УДК 519.7+681.3

ИНТЕЛЛЕКТУАЛЬНАЯ СИСТЕМА ИДЕНТИФИКАЦИИ ПРОБЛЕМНОЙ ОБЛАСТИ ТЕКСТА НА БАЗЕ НЕЧЕТКИХ ПРАВИЛ

И.В. Арзамасцева

ГОУ ВПО «Ульяновский государственный технический университет», г. Ульяновск

Рецензент В.Е. Подольский

Ключевые слова и фразы: веса терминов; идентификация проблемной области; интеллектуальная система; терминосистема «нечеткая логика».

Аннотация: Дано описание интеллектуальной системы идентификации проблемной области на примере терминосистемы «нечеткая логика».

Введение. Функции интеллектуальной системы

Для реализации интеллектуальной системы логического вывода по базе нечетких правил необходимо определить функции:

– представления в системе нечетких понятий (функций принадлежности);

– вычисления логических выражений условных частей правил с логическими связками И, ИЛИ;

– вычисления импликации;

– усреднения результата, получаемого по разным правилам путем композиции [2].

Описание области использования интеллектуальной системы

Оценим вышеописанный метод идентификации проблемной области на примере проблемной области нечеткой логики (**НЛ**). Терминосистема НЛ состоит из шести подсловарей, термины которых используются в научных текстах по нечеткой логике: «Нечеткая логика» (Fuzzy), «Логика» (Logik), «Математика» (Mathematik), «Управляющие системы» (LT), «Искусственный интеллект» (KI), «Компьютер» (C) [1].

В качестве входных данных были использованы тексты на немецком языке по нечеткой логике, которые образовали корпус текстов по НЛ.

Алгоритм действия программы представлен на рис. 1.

Арзамасцева И.В. – старший преподаватель кафедры «Прикладная лингвистика», аспирант кафедры «Информационные системы» УлГТУ, г. Ульяновск.

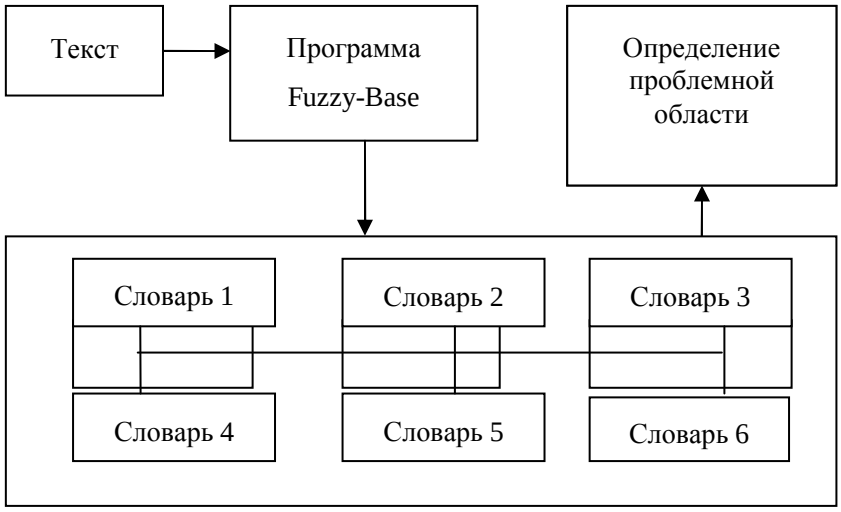

Рис. 1. Алгоритм действия программы

Реализация интеллектуальной системы идентификации предметной области текста

В ходе выполнения исследования была создана программа Fuzzy-Base, которая может быть использована по нескольким направлениям.

В программе даны все значения каждого слова, но на первом месте всегда будет находиться специальное значение. Например, die Ausgabe: 1) выход (Fuzzy); 2) выходное устройство (комп. тех.); 3) издание. В универсальном словаре на первом месте практически всегда будет стоять общее значение – «издание».

Программа легко обновляется и позволяет фиксировать самую последнюю словарную информацию по выбранной проблемной области. В дальнейшем планируется использовать программу для разработки автоматизированных словарей систем машинного перевода, а также для систем автоматического аннотирования и реферирования текстов.

