Технология автоматизированного формирования понятийной структуры научного контента

© О.В. Окропишина

Национальный исследовательский ядерный университет «МИФИ»

kvizar@rambler.ru

Аннотация

Рассматривается технология автоматизированного формирования понятийных структур (представляющих предметные области на разных описательных уровнях) в виде последовательности статистических и лингвистических процедур обработки текста научной работы, основанная на уровневой семиотической модели представления предметной области научного исследования.

1 Введение

Электронная информация играет все большую роль во всех сферах жизни современного общества, в том числе в процессах образования и научной деятельности. Как показывают оценки, около 90 % документированной информации, циркулирующей в обществе, сосредоточенно в текстах на естественных языках, причем объем текстовой информации продолжает расти [4]. Как следствие, возникла повсеместная потребность в программных средствах, способных извлекать и обрабатывать информацию. Основу для создания таких программных средств, как правило, составляют различные методы формализованного представления текстовой информации, связанного с извлечением знаний [1 – 3, 7]. Основным назначением таких представлений является идентификация содержания, т. е. создание таких образов, которые за счет унифицированной формы представления обеспечивают «узнавание» оригинала в различных его проявлениях.

Особенности данного исследования в достижении цели идентификации содержания научного контента определяются, главным образом, классом задач, в рамках которого разрабатываются описываемые подходы, а также рассматриваемыми предметными областями, для которых строятся формализованные представления.

Предложенный в работе подход к построению специализированного представления знаний, который заключается в формировании структурированного представления текста в виде понятийной структуры, представленной на уровне терминов и связей между ними, разработан в рамках задач полнотекстового поиска научной информации: формулирования и реформулирования поискового запроса, создания поискового образа документа, представления и формирования предметных областей научных исследований. Предметную область исследований составляют определенным образом структурированные тексты научного стиля (авторефераты диссертаций, рефераты НИР, аннотации и т. д.).

Формирование набора функциональных отношений, которыми связаны термины понятийной структуры, производилось на основе анализа научно-производственной деятельности, сведения о процессах и результатах которой отражены в научном тексте.

Так как методы извлечения фактов при помощи сопоставления текста с набором регулярных выражений (образцов, шаблонов и т. д.), наиболее часто применяемые в рамках задач информационной разведки и др., не подходят для решения проблемы идентификации содержания научного контента (о чем будет сказано далее), необходимо было разработать технологию, отражающую особенности рассматриваемых класса задач и предметной области.

В основу технологии формирования понятийной структуры научного контента была положена разработанная уровневая семиотическая модель дескриптивного представления предметных областей научных исследований, на которой был основан ключевой этап технологии – этап семантической интерпретации полученных результатов анализа текста, также отличающий данный подход от других подходов по построению понятийных структур по тексту.

2 Основы технологии

В основу построения дескриптивного представления предметных областей в виде совокупности аспектных описаний положена уровневая семиотическая модель представления сведений о результатах познавательной деятельности, включающая уровни предметной области, концептуальный и знаковый.

На каждом из них описание выбранной предметной области может быть представлено как совокуп-

Труды 12[й] Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2010, Казань, Россия, 2010

ность элементов, которые определены на множестве характеристических признаков и связаны отношениями (объекты и отношения предметного уровня, концепты и концептуальные отношения, лексемы и лингвистические отношения) в рамках закона композиции, т. е. в рамках одного закона композиции получается одно представление предметной области, в рамках другого – другое, в котором есть как совпадающие с первым представлением, так и различающиеся элементы и отношения.

Концепция уровневой организации объектов использована в модели, исходя из общности физического и языкового описаний явлений внешнего мира. Текстовое описание предметной области научного исследования, отражая в своей структуре структуру явлений внешнего мира, подпадает под действие уровневого принципа.

Существенной особенностью информационного представления знаний автора текста является его семиотическая природа: свойства объекта и его поведение должны быть представлены знаковой системой.

Множество объектов реального мира представляется в языке множеством лексических единиц (знаков), каждая из которых именует предмет или ряд предметов и выражает некоторое понятие. Отражение предметов и понятий в языке иллюстрирует семиотический треугольник Фреге [11].

