ОРГАНИЗАЦИЯ КАРТОГРАММЫ ЗНАНИЙ В СИСТЕМЕ ЛОГИСТИКИ ЗНАНИЙ «ИНТЕГРАЦИЯ»

М. П. Пашкин

Санкт-Петербургский институт информатики и автоматизации РАН
199178, Санкт-Петербург, 14-я линия В.О., д.39

<michael@iias.spb.su>

УДК 681.3
М. П. Пашкин. **Организация картограммы знаний в системе логистики знаний «Интеграция»** // Труды СПИИРАН. Вып. 2, т.1. — СПб.: СПИИРАН, 2004.
Аннотация. *Показана актуальность задачи индексирования знаний для их оперативного поиска в системах управления знаниями. Представлен подход решения задачи индексирования знаний при помощи картограмм, разработанный в рамках подхода к построению систем логистики знаний «Сеть Источников Знаний» («СИЗ»). Представлена формальная модель картограммы знаний и требования к её построения.* — Библ. 11 назв.

UDC 681.3
M. P. Pashkin. **Organization of knowledge maps in the knowledge logistic system “Integration”** // SPIIRAS Proceedings. Issue 2, vol. 1. — SPb.: SPIIRAS, 2004.
Abstract. *The paper describes importance of the task of knowledge indexing for knowledge search in the knowledge management systems. It presents a decision of this task with use of knowledge map in the framework of the developed scientific approach to knowledge logistics “Knowledge Source Network” (“KSNet”). The formal model of knowledge map and requirements to its building are presented*. — Bibl. 11 items.

Введение

Развитие информационных технологий привело к накоплению большого количества разрозненных и разнородных электронных информационных ресурсов (документов, баз данных и знаний, электронных библиотек и т.п.). С появлением сетевых технологий и Интернета стал возможен обмен информацией, находящейся в данных ресурсах, с целью ее повторного и совместного использования. Одновременно с этим процессом возникла проблема выбора и отбраковки информации для решения конкретной задачи или принятия решений. Одна из основных подзадач, решаемых в данной проблемной области — быстрый поиск релевантной информации в постоянно расширяющихся относительно типов и содержимого информационных ресурсов и предоставление её по запросу в кратчайшее время.

В ходе исследования, посвящённого построению систем интеграции знаний, был разработан подход «Сеть Источников Знаний» («СИЗ») [1–3]. Предложенная в рамках проекта методология — «логистика знаний» — представляет собой новое научное направление в области управления знаниями. Она ориентирована на извлечение/приобретение, интеграцию/ обработку и передачу/транспортировку адекватных знаний из распределенных источников в нужном контексте нужным пользователям в нужное время для решения актуальных

[1]Фрагменты данного научного исследования были выполнены в рамках проекта 2.44 научно-исследовательской программы «Математическое моделирование и интеллектуальные системы» и проекта 1.9 научно-исследовательской программы «Фундаментальные основы информационных технологий и компьютерных систем» РАН, гранта 02-01-00284 РФФИ и партнерского проекта 1993Р с МНТЦ, спонсируемого EOARD.

задач. На основе подхода «СИЗ» были разработаны организационные принципы, архитектура и исследовательский прототип системы интеграции знаний.

Процесс поиска знаний инициируется запросом пользователя, описывающим существующую проблему. Принимая во внимание его предпочтения, доступность существующих источников знаний и анализ текущей ситуации, создаётся сеть источников знаний для решения проблемы. Знания извлекаются из источников, интегрируются, проверяются на непротиворечивость и доставляются пользователю. Основа подхода состоит в том, что знание может представлено двумя уровнями. Верхний уровень содержит знания о том, как проблема может быть решена (интенсиональное или описательное знание). Нижний уровень содержит наполнение знаний верхнего уровня, для решения проблемы (эктенсиональное или содержательное знание). Знания на обоих уровнях могут быть интегрированы для получения нового знания (рис. 1).

