СИСТЕМА “ONTOGRID” ДЛЯ ПОСТРОЕНИЯ ОНТОЛОГИЙ

В.Д. Гусев, А.В Завертайлов, Н.Г. Загоруйко, С.П. Ковалёв, А.М. Налётов, Н.В.Саломатина

Институт Математики СО РАН, Новосибирск

zag@math.nsc.ru

Описывается проект инструментальной системы “OntoGRID” для автоматизации построения онтологий предметных областей с использованием GRID–технологий и анализа текстов на естественном языке. Рассматривается содержание и текущее состояние разрабатываемых блоков системы “OntoGRID”.

Введение

Онтологией (О) называется краткое описание структуры предметной области (ПрО), которое включает в себя термины (Т), обозначающие объекты и понятия ПрО, отношения (R) между терминами и определения (D) этих понятий и отношений:

O=<T,R,D>.

Построенная онтология предметной области будет полезна для совершенствования следующих областей деятельности:

1. **Системы обучения.** Действительно, для первого знакомства с предметной областью было бы очень полезно иметь в качестве «опорного сигнала» легко воспринимаемую структуру этой области. С помощью онтологии можно быстро находить ссылки на источники информации.
2. **Поисковые системы**. Наметившийся сейчас переход от поиска информации по ключевым словам к использованию семантически значимых фрагментов текстов существенно облегчается, если используется онтология ПрО.
3. **Научные исследования.** Большое значение имеет унификация терминологии ПрО. Наличие онтологии ПрО позволит автоматизировать процесс отслеживания полезных данных и знаний в потоке текущей информации.
4. **Системный анализ предметной области.** Онтология предоставляет структурированную и частично формализованную основу для проведения системного анализа предметной области.
5. **Интегрирование данных и знаний.** При объединении информационных баз онтология будет помогать устанавливать семантическую эквивалентность одинаковых фактов и понятий, сформулированных в разных терминах.

В отличие от большинства инструментов для построения онтологий [1] система OntoGrid оснащается двуязычным лингвистическим процессором для извлечения знаний из текстов на естественном языке.

Онтология только тогда будет принята научным сообществом, если в ее разработке участвовали широкие коллективы экспертов данной ПрО, географически удаленные друг от друга. Удобной технологической средой для реализации такого сотрудничества является GRID-сеть – распределенная информационно-вычислительная инфраструктура, построенная на основе технологии динамической интеграции вычислительных ресурсов.

Ниже описываются отдельные блоки разрабатываемой системы OntoGRID.

Создание лингвистической базы знаний

Работы, связанные с автоматическим анализом текстов, требуют определенного набора лингвистических и алгоритмических ресурсов. В настоящий момент нами реализованы: морфологическая база русского языка; блоки морфологического и статистического анализа; программы выделения устойчивых словосочетаний в тексте и выявления аномалий в позиционном распределении лексем по тексту.

Базой для морфологического анализа послужил электронный словарь Д.Уорта [2]. Процесс индексации (по Зализняку) большей части словаря был автоматизирован, он включает около 200 правил. Полученная *морфологическая база* содержит 3,2 млн. словоформ.

Статистический анализ основан на процедуре вычисления L–граммных характеристик текста (L–грамма – это цепочка из L подряд следующих слов). Частотной характеристикой порядка L текста T называется совокупность $\Phi_L(T)$ всевозможных L–грамм с указанием частот их встречаемости. Набор частотных характеристик $\Phi(T) = \{\Phi_1(T), \Phi_2(T), ..., \Phi_{L_{\max}}(T)\}$ определяет полный частотный спектр текста ($L_{\max}$ – длина максимальной повторяющейся цепочки слов в T). Аналогом $\Phi_L(T)$ для группы текстов $\overline{T} = \{T_1, T_2, ..., T_m\}$ является совместная частотная характеристика L–го порядка $\Phi_L(\overline{T})$, содержащая информацию об L–граммах, *общих* хотя бы для пары текстов. Совокупность $\Phi_L(\overline{T})$ ($L = 1, ..., L_{\max}(\overline{T})$) образует совместный

частотный спектр подборки $\overline{T}$ ($L_{\max}(\overline{T})$ – длина максимального межтекстового повтора). Совместные частотные характеристики служат основой для вычисления теоретико–множественных мер близости для пар и групп текстов [3].