Семиотическая модель имеет свою проекцию на двухуровневое представление мира в рамках созидательной деятельности, состоящее из сознания и материи. Понятие – это смысл языкового знака или соответствующий знаку образ, формирующийся в сознании человека. Объекты, воплощающие соответствующие понятия, находятся на материальном уровне. Знаки в свою очередь также являются объектами действительности, но для дескриптивного представления предметной области необходимо рассмотрение знаков на отдельном уровне и не требуется рассмотрения связи, существующей между объектами на материальном уровне.

Таким образом, представление предметной области научного исследования можно построить на основании трехуровневой модели. В этом случае знаковое представление объекта является моделью, не изоморфной самому объекту, но, тем не менее, позволяющей идентифицировать его существенные свойства и связи.

С точки зрения общей теории систем, согласно [12], описание научного исследования может быть представлено как система $S = \{S_i\},\ i=1,..,n$, уровневых моделей дескриптивного представления предметной области научного исследования:

$$S_i = <M_i, A_i, R_i, Z_i>,$$

где i – аспект, который формально соответствует требованиям назначения и формы представления (и, таким образом, идентифицирован своим законом композиции Z_i), M_i – множество элементов, A_i – множество атрибутов, R_i – множество отношений. Иными словами, описание научного исследования в каждом из аспектов может быть представлено как совокупность элементов M_i, которые в рамках закона композиции Z_i определены на множестве характеристических признаков A_i и связаны отношениями R_i. Согласно [5, 18], такое описание научного исследования в частности можно назвать его онтологией.

Как было показано, знаковое представление предметной области позволяет идентифицировать существенные свойства и связи самого объекта. Для такого рода идентификации необходимо решить задачи анализа, формализации и структуризации текста и представления его (на знаковом уровне) как системы взаимосвязанных самостоятельных объектов, обладающих своими характеристиками. Затем становятся актуальными задачи перехода от знакового уровня к другим уровням представления, точнее, в связи с исключением из рассмотрения отношения знака к своему денотату, перехода от знакового уровня к концептуальному. Необходимость решения этих задач определила основу технологии формирования понятийной структуры научного контента.

3 Обзор подходов к формализации текстов

Как было выяснено, существует два подхода к решению задачи анализа, формализации и структуризации текстовой информации: статистический и лингвистический. Статистический подход включает различные методы, основанные на расчете весов слов; на определении частых наборов слов и объединении их в ключевые понятия и др. [2]. Лингвистический подход предполагает проведение различных видов лингвистического анализа текста с целью идентификации фактов в тексте и извлечении их характеристик [3].

Применение только статистического подхода позволяет получить лишь последовательность ключевых слов или ключевых понятий, составленных из частых наборов слов, и не дает возможности выделить связи в тексте, без которых невозможно отражение смысла. При данном подходе минимально сохранение семантики текста. Чаще всего статистический подход к извлечению ключевых понятий из текста используется как предварительный этап анализа текста для решения различных задач: классификации, кластеризации и др. Примерами программных средств, в основу работы которых положен этот подход, являются Autonomy IDOL Server, TextAnalyst и др. [14, 20, 22].

Применение только лингвистического подхода не позволяет выделить в тексте характеристические для данного текста признаки, а также замедляет и затрудняет решение задачи, так как нет возможности исключить из рассмотрения и не подвергать анализу слова, изначально не являющиеся кандидатами на информативные термины. При данном подходе под ключевыми понятиями подразумеваются все факты, встречающиеся в тексте. В лингвистическом подходе существует несколько направлений

[1, 3, 7 и др.], которые в большинстве своем являются модификациями одного наиболее распространенного, который заключается в извлечении фактов при помощи сопоставления текста с набором регулярных выражений (образцов, шаблонов и т. д.). Семантика текста при данном подходе сохраняется с помощью ключевых понятий, которые в основном являются именами существительными: имена и фамилии людей, названия организаций и др. Иногда выделенные факты связываются ссылками друг на друга. Этот способ наилучшим образом подходит для решения задач информационной разведки (сбора разрозненной и распределенной информации по определенной теме, проблеме, персоналиям). Для задач поиска научной информации эти факты в большинстве случаев даже не являются информационно значимыми (характеристическими) объектами, а представляют собой второстепенные сведения, служащие для более полного представления предметной области и используемые на самых низких уровнях абстракции представления. При этом связей типа «ссылок» не достаточно для построения понятийного образа научного текста, т. к. важны смысловые связи между ключевыми понятиями, которые отражают функциональные зависимости и тем самым несут дополнительную смысловую информацию по предметной области, которая зачастую не присутствует в тексте в явном виде. Программные средства КРОТ, Криминал, Аналитик и др., реализующие данный подход, представлены в [7, 15].