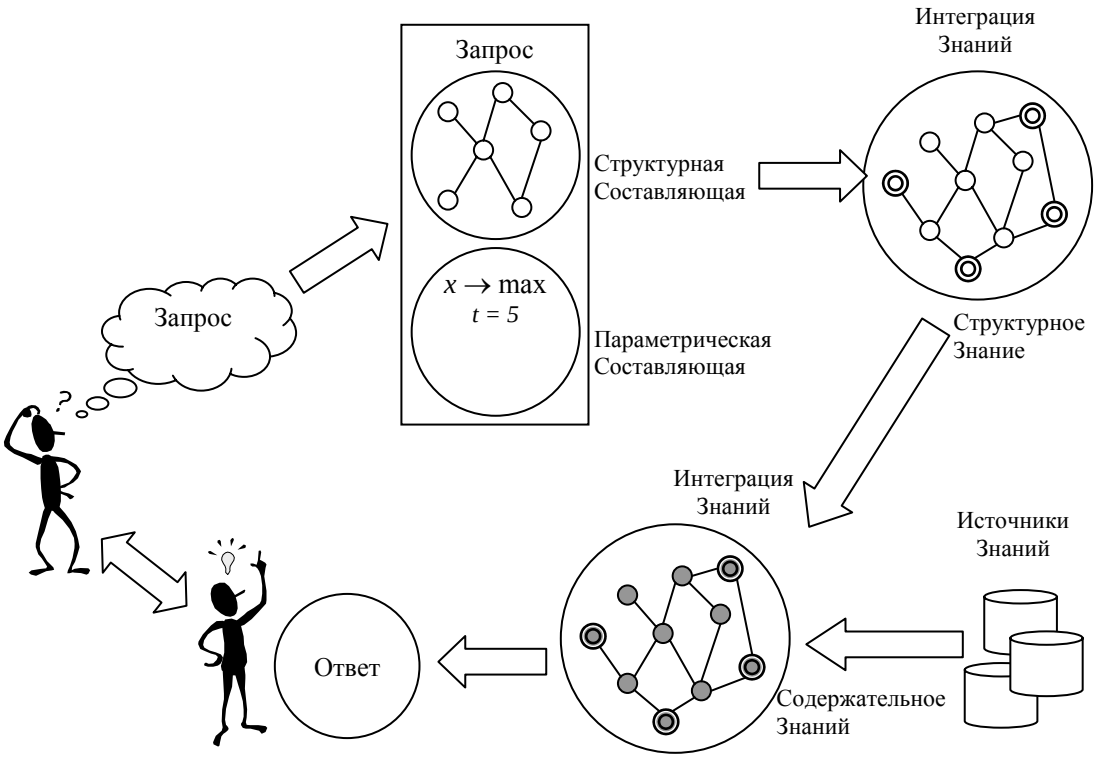

Рисунок 1. Общий сценарий процесса интеграции знаний

Структурное знание представляется при помощи онтологий хранящихся в библиотеке. Нотация объектно-ориентированных сетей ограничений была выбрана для описания онтологий [4]. Согласно этой нотации, онтология A описывается как кортеж $A = (O, Q, D, C)$, где

- O — классы — множество классов сущностей — типовых множеств реальных сущностей, обладающих общими признаками (объекты в нотации сети ограничений). Сущность определяется совокупностью значений свойств класса, которому она принадлежит;
- Q — атрибуты — множество атрибутов классов (свойства в нотации сети ограничений);
- D — домены — множество доменов атрибутов (области допустимых значений свойств в нотации сети ограничений);

Труды СПИИРАН. 2004. Вып. 2. Т. 1. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2004. Issue 2. V. 1. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

- C — множество ограничений, описывающих: принадлежность атрибутов классам; принадлежность доменов атрибутам; совместимость классов (структурные ограничения совместимости классов); иерархические отношения (“быть экземпляром” и “часть-целое”) между классами (иерархические структурные ограничения); ассоциативные отношения между классами (структурные ограничения одного уровня); функциональные ограничения, заданные над значениями атрибутов.

Используемые в описании онтологии компоненты: классы, атрибуты, домены и ограничения, являются элементами онтологии.

Система интеграции знаний работает в терминах общей онтологии-приложения, описывающей проблемную область и хранящейся в библиотеке онтологий. Онтологии-приложения основываются на онтологиях предметных областей и онтологиях задач и методов, также хранящихся в библиотеке онтологий. Каждый пользователь работает в терминах своей расширяемой онтологии запросов, и посредством её, с частью онтологии-приложения, соответствующей интересам данного пользователя/группы пользователей. Перевод между нотациями и терминами источников и системы осуществляется с использованием онтологий источников знаний [5].

