АВТОМАТИЗИРОВАННЫЙ АНАЛИЗ ГЛОССАРИЯ УЧЕБНОЙ ДИСЦИПЛИНЫ В ИНТЕРАКТИВНЫХ ОБУЧАЮЩИХ СИСТЕМАХ

Н.В. Лутошкина, А.А. Высотин, М.А. Высотин

Сибирский государственный технологический университет, Красноярск, Россия, lutoshkinanv@list.ru

Abstract. This article describes automatized analysis of glossary's text, which was developed for the purpose of checking the glossary's adherence to the standards. In the course of study Porter's Stemmer algorithm was altered and the modified version was implemented. That enabled to carry out text analysis and, in consequence, to generate and visualize semantic network of the scientific concepts. As a result, the analysis of structure and hierarchy of the concepts given in the interactive course can be carried out.

Современные обучающие системы представляют собой интеллектуальные системы, основанные на парадигме обработки знаний. Понятия составляют содержание знаний, владение системой понятий необходимо в любой образовательной технологии. Научные понятия, включенные в программу дисциплины, образуют состав учебной системы знаний дисциплины. Для работы с понятийным аппаратом учебной дисциплины используется глоссарий, являющийся встроенным инструментом многих современных интерактивных обучающих систем.

Под глоссарием понимается контролируемый словарь[1], содержащий толкования специфичных терминов некоторой предметной области и поддерживающий таксономию. По сути, глоссарий – это иерархически структурированное множество терминов, описывающих предметную область, которое может быть использовано как исходная структура для базы знаний в дистанционном обучении. Отсюда следует важность контроля структуры глоссария и его составляющих.

Структура, правила разработки и форма представления одноязычных информационно-поисковых тезаурусов (глоссария как частного случая), ориентированных на лексику русского языка, изложены в [2]

Базовая модель одноязычного тезауруса введена в международном стандарте [3].

С целью автоматизации анализа глоссария была разработана программа, позволяющая, выделять связи «определяемое понятие – определяющее понятие» между содержащимися в нем понятиями. В результате этого может быть построена семантическая сеть в виде ориентированного графа, вершины которого обозначают понятия, изучаемые в курсе, а рёбра указывают на наличие указанной выше связи между понятиями.

В качестве входных данных для анализа используется текстовый файл, содержащий в себе глоссарий рассматриваемого учебного курса.

Во входном файле каждое определение должно быть записано в формате <определяемый термин> – <определение> и состоять из одного абзаца; термин может быть составным - в несколько слов, может содержать скобки (причём их содержимое не учитываются в дальнейшем анализе), регистр букв не имеет значение. Остальные строки, не подходящие под приведённый формат, игнорируются.

Термины, выделенные в ходе анализа, формируют множество концептов будущей семантической сети. Для определения связей между концептами производится анализ текста на предмет включения в определении одного из терминов других понятий этого глоссария. Любое понятие считается определяющим для данного термина, если оно встречается в тексте определения, не зависимо от словоформы (падежа, рода, числа, лица). Для этого осуществляется поиск не самих слов, а их основ (стемов), выделяемых

с помощью модифицированного алгоритма Портера. Алгоритм стемматизации Портера (или стеммер Портера) основывается на особенностях языка (в данном случае русского), отсекает окончания и суффиксы, применяя последовательно ряд правил. Специально для целей данной программы этот алгоритм был модифицирован в сторону повышения чувствительности к производным словам.

На первом этапе анализа текста формируется набор всех терминов, определённых в данном глоссарии. Далее, для каждого из них находятся основы (стемы). Для терминов, состоящих из нескольких слов – стемы каждого из слов. После этого формируются строки, содержащие стемы всех слов, из которых состоят определения, причём порядок слов (стемов) сохраняется, знаки препинания опускаются. В завершении, выделятся совпадения стемов концептов и фрагментов стемов определений, на основе чего строится матрица смежности ориентированного графа семантической сети.