4 Технология построения понятийных структур

В результате сделанного обзора существующих решений был выбран смешанный подход, объединяющий статистический и лингвистический подходы. Он основывается на идентификации ключевых понятий в текстах и выделении отношений (связей) между ними и включает статистические методы расчета полезности слов.

На основании выбранного подхода для аналитической обработки научного контента разработана технология построения понятийной структуры, состоящая из трех этапов: анализа текста, семантической интерпретации результатов анализа и анализа понятий, составляющих понятийную структуру.

Схема представлена на рис. 1.

4.1 Этап анализа текста

На этапе анализа текста решается задача получения представления текста в виде семантической сети, в вершинах которой находятся лексемы, обозначающие информативные (ключевые) для данного текста слова и разделители, а дуги, соединяющие вершины, имеют смысл отношений между ними. Лексемам и разделителям приписываются графематическая информация (положение в тексте, длина, шрифт, язык и т. д.), морфологическая информация (часть речи и множество наборов граммем), статистическая информация (полезность слова) и синтаксическая информация (принадлежность к именной или глагольной группе). Меткам дуг присваиваются названия лингвистических отношений между ключевыми словами, обозначаемыми лексемами или разделителями, находящимися в вершинах, соединенных этими дугами.

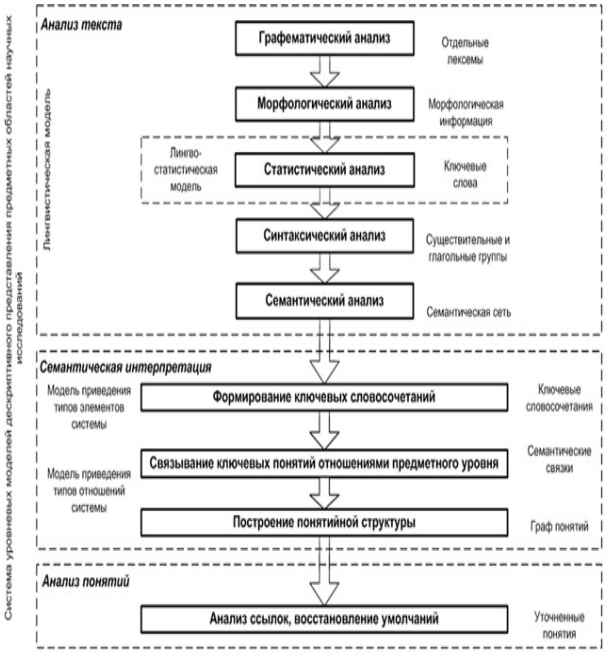

Рис. 1. Схема построения понятийной структуры научного контента

Процесс анализа текста описан лингвистической моделью. Анализ текста производится с целью определения характеристик каждого графического элемента текста, позволяющих однозначно идентифицировать его в пределах текста и относить к тому или иному классу, а также определения связей между элементами текста и на основании полученной информации построения семантической сети. Таким образом, объект анализа декомпозируется на систему взаимосвязанных самостоятельных объектов, обладающих своими характеристиками.

Для перехода от представления текста в виде цепочки ASCII-символов к представлению семантической сетью исходный текст должен последовательно пройти стадии графематического, морфологического, синтаксического и семантического анализа.

После прохождения стадии морфологического анализа проводится статистическая обработка полученной информации с целью выделения характеристических для данного текста признаков и исключения из рассмотрения и дальнейшего анализа слов, изначально не являющихся кандидатами на информативные термины.