Важную роль при анализе текстов играют устойчивые словосочетания [4]. В основе предложенного нами алгоритма выделения словосочетаний лежит последовательное вычисление частотных характеристик ($L = 2,3,...,L_{\max}$) и фильтрация повторяющихся L –грамм по критерию устойчивости [5].

Существенное значение при выявлении «ключевой лексики» играет информация о распределении слов в тексте. Слова с неравномерным распределением обычно оказываются более значимыми, чем распределенные равномерно. Повторы, многократно встречающиеся только в одном фрагменте текста, могут служить основой для выделения т.н. «сверхфразовых единств» [6]. Нами предложен новый метод выявления сверхфразовых единств, образуемых сгущениями лексических единиц определенного типа. Проведена его апробация на литературных и научных текстах [7].

Построение семантических сетей текстовых документов

Система анализа текста в интересах построения онтологии ПрО. Под системой анализа текста обычно понимается система, для которой определены следующие элементы: формализм для представления смысла текста; база лингвистических знаний (БЛЗ); отображение, переводящее текст в выбранный формализм; набор алгоритмов решения задач анализа текстов, использующих в качестве данных полученное семантическое представление; интерфейс эксперта, если предусмотрено его участие.

Среди классических задач, на решение которых ориентированны такие системы, можно упомянуть классификацию текстов, реферирование, семантически ориентированный поиск текстов по заданным концептам и др. Достаточно широкое распространение получил подход к анализу текста, опирающийся на онтологию, как на формальную модель ПрО. При этом система анализа текста проецирует онтологию на текст, выделяет в нем объекты из объема понятий ПрО и связи между ними. Для этого необходимо, чтобы в онтологию входило описание способов реализации понятий и отношений ПрО в текстах. Основной задачей системы анализа текстов при построении онтологии видится как раз автоматизация формирования проекции онтологии на ЕЯ тексты (ПроекОнт). Исходя из этого, авторы накладывают следующие требования на БЛЗ и систему в целом.

1. На начальном этапе БЛЗ должна представлять собой зачаток ПроекОнт, необходимый для начала функционирования системы. Этот зачаток вносится экспертом.
2. В системе должны быть реализованы механизмы развития БЛЗ в ходе анализа потока текстов ПрО, а также возможность контроля этого развития экспертом. На каждом этапе развития, БЛЗ должна являться некоторым приближением к ПроекОнт, на основе которого можно решать задачи анализа текста. На некотором уровне развития, БЛЗ должна содержать в себе ПроекОнт.
3. Структура и содержание БЛЗ системы должны быть удобны как при построении семантических представлений текстов, так и при дальнейшем анализе этих представлений.