Настройка интеллектуальной системы на проблемную область

Проанализируем один из файлов, который входит в корпус текстов по нечеткой логике. Если после обработки программа обнаружит несколько новых терминов, то можно сформулировать следующее нечеткое правило, на основе которого работает программа: ЕСЛИ слово состоит из терминов из подсловарей программы, ТО оно является термином.

В log-файле программы это выглядит следующим образом:

- добавь выражение: Fuzzy-Regelung;
- добавь выражение: Vor-und.

Теперь необходимо расставить коэффициенты значимости словарей, поскольку вес терминов из словаря Fuzzy превышает вес терминов из словаря Logik. Процентное соотношение встреченных терминов из разных подсловарей по отношению к общему количеству слов в исследуемом тексте высчитывается по формулам:

$$P_i = \frac{X_i\,100}{S} K_i; \quad P_1 = \frac{11 \cdot 100}{1775} \cdot 1 = 0{,}62; \quad P_2 = \frac{11 \cdot 100}{1775} \cdot 1{,}70 = 1{,}05,$$

где P_i – процентное соотношение терминов из различных подсловарей между собой по отношению к общему количеству слов в обработанном тексте; K_i – изменяемый вручную коэффициент веса терминов; X_i – количество терминов определенного подсловаря; S – общее количество (сумма) терминов, обнаруженных в тексте.

Посчитать веса терминов можно по следующей формуле:

$$V_i = \frac{P_i \cdot 100}{\sum_1^n P_i}; \qquad V_2 = \frac{1{,}05 \cdot 100}{2{,}82} = 37{,}2,$$

где V_i – процентное соотношение терминов из различных подсловарей между собой в тексте.

Изменив вручную коэффициент веса терминов НЛ, можно изменить идентификацию предметной области (рис. 2, 3)

$$V_1 = \frac{0{,}62 \cdot 100}{2{,}43} = 25{,}5.$$

Определить принадлежность текста к определенной проблемной области на основе наибольшего значения терминов из определенного словаря можно по формуле

$$\max(V_i) \in \mathrm{SL}_i,$$

где SL_i – определенный подсловарь.

Нам необходимо также определить систему нечетких правил для нашей программы. В качестве входной переменной было взято количество найденных программой терминов обработанного текста, разделенного по подсловарям. Выходной переменной является определение проблемной области текста, то есть относится ли текст к проблемной области НЛ или нет. Были разработаны интуитивные правила, которые были занесены в среду MatLab.

Таким образом, изменяя коэффициент веса терминов, можно получить нужный результат и правильно идентифицировать проблемную область.

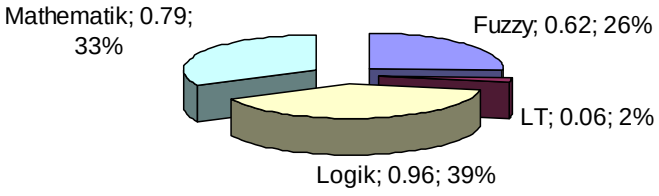

Рис. 2. Распределение терминов по словарям и вес каждого словаря

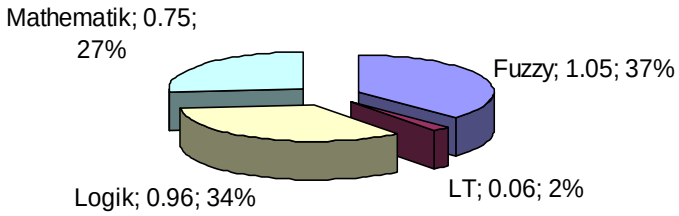

Рис. 3. Изменение веса терминов

Список литературы

1. Арзамасцева, И.В. Статистические исследования немецкоязычной терминосистемы «нечеткая логика» при помощи программы Fuzzy-Base / И.В. Арзамасцева. – Новосибирск, 2005. – С. 65–76.

2. Ярушкина, Н.Г. Основы теории нечетких и гибридных систем : учеб. пособие / Н.Г. Ярушкина. – М. : Финансы и статистика, 2004. – 320 с.

Intelligent System for Identification of Text Problem Domain on the Basis of Fuzzy Rules

I.V. Arsamastseva

Ulyanovsk State Technical University, Ulyanovsk

Key words and phrases: importance of terms; identification of problem domain; intelligent system; term system “Fuzzy-Logic”.

Abstract: The paper describes an intelligent system for identification of data domain by the example of the term system “Fuzzy-Logic”.