Определение статистической информации и формирование множества информативных признаков происходят на основании лингво-статистической модели анализа текста, согласно которой ключевые слова определяются на основании расчета полезности слов-кандидатов на информативные термины. Все слова, которые не являются стоп-

словами, автоматически становятся словами-кандидатами на информативные термины. На основании результатов проведенного анализа методов расчета весов терминов [2, 16, 17, 21] для расчета полезности слов за основу была взята методика *TL*TF (Term Length*Term Frequency)*, одним из конкретных применений которой является мера Чена. Построенная функция полезности слов отражает зависимость «полезности» слова для читателя с точки зрения информативности, характеристичности данного слова от значений длины и частоты появления слова в тексте и имеет следующий вид:

$$u_j = \frac{l_{norm_j}}{\max\limits_{1 \le i \le N} l_{norm_i}} \cdot \frac{m_{norm_j}}{m},$$

где u_j – полезность слова j; l_{norm_j} – длина начальной (словарной) формы слова j; N – количество начальных форм слов документа; m_{norm_j} – количество предложений, в которых присутствуют слова с начальной формой слова j; m – число предложений в документе.

Выделение словаря информативных слов делается на основе граничных значений частот. После определения интервала термины, попадающие в него, составляют множество информативных терминов (ключевых слов). Множество информативных признаков расширяется за счет добавления слов, непосредственно связанных с ними лингвистическими отношениями.

Семантический анализ заключается в преобразовании обобщенной синтаксической структуры каждого предложения, полученной на этапе синтаксического анализа с учетом всего комплекса информации, приписанного каждой лексеме, в представление предложения семантической сетью.

Для связывания лексем, полученных в результате обработки текста, используются знания о лингвистических отношениях, которые берутся из русского общесемантического словаря (РОСС), описание которого приведено в [8, 19]. Семантическая сеть предложения получается путем соединения пар слов (в которых слова связаны лингвистическими отношениями) по совпадающим словам в пределах предложения.

4.2 Этап семантической интерпретации

На данном этапе происходит идентификация (восстановление) объектов и связей предметной области по ее знаковому представлению как системы взаимосвязанных самостоятельных объектов, обладающих своими характеристиками, путем перехода от знакового уровня к концептуальному, а именно, сведения элементов и отношений знакового уровня к понятиям и концептуальным отношениям. Также решается обратная задача визуализации отображения концептуального уровня представления описания предметной области на знаковом уровне.

Множества элементов всех уровней представления предметных областей представляют собой четверку следующего вида (или, в соответствии с определением Ю.А. Шрейдера [13], знаковую систему):

$$E = \langle L, K, D, I \rangle,$$

где L – множество лексических единиц (знаков); K – множество понятий, в которых описываются (представляются) концепты; D – множество объектов предметной области (денотатов); I – интерпретации, соотносящие знаку его концепт (в связи с исключением из рассмотрения отношения знака к своему денотату).

Для перехода от элементов знакового уровня к элементам концептуального уровня используется подход группирования лингвистических отношений по признакам их роли при связывании лексических единиц. Соотнесение знакам (согласно [10], именам понятий) их денотатов происходит в два этапа.

На первом этапе каждому знаку ставится в соответствие выражаемое им понятие (концепт имени понятия или несколько концептов в случае омонимии). Каждое понятие обозначает класс объектов предметной области, границы которой шире, чем границы предметной области описываемого научного исследования, или, согласно [10], класс денотатов имени этого понятия.

На втором этапе лингвистические отношения объединяются в кластеры по близости значений этих отношений, а затем получившиеся кластеры объединяются в две группы:

1) ***Группа отношений сужения объема понятия***

Объединение слов, связанных лингвистическими отношениями из этой группы, в словосочетание приводит к увеличению содержания нового понятия по отношению к обоим исходным понятиям, выражаемым словами, что соответствует уменьшению объема полученного понятия, т. е. класс объектов, воплощающих данное понятие, будет являться пересечением классов объектов, воплощающих два исходных понятия, таким образом, он будет меньше, чем классы объектов, воплощающих два исходных понятия (при условии, что исходные понятия не совпадают). Словосочетание, образуемое из имен этих понятий, будет именем нового понятия и будет обозначать класс объектов предметной области, отраженной в анализируемом тексте.