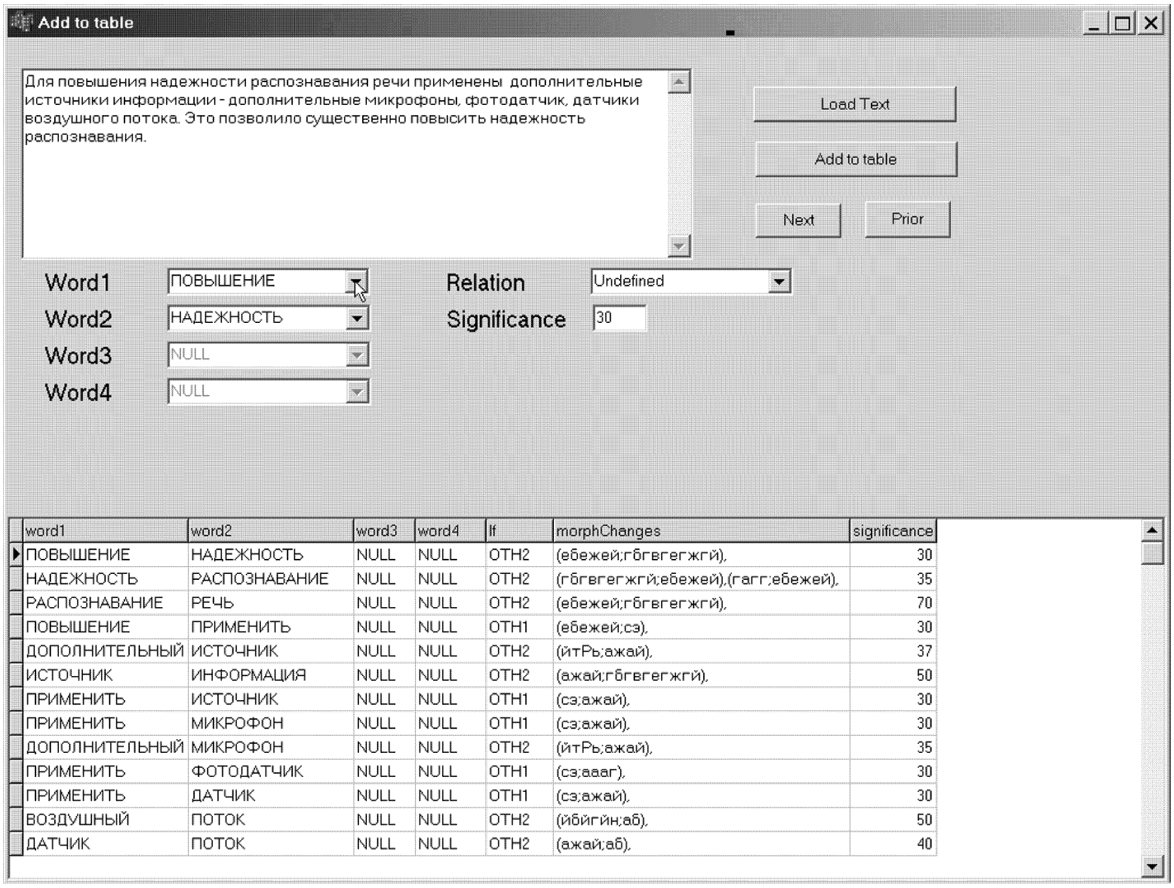
Система анализа текста САТ. В соответствии с изложенными требованиями, авторами разрабатывается и реализуется система анализа текста САТ. В качестве формализма для представления смысла текста в ней используются семантические *Q*-сети [8], в основе которых лежат аппарат пирамидальных сетей (ПС) Гладуна В.П. [9] и семантические представления Кузнецова И.П. [10]. Среди достоинств ПС следует упомянуть развитые ассоциативные свойства, иерархичность и, что особенно важно, в них реализованы процессы формирования связей между семантическими объектами, выделения классов объектов и ситуаций, а также процессы формирования обобщенных определений этих классов. В семантических представлениях Кузнецова И.П., в свою очередь, все части текста, соответствующие существенным единицам ПрО, вне зависимости от частоты их появления в текстах, отражены в сети соответствующими фрагментами. Принадлежность *Q*-сетей к обоим упомянутым классам дает возможность реализовать на них удобные алгоритмы анализа текста, предложенные Гладуном В.П. и Кузнецовым И.П..

Элементы ПрО описываются в ЕЯ текстах элементарными и составными словосочетаниями. Первые обычно состоят из двух слов (анализ данных) и являются реализациями элементарных отношений (r = свойство - объект). Вторые (интеллектуальный анализ данных) можно представить в виде комбинации элементарных словосочетаний (интеллектуальный анализ, анализ данных). Они, в свою очередь, являются реализациями составных отношений. Понятия ПрО также выражаются словосочетаниями (одним словом – наименованием понятия, элементарным словосочетанием или комбинацией элементарных словосочетаний).

БЛЗ САТ представляет собой набор элементарных и составных словосочетаний предметной области. БЛЗ удобно использовать в том же виде, который имеют семантические представления текстов, т.е. в виде *Q*-сети, каждый фрагмент которой соответствует реализации

некоторого отношения (т.е. некоторому словосочетанию). Если реализация отношения (словосочетание A') включает в свой состав реализацию другого отношения (словосочетание B'), то фрагмент сети A включает в себя фрагмент сети B. Условно БЛЗ САТ можно разделить на базу реализаций элементарных отношений (БРО) и набор критичных фрагментов (НКФ), по которым можно определить, какие элементы онтологии затрагиваются в данном тексте.