В результате анализа текста глоссария может быть получен ориентированный граф – прототип будущей семантической сети. В случае, если исходный глоссарий был составлен неверно, данный граф может содержать циклы. С целью создания более удобного представления, граф рисуется с помощью утилиты dot из пакета Graphviz, а также, приводится текст глоссария со специально подсвеченными фрагментами, соответствующими обнаружению концептов в определениях.

Для удаления циклов из графа-прототипа удаляются некоторые рёбра. Программа сама способна предложить набор рёбер для удаления. При этом она последовательно, в несколько шагов, удаляет по одному ребру, причём на каждом шаге выбирается ребро, удаление которого приведёт к минимизации размеров циклов (по числу рёбер) на следующем шаге.

На выходе пользователь получает семантическую сеть в виде ациклического ориентированного графа. На его основе можно проводить анализ учебного материала данного курса.

На рисунке 1 показано окно программы на этапе нахождения и удаления циклов.

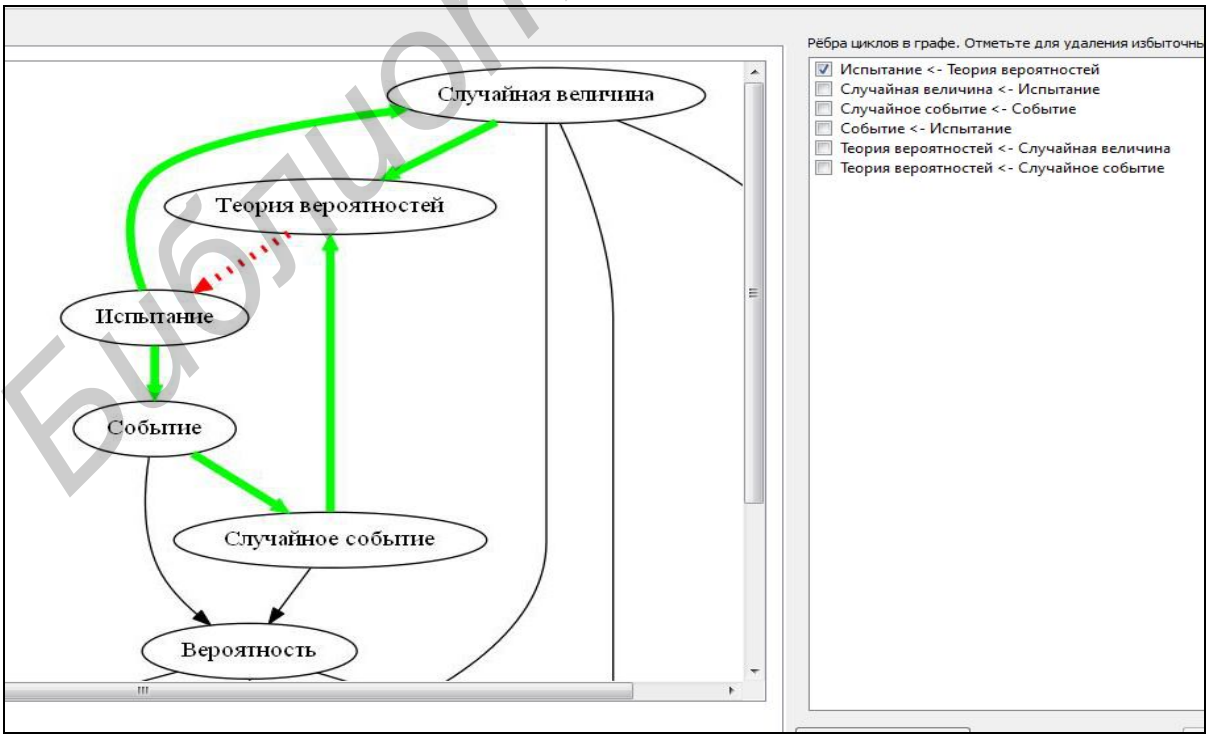

Рисунок 1 - Окно программы на этапе нахождения и удаления циклов.