2) ***Группа отношений между разными понятиями***

Лингвистические отношения из этой группы показывают связь объекта, воплощающего данное понятие, с другим объектом, воплощающим другое понятие. Объединение этих понятий также может привести к увеличению содержания нового понятия, но оно выйдет за границы отраженной в тексте предметной области, т. е. пересечение классов объектов, воплощающих два исходных понятия, будет пустым в пределах предметной области описываемого научного исследования. Также из имен этих понятий не всегда можно составить словосочетание, так как слова, выражающие имена этих понятий, не

всегда непосредственно связаны в предложении грамматически.

Лингвистические отношения, не вошедшие в эти группы, и связанные ими слова исключаются из рассмотрения, так как слова, с которыми связано слово, воплощающее некоторое понятие, при помощи этих лингвистических отношений, являются именами отвлеченных понятий и служат в тексте для построения грамматически верных конструкций.

Таким образом, полученным на данном этапе словам и словосочетаниям (именам понятий) соотносятся их концепты через идентификацию информации, которую эти имена несут о денотатах.

Множества отношений всех уровней представления предметных областей научных исследований, по аналогии со знаковой системой, представляют собой четверку следующего вида:

$$O = <C, P, V, J>,$$

где C – множество лингвистических отношений (отношений знакового уровня); P – множество концептуальных отношений; V – множество отношений предметного (вещного) уровня; J – интерпретации, соотносящие лингвистическим отношениям концептуальные отношения (в связи с исключением из рассмотрения отношения знака к своему денотату).

Список концептуальных отношений был получен на основании функционального подхода к представлению научно-производственной деятельности, сведения о процессах и результатах которой отражены в научном тексте, а также на основании рассмотрения событий (действий, процессов и т. д.) предметной области с точки зрения функционального моделирования [6]. Важно учесть, что функциональные модели процессов разных предметных областей (а также в разных аспектах представления) отличаются между собой и, следовательно, перечни полученных функциональных связей также будут отличными. Список лингвистических отношений получен из специальной зоны словарных статей семантического словаря, в которой хранятся знания о способностях слов подчинять себе другие слова. Переход от лингвистических отношений к концептуальным происходит на основании бинарного нечеткого соответствия одних отношений другим.

Соответственно, на знаковом уровне визуализируется отображение концептуального уровня представления описания предметной области в виде графа понятий, представляющего собой ключевые компоненты понятийной структуры, связанные смысловыми отношениями.

Граф понятий наглядно представляет результаты выполненного исследования, а также дает представление о понятийной структуре предметной области научного исследования.

4.3 Этап анализа понятий

На этапе анализа понятий происходит уточнение полученных ранее ключевых компонент понятийной структуры путем восстановления умолчаний и анафорических ссылок, а также происходит уточнение смысловых отношений, связывающих ключевые компоненты понятийной структуры.

5 Текущие результаты

Рассмотренные подходы реализует созданный комплекс программно-информационных средств автоматизированного построения структурно-аналитических описаний диссертаций и НИР, языковая компетентность которого предусматривает возможность сопоставлять научному тексту его понятийную структуру в виде графа, в вершинах которого находятся термины, а дуги имеют смысл связей между ними.

Тестирование реализованных программных средств производилось на текстах авторефератов диссертаций и рефератов НИР. Был сформирован тестовый набор из ряда текстов (порядка 30) разной степени сложности по различным предметным областям. Тестирование проводилось авторами текстов или экспертами в соответствующих предметных областях. С точки зрения экспертов, полученные графы соответствовали тексту с точностью до наличия в нем скрытых связей и зависимостей, которые не могли быть выявлены автоматически. При этом подчеркивалаись правильность построения словосочетаний и корректность установления функциональных связей между построенными словосочетаниями, что экспериментально подтверждает достоверность полученных результатов. В качестве примера на рис. 2 приведен понятийный граф, автоматически построенный по тексту автореферата диссертации, где выделен аспект назначения.