Формирование БРО. Начальный объем знаний в виде реализаций элементарных отношений ПрО вносится в БРО экспертом либо в ходе интерактивного анализа текстов предметной области (рис. 1), либо отсеиванием из набора устойчивых словосочетаний предметной области, полученного статистическими методами обработки текстов [5].



word1	word2	word3	word4	lf	morphChanges	significance
ПОВЫШЕНИЕ	НАДЕЖНОСТЬ	NULL	NULL	OTH2	(ебежей;гбгвгегжгй),	30
НАДЕЖНОСТЬ	РАСПОЗНАВАНИЕ	NULL	NULL	OTH2	(гбгвгегжгй;ебежей),(гагг;ебежей),	35
РАСПОЗНАВАНИЕ	РЕЧЬ	NULL	NULL	OTH2	(ебежей;гбгвгегжгй),	70
ПОВЫШЕНИЕ	ПРИМЕНИТЬ	NULL	NULL	OTH1	(ебежей;сэ),	30
ДОПОЛНИТЕЛЬНЫЙ	ИСТОЧНИК	NULL	NULL	OTH2	(йтРь;ажай),	37
ИСТОЧНИК	ИНФОРМАЦИЯ	NULL	NULL	OTH2	(ажай;гбгвгегжгй),	50
ПРИМЕНИТЬ	ИСТОЧНИК	NULL	NULL	OTH1	(сэ;ажай),	30
ПРИМЕНИТЬ	МИКРОФОН	NULL	NULL	OTH1	(сэ;ажай),	30
ДОПОЛНИТЕЛЬНЫЙ	МИКРОФОН	NULL	NULL	OTH2	(йтРь;ажай),	35
ПРИМЕНИТЬ	ФОТОДАТЧИК	NULL	NULL	OTH1	(сэ;аааг),	30
ПРИМЕНИТЬ	ДАТЧИК	NULL	NULL	OTH1	(сэ;ажай),	30
ВОЗДУШНЫЙ	ПОТОК	NULL	NULL	OTH2	(йбйгйн;аб),	50
ДАТЧИК	ПОТОК	NULL	NULL	OTH2	(ажай;аб),	40

Рис. 1 Формирование БРО в ходе интерактивного анализа текста

В первом случае эксперт имеет возможность, переходя от одного предложения текста к другому, выбирать наименование элементарного отношения («Relation») и его аргументы, число которых может варьироваться («Word1»,…) в выпадающих списках. Каждое словосочетание (реализация отношения) характеризуется значениями набора признаков. В текущей версии в него кроме лексем – аргументов и наименования отношения входят набор сочетаемости аргументов по морфологическим признакам (заполняется системой с помощью компоненты морфологического анализа [12]) и экспертная оценка значимости словосочетания в рамках предметной области. В ходе анализа текста в первую очередь следует выделять наиболее значимые для ПрО фрагменты, далее расширяя их менее важными подробностями. Для обеспечения устойчивости экспертных оценок можно применить интерфейс «Визуализатор отношений» [8,13]. Авторами реализован алгоритм построения *Q*-сетей, опирающийся на БРО. Пример его работы и соответствующая *Q*-сеть показаны на рис. 2.

Формирование НКФ. Для формирования НКФ составляется обучающая выборка текстов, для каждого из которых экспертом указывается, какие элементы онтологии в нем затронуты. По этим текстам строится общая *Q*-сеть. Далее, на основе алгоритма формирования понятий [9], происходит разделение семантических представлений текстов, затрагивающих разные элементы онтологии ПрО. В ходе этого разделения и получается упомянутый набор критичных фрагментов. Таким образом, каждый класс текстов описывается набором фрагментов, которые типичны или, наоборот, нетипичны для семантических представлений текстов данного класса. При этом соблюдается принцип наибольшей краткости описания (используются фрагменты, занимающие

минимально возможное по высоте место в иерархии сети).

