода по нечетким правилам
.....II. Приложения
.....5. Нечеткий контроль
.....6. Роботика
.....7. Экспертные системы
.....8. Информационные системы
.....8с. Нечеткие временные ряды
.....8а. Интернет
.....8б. Базы данных
.....9. Нечеткая кластеризация
.....III. Гибриды
.....10. Нечеткие системы + нейронные сети
.....11. Нечеткие системы + вероятностные сети
.....12. Нечеткие системы + генетические алгоритм

Теперь определим предметную область документов на основе фреймового словаря-тезауруса.

C:\IVP\FUZZY.BASE\Тексты на обработку\Neue Texte\199. Bratz.doc

1.Определение множеств	72
10. Нечеткие системы + нейронные сети	14
11. Нечеткие системы + вероятностные сети	2
2. Виды алгебр	4
2a. Операции	14
3.Нечеткие правила (базы)	32
4.Схемы вывода по нечетким правилам	16
5.Нечеткий контроль	10
7. Экспертные системы	3
8. Информационные системы	1
9. Нечеткая кластеризация	4
I. Теория	21
II. Приложения	1
Итого:	**368 из 6835**

Отчет по фреймам одного из обработанных текстов выглядит образом:

То есть можно сделать вывод, что текст не только принадлежит к предметной области Нечеткой логики, но и относится к фрейму “Определение множеств”, поскольку терминов этого фрейма существенно больше (табл. 4).

Таким образом, на основе иерархического фреймового словаря-тезауруса определение предметной области текстов стало более тонким, где документы можно относить не только к определенной предметной области, но и распределять их внутри нее.

Таблица 4. Определение ПО текста при помощи иерархического словаря-тезауруса

№	1	1a	2	2a	3	4	5	6	7	8	9	10	11	12	13	14	8a	8b	8c	I	II	Т-в	ПО
1	6	0	0	0	6	0	16	0	0	0	0	1	0	0	0	0	0	0	0	4	0	**33**	**5**
2	5	0	0	0	36	2	9	0	0	1	0	6	0	0	0	0	1	0	0	10	15	**85**	**3**
3	6	0	2	0	44	2	5	0	2	4	0	0	0	0	0	0	0	0	0	9	19	**93**	**3**
4	41	0	1	3	34	18	6	0	0	0	0	1	0	0	0	0	0	0	0	11	8	**123**	**3**
5	12	0	5	12	36	3	41	0	0	0	1	1	0	0	0	0	0	0	0	8	11	**130**	**5**
6	18	0	9	13	34	4	12	0	0	0	0	4	0	0	0	0	0	0	0	10	4	**108**	**3**
7	0	0	0	0	11	1	7	0	0	0	0	0	0	0	0	0	0	0	0	9	1	**29**	**3**
8	10	0	0	0	21	7	19	0	0	0	0	1	0	0	0	0	0	0	0	20	3	**81**	**3**
9	2	0	0	0	20	5	7	0	0	0	0	1	0	0	0	0	0	0	0	16	4	**55**	**3**
10	7	0	0	0	17	1	10	0	0	0	0	0	0	0	0	0	9	0	0	14	2	**60**	**3**

СПИСОК ЛИТЕРАТУРЫ

1. *Арзамасцева И.В., Евсеева О.Н.* Построение правил нечеткого вывода для идентификации текстов проблемной области // Информационные технологии: межвузовский сборник научных трудов. Ульяновск: УлГТУ, 2008. С. 14-21.
2. *Арзамасцева И.В., Евсеева О.Н.* Построение частотного портрета текстов проблемной области // Информационные технологии: межвузовский сборник научных трудов. Ульяновск: УлГТУ, 2008. С. 21-26.
3. *Наместников А.М.* Интеллектуальные проектные репозитории. Ульяновск: УлГТУ, 2009. 110 с.
4. Свидетельство о государственной регистрации программы для ЭВМ №2008615366 от 10.11.2008г. / *Арзамасцева И.В., Подгорный И.В.* М.: Федеральная служба по интеллектуальной собственности, патентам и товарным знакам.
5. *Штовба С.Д.* Идентификация нелинейных зависимостей с помощью нечеткого логического вывода в системе MATLAB // Exponenta Pro. Математика в приложениях. 2003. №2. С.9-15. URL: http://soft.mail.ru/journal/pdfversions/519588.pdf (дата обращения 25.02.2010).

MATHEMATIC MODELING OF THE TERMINOLOGY SYSTEMS OF DOCUMENTS IN CAD

© 2010 I.V. Arzamastseva

Ulyanovsk State Technical University

In this article were described mathematic models of the terminology systems and their use in CAD.

Key words: mathematic models, terminology systems, CAD.

Ivetta Arzamastseva, Senior Lecturer at the Applied Linguistics Department, Graduate Student at the Information Systems Department. E-mail: lingua@ulstu.ru.