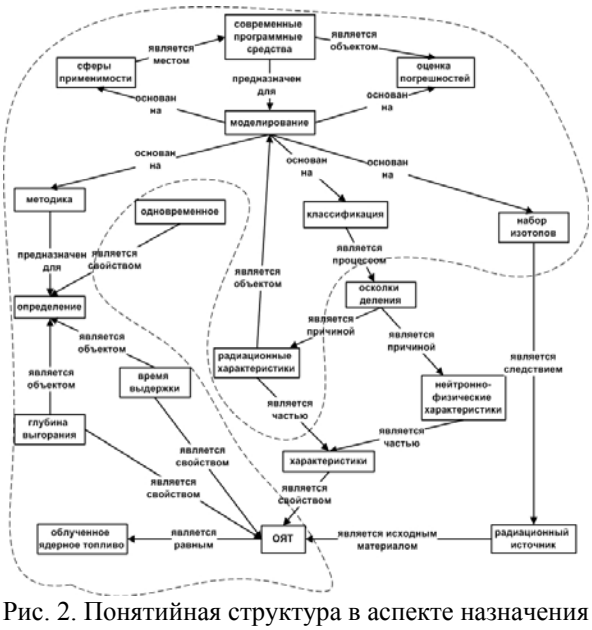

Рис. 2. Понятийная структура в аспекте назначения

Существенно, что при соотнесении автоматически построенного представления текста с реальным представлением автора или эксперта (его знаниями в данной предметной области) и внесении изменений и дополнений в автоматически полученную структуру выявлялись скрытые (неявные, не отра-

женные в тексте) знания автора или эксперта. Таким образом, происходило выделение понятийных связей, которые явно не присутствовали в анализируемом тексте.

Также изложенные подходы к построению системы формирования и использования научной информации были реализованы в рамках проекта «Разработка и внедрение информационно-аналитической системы регистрации, учёта, обработки и хранения отчётных документов по НИОКР с целью проведения мониторинга состояния и основных тенденций и направлений развития научных исследований и разработок, выполняемых компаниями государственного сектора, в том числе направленных на реализацию приоритетных направлений развития науки, технологий и техники в Российской Федерации, а также критических технологий Российской Федерации»[1].

В качестве программной основы использовалась документальная информационно-аналитическая система xIRBIS [9], интегрированная с системой лингвистического анализа АОТ.

6 Направления дальнейшей работы

В дальнейшем планируется разработать модели и методы представления построенной понятийной структуры в виде так называемого «навигационного» графа. Такие представления позволят пользователю перемещение (как по различным графам, так и по системе аспектных описаний предметных областей («в ширину»), так и в пределах разных описательных уровней представления предметных областей («в глубину») с запоминанием траектории.

Предполагается использование разработанного комплекса в составе информационно-поисковой системы, что сделает возможным представление всех текстов отдельно взятой документальной базы «навигационными» графами и формирования единого графа предметной области. Это, в свою очередь, позволит осуществить нахождение соответствий пользовательского запроса частям графа предметной области. При этом структура «навигационного графа» позволит сначала выделить ту его часть, которая наиболее точно соответствует запросу потребителя, а затем сформировать «траекторию прохождения» потребителем информационного массива для решения конкретной задачи.

7 Заключение

В машинном виде существуют различные, с точки зрения назначения, технологии создания и использования, формы (и виды публикаций) представления знаний, но все они, так или иначе, представляют собой тексты на естественных языках. Представленные таким образом знания существуют объективно и независимо от истории (контекста) их получения. Для того чтобы информация адекватно передавала в машинной форме знания автора, она должна фиксироваться в виде контекстно-обусловленных данных. Это возможно благодаря использованию интерактивных методов и средств построения формализованных представлений информации.

Использование уровневой семиотической модели дескриптивного представления предметных областей научных исследования позволяет нам утверждать, что знаковое представление предметной области дает возможность идентифицировать существенные свойства и связи самого объекта. Следовательно, применение технологии формирования понятийной структуры научного контента, в основу которой положена данная модель, позволяет строить средства автоматизированного построения формализованных представлений информации.

Литература

[1] Алексеев С.С., Морозов В.В., Симаков К.В. Методы машинного обучения в задачах извлечения информации из текстов по эталону// Электронные библиотеки: перспективные методы и технологии, электронные коллекции – RCDL'2009: Труды Всерос. науч. конф. – 2009.