Между критичными фрагментами, входящими в описание класса текстов, и элементами онтологии, которые в этом классе затрагиваются, можно провести параллель. При этом логично предположить, что сочетания слов, характеризующие класс текстов, затрагивающих определенные элементы онтологии, и будут отображением этих элементов на текст. Когда онтология еще не построена, лингвистическая база знаний используется системой анализа текстов (САТ) как приближение ПроекОнт. Когда онтология построена, такая база становится ее частью.

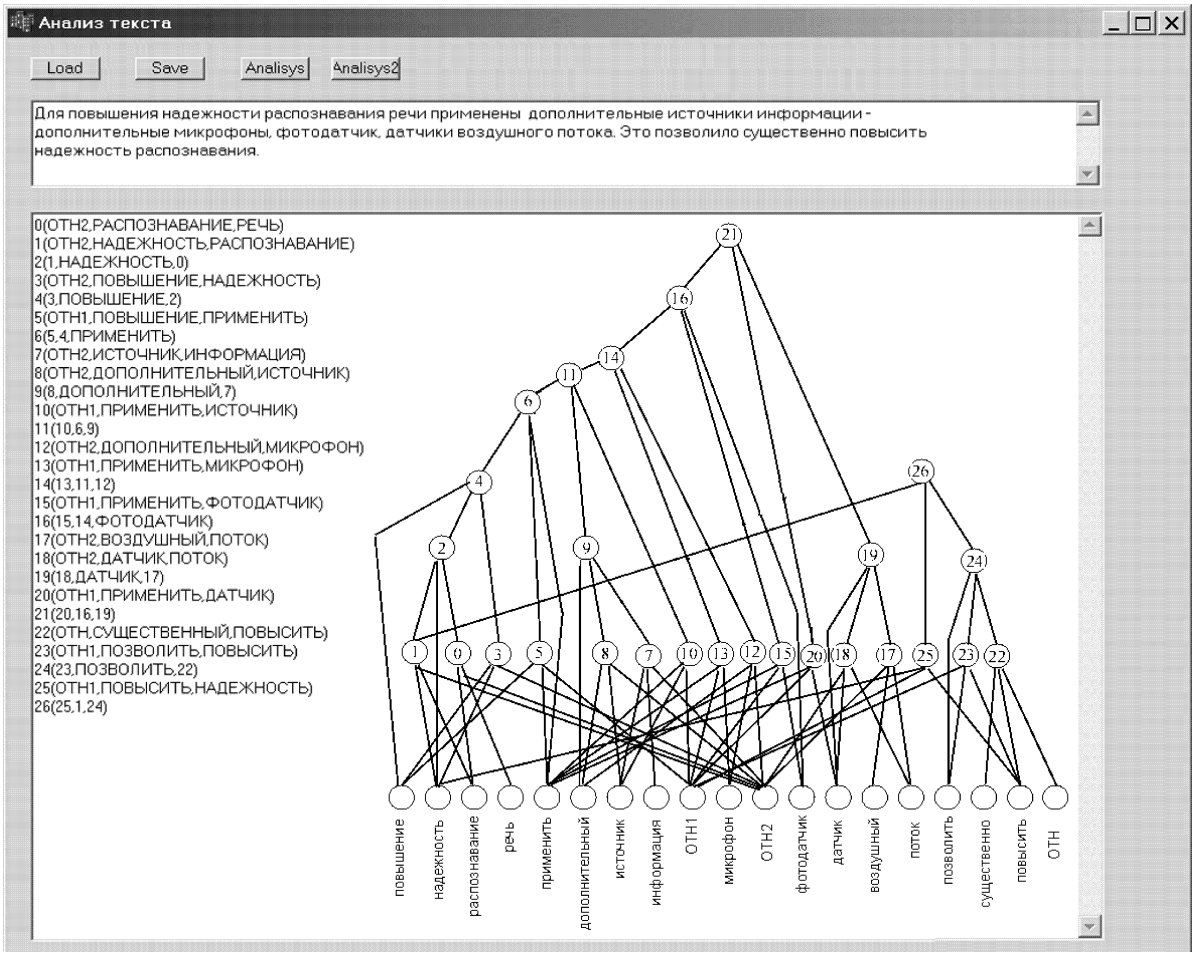


Рис. 2 Анализ фрагмента текста и соответствующая этому фрагменту Q-сеть

САТ и набор критичных фрагментов могут использоваться и для поддержки существующей онтологии. Соотнесение потока семантических портретов новых текстов с базой значимых фрагментов осуществляет наполнение элементов онтологии ссылками на текстовые документы. По степени «наполнения» эксперт может принимать решение о разделении «перегруженных» элементов сети и объединении «недогруженных».