Слева пользователь может видеть рисунок графа, на котором специально выделены (жирными линиями) найденные программой циклы; пунктирными линиями изображены рёбра, предложенные для удаления самой программой. Переключив вкладку, пользователь может свериться с исходным текстом глоссария, а затем произвести его коррекцию. Справа расположено перечисление всех рёбер, входящих в циклы, часть из которых (возможно и все) пользователь может удалить, чтобы получить ацикличный граф.

На рисунке 2 изображён конечный результат анализа текста глоссария – семантическая сеть понятий, представленная в виде ацикличного ориентированного графа.

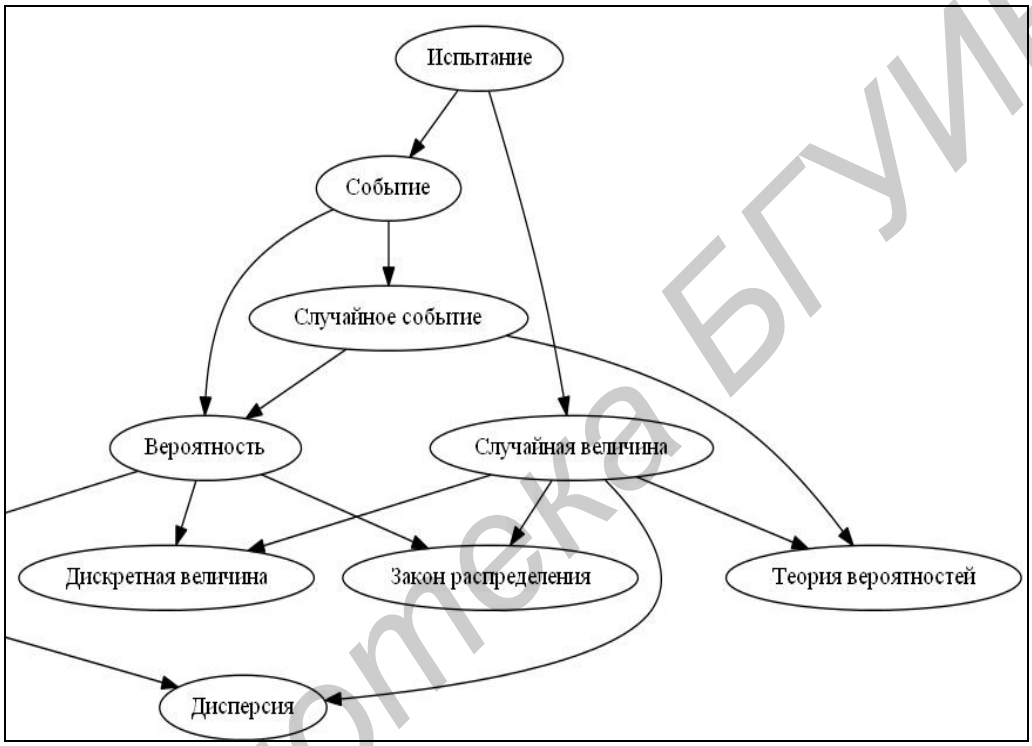

Рисунок 2 - Семантическая сеть понятий (фрагмент).

Данную программу можно использовать:

1. для контроля структуры глоссария на предмет нарушения структуры или ограничений требований ГОСТ 7.25-20011;

2. для построения семантической сети учебной дисциплины;

3. для отображения междисциплинарных связей, используя в качестве входного файла объединение нескольких глоссариев.

Литература

1. ANSI/NISO Z39.19-2005. Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies.

2. ГОСТ 7.25-2001. Тезаурус информационно-поисковый одноязычный. Правила разработки, структура, состав и форма представления.

3. ISO 2788:1986. Documentation – Guidelines for the establishment and development of monolingual thesauri.