[2] Алыгулиев Р.М. Математическое программирование в Text Mining// Электронные библиотеки: перспективные методы и технологии, электронные коллекции - RCDL'2005: Труды Всерос. науч. конф. – 2005.

[3] Барсегян А.А., Куприянов М.С., Степаненко В.В., Холод И.И. Технологии анализа данных: Data Mining, Visual Mining, Text Mining, OLAP. – СПб.: БХВ-Петербург, 2007.

[4] Вихнин А.Г., Сакипов Н.З. Штурм четвертого мегапроекта: кто будет новым Биллом Гейтсом? Системный анализ и выбор стратегии. – М.: «Диалог МИФИ», 2008.

[5] Джарратано Д., Гайли Г. Экспертные системы. Принципы разработки и программирования. – М.: Издательский дом «Вильямс», 2007.

[6] Дубейковский В.И. Практика функционального моделирования. – М.: «Диалог МИФИ», 2004.

[7] Кузнецов И.П., Мацкевич А.Г. Лингвистические и алгоритмические аспекты выделения объектов и связей из предметно-ориентированных текстов // Труды межд. конф. по компьютерной лингвистике и интеллектуальным технологиям «Диалог 2007», Бекасово, 2007. – С. 333-342.

[8] Леонтьева Н.Н., Кудряшова И.М., Соколова Е.Г. Семантическая словарная статья в системе ФРАП//ПГЭПЛ. – М.: Ин-т русского языка АН СССР, 1979. – Вып. 121. – С. 64.???

[9] Максимов Н.В. Документальная информационно-аналитическая система xIRBIS: программа для ЭВМ. / Максимов Н.В., Васина Е.Н., Голицына О.Л. и др. // Свидетельство о гос. регистрации №2008611511 от 25.03.2008.

[1] Федеральная целевая программа «Исследования и разработки по приоритетным направлениям развития научно-технологического комплекса России на 2007– 2012»

[10] Мейен С.В., Шрейдер Ю.А. Методологические аспекты теории классификации //Вопросы философии. – 1976. – №12. – С. 67-79.
[11] Попов Э.В. Общение с ЭВМ на естественном языке. – М.: Наука, 1982.
[12] Урманцев Ю.А. Общая теория систем: состояние, приложения и перспективы развития// Сборник «Система, Симметрия, Гармония». – М.: Мысль, 1988. – С. 38-124.
[13] Шрейдер Ю.А. Элементы семиотики. – М.: Знания, 1974.
[14] Autonomy IDOL. – http://www.autonomy.com/content/Products/IDOL.
[15] Avicomp. – http://www.avicomp.ru/.
[16] Buckley C., Allan J., Salton G. Automatic routing and retrieval using SMART: TREC-2// Inf. Proc.& Manag. – 1986. – V. 31, No 3. – P. 315-326.
[17] Chen Hsinchun, Ng Tobun D., Martinez Joanne, Bruce R., Schatz. A concept space approach to addressing the vocabulary problem in scientific information retrieval: an experiment on the Worm Community System// J. of the American Society for Information Science. – January 1997. – V. 48, No 1.
[18] Gruber T.R. Towards principles for the design of ontologies used for knowledge sharing// Int. Workshop on Formal Ontology, March, Padova, Italy, 1993.
[19] Leontyeva N.N. ROSS: semantic dictionary for text understanding and summarization //META. – 1995. – V. 40, No 1. – P. 178-184.
[20] Megaputer. – http://www.megaputer.com/products/ta/index.php3.
[21] Salton G., Zhang Y. Enhancement of text representations using related document titles// Inf. Proc. & Manag. – 1986. – V. 22, No 5. – P. 385-394.
[22] TextAnalyst. – http://www.analyst.ru/.

Technology of the aided formation of conceptual structure of scientific content

O.V. Okropishina

Technology of the aided formation of conceptual structures (that represents knowledge domains on different descriptive levels) in form of sequence of statistic and linguistic scientific work text processing procedures, based on level semeiotic model of representation of knowledge domain of scientific investigation are considered.