При анализе текста часты ситуации, когда вершины фрагментов сети именуются формально разными словами, связанными, однако, отношениями синонимии, гиперонимии/гипонимии (родовидовые), меронимии (часть - целое). Поэтому в системе должен быть предусмотрен выход на тезаурусы WordNet, RussNet (для обще языковой лексики), а также на специальные тезаурусы для предметной лексики.

Создания и развития онтологии в GRID-сети

Структура системы автоматизированного построения онтологий «OntoGRID» должна отражать специфику трех типов ее клиентов: Эксперт, Пользователь и Администратор [14]. С точки зрения представления, создаваемая онтология - это комплект документов определённой структуры. Процесс построения онтологии состоит из итераций по дополнению и изменению этого комплекта документов. По результатам проведения ряда итераций администратор принимает решение о завершении очередного этапа процесса построения онтологии и публикации ее очередной стабильной версии.

Требования к системе, поддерживающей создание и обработку документов онтологии в

распределенном режиме, определяются коллективной работой разрозненных коллективов экспертов. Адекватную основу для построения систем, удовлетворяющих таким требованиям, предоставляют GRID-технологии [15], из которых наиболее интересной представляется архитектура OGSA (Open Grid Services Architecture), основанная на концепции веб-сервисов.

Для представления структуры онтологии был принят стандарт **OWL** (Ontology Web Language) [16], разработанный и рекомендованный консорциумом W3C. Язык OWL предназначен для представления информации, которая содержит **знания,** а не только **представление,** и предназначена для автоматической обработки **компьютер**ными программами, в противоположность использованию знаний непосредственно **человек**ом.

OWL обладает большей выразительной силой, чем такие структурные языки как XML, RDF и RDF-S, и может быть представлен в их форме. OWL-документ позволяет, используя лежащую в основе OWL дескриптивную логику, выводить такие факты о сущностях предметной области, которые не содержатся непосредственно в этом документе. В нашем проекте используется представление онтологии в нотации OWL-RDF.

Для упрощения разработки новых онтологий удобно создавать шаблоны онтологий различных групп предметных областей. Данный проект ориентирован на построение шаблона онтологий научно-технических предметных областей, связанных с процессами анализа, синтеза и преобразования информации о произвольных фрагментах реального мира. К числу таких процессов относятся измерение и накопление данных, обнаружение закономерностей (знаний), хранение, обработка и передача данных и знаний, использование знаний для прогнозирования и синтеза. На рис. 3 приведен перечень базовых категорий онтологий проблемных областей такого рода.

Рис. 3 Базовые категории онтологий научно-технических предметных областей

В дополнение к основному содержанию онтологии возможно формирование и хранение метаинформации о реквизитах автора и соавторов, времени создания и публикации, источниках информации и т.д.

В ходе коллективной работы над онтологией формируется распредёленная информационно-вычислительная среда, топология которой отражает структуру глобальных потоков информации, порождаемых в ходе исследования предметной области. В качестве индивидуального средства работы эксперта с фрагментом онтологии планируется использовать редактор, разработанный группой Protege Project [18], который обеспечивает удобный визуальный контроль процесса разработки фрагментов онтологии. Он дополняется средствами доступа к системе автоматизированного анализа текстов, описанной в предыдущих разделах.

Текущая рабочая версия онтологии хранится в центральном репозитории под управлением сервиса, который обеспечивает выделение фрагментов онтологии экспертам для разработки. После проведения доработок эксперт возвращает модифицированный фрагмент онтологии репозиторию для внесения изменений в рабочую версию онтологии. По решению администратора

сервис репозитория публикует очередную стабильную версию онтологии. Этот документ предоставляется пользователям в качестве текущей версии онтологии.

Описанная структура представления информации и архитектура ключевых сервисов были успешно апробированы в ходе создания **прототипа.** В настоящее время ведутся работы по дальнейшей реализации системы OntoGRID.

Заключение

Параллельно с описанными выше исследованиями по созданию инструментальной системы OntoGRID ведется подготовительная работа по организации виртуального коллектива экспертов из различных исследовательских центров, занимающихся проблемой «Интеллектуальный Анализ Данных» (Data Mining) для совместной разработки онтологии этой предметной области. В настоящее время для обслуживания коллективной работы на сервере Института Математики СО РАН создается сайт на русском и английском языках.

Авторы выражают благодарность Борисовой И.А., Дюбанову В.В., Кутненко О.А., Соколовой А.П. и Чуриковой В.А. за активное и полезное участие в обсуждении вопросов, затронутых в докладе.

Список литературы:

1) http://xml.com/2002/11/06/Ontology_Editor_Survey.html
2) Worth D.S, Kozak A.S., Johnson D.B. Russian Derivational Dictionary. //. New York: American Elsevier Publishing Company Inc, 1970.
3) Гусев В.Д. Механизмы обнаружения структурных закономерностей в символьных последовательностях // Проблемы обработки информации. Новосибирск: Вычислительные системы,1983. Вып. 100. С. 47–66.
4) Белоногов Г.Г., Быстров И.И., Новоселов А.П. и др. Автоматический концептуальный анализ текстов // НТИ, 2002. сер. 2. № 10. С. 26–32.
5) Гусев В.Д., Саломатина Н.В. Алгоритм выявления устойчивых словосочетаний с учетом их вариативности (морфологической и комбинаторной) // Труды международной конференции Диалог'2004. М.: Наука, 2004. С. 530 – 535.
6) Пащенко Н.А., Кнорина Л.В., Молчанова Т.В. и др. Проблемы автоматизации индексирования и реферирования // Итоги науки и техники. Информатика, 1983, Т. 7. С. 7–165.
7) Гусев В.Д., Немытикова Л.А., Саломатина Н.В. Выявление аномалий в распределении слов или связных цепочек символов по длине текста // Интеллектуальный анализ данных. – Новосибирск: Вычислительные системы, 2002, Вып. 171. С. 51–74.
8) Загоруйко Н.Г., Налетов А.М., Соколова А.А., Чурикова В.А.. Формирование базы лексических функций и других отношений для онтологии предметной области // Труды международной конференции Диалог-2004. М.: Наука, 2004. С.202-204.
9) Гладун В.П. Планирование решений //. Киев.: .Наукова думка, 1987. С.17-51.
10) Кузнецов И.П. Семантические представления. // М.:Изд. Наука, 1986.
11) Саломатина Н.В. Количественные характеристики вариативности морфемных моделей (на материале словаря канонических форм русского языка) // Методы обнаружения эмпирических закономерностей .Новосибирск: Вычислительные системы, 2001. Вып.167. С. 93–114.
12) Сокирко А.В. Морфологические модули на сайте www.aot.ru // Труды международной конференции Диалог-2004. М.: Наука, 2004. С.559-564.
13) Загоруйко Н.Г. Метрологические свойства эксперта.// Обнаружение эмпирических закономерностей. Новосибирск: Вычислительные системы, 1999. Вып. 166. С.119-128.
14) Завертайлов А.В., Ковалев С.П. Система поддержки деятельности распределенных экспертных групп по разработке онтологий предметных областей // Труды Международной конференции по вычислительной математике <МКВМ-2004>. Рабочие совещания. Новосибирск: ИВМиМГ СО РАН, 2004. С. 56-65.
15) Grid Computing: Making the Global Infrastructure a Reality. // N. Y.: Wiley & Sons, 2003. 16. Smith M.K., Welty C., McGuinness D.L. OWL Guide. W3 Consortium, 2004.
16) http://www.w3.org/TR/owl-guide/.
17) Powell A., Johnston P. Guidelines for implementing // Dublin Core in XML. DCMI, 2003. http://dublincore.org/documents/dc-xml-guidelines/.
18) Protege Project. http://protege.stanford.edu