Как было указано, одной из важных характеристик поиска знаний является быстрота. В предложенном подходе для обработки запроса конфигурируется вспомогательная сеть источников знаний, которая определяет, когда и какие ИЗ должны быть использованы для обработки запроса наиболее эффективно. Для этого используется картограмма знаний [6–9], включающая информацию о расположении и характеристиках источников знаний.

Статья посвящена вопросам организации, поддержки и использования картограммы знаний. Во втором разделе статьи описаны свойства источников знаний, используемых в системе интеграции знаний и хранящиеся в картограмме. В третьем разделе представлена формальная модель картограммы знаний. В четвёртом разделе рассмотрены проблемы построения картограммы знаний.

Характеристики источников знаний

Подход «СИЗ» разработан для конфигурирования сетей источников знаний, обладающих следующими свойствами: (i) источники знаний расположены в различных местах, (ii) источники знаний в определённый момент времени объединяются в сеть для решения конкретной задачи, (iii) конфигурация сети источников знаний существует только некоторое, предварительно определённое время обработки запроса пользователя.

Были определены следующие источники знаний:

1. Внутренняя база знаний системы интеграции знаний, аккумулирующая генерируемые новые знания при помощи встроенных средств.
2. Эксперты, вводящие знания напрямую по запросу пользователя при помощи встроенных средств.
3. Доступные внешние базы знаний, из которых путём трансляции используемых онтологий между системой интеграции знаний и источником знания, извлекается необходимая информация.
4. Доступные внешние базы данных, из которых извлекается структура таблиц, определяются отношения между ними и области допустимых значений.

Труды СПИИРАН. 2004. Вып. 2. Т. 1. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2004. Issue 2. V. 1. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

5. Структурированные документы (текстовые, офисные, HTML, XML и другие документы), в которых путём поиска информации по ключевым словам определяется соответствие документа онтологии приложения.
6. Прочие источники, для которых разработаны механизмы распознавания и извлечения знаний.

Источники знаний разделяются на две группы (рис. 2): (i) пассивные (базы знаний, базы данных, репозитории и т.д.), которые опосредованно воздействуют на знания, хранящиеся в системе, и (ii) активные, которые непосредственно могут вносить изменения в знания системы (эксперты, специальные средства управления знаниями).

Рисунок 2. Источники знаний

Для описания и быстрого поиска источников знаний в системе интеграции знаний используется картограмма, хранящая следующие характеристики источников знаний [9]:

1. Количественные: (i) стоимость получения знаний из платных источников; (ii) полнота — степень перевода понятий источника знаний в понятия онтологии приложения; (iii) рейтинг, получаемый по итогам экспертной оценки при ранжировании альтернативных источников по заданным критериям.
2. Качественные: (i) расписание работы источника; (ii) наличие требуемых знаний; (iii) дата последнего обновления источника; (iv) рейтинг, вычисляемый по результатам работы системы.

Картограмма знаний содержит информацию о расположении элементов сети источников знаний, используемых для решения задач.

Формальная модель картограммы знаний

В процессе добавления/удаления и модификации источников знаний, система проверяет соответствие между содержимым источников знаний и содержимым онтологии приложения. Набору пар из онтологии приложения $\{(o_i, q_j) \mid o_i \in O, q_j \in Q, \exists c_n \in C^l : c_n = (o_i, q_j)\}$, где o_i — класс, q_j — атрибут, и c_n — отношение принадлежности атрибута классу, ставится в соответствие информация об источниках знаний.

При этом, возможны следующие ситуации:

1. два и более альтернативных источников знаний соответствуют одной паре (o_i, q_j) из онтологии приложения: знание может быть извлечено из любого

Труды СПИИРАН. 2004. Вып. 2. Т. 1. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2004. Issue 2. V. 1. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

из них, в соответствии с ограничениями на параметры исполнения запроса, заданных пользователем (время, стоимость, надёжность и т.д.);

2. два и более источника знаний содержат различное знание, относящиеся к паре (o_i, q_j) из онтологии приложения: напр., необходимо получить знание из нескольких источников и обработать его предварительно определёнными (разработанными) процедурами (конъюнкция, дизъюнкция, композиция, и т.д.) для того, чтобы получить значения соответствующих атрибутов ([10–11]). В этом случае, в соответствии с информацией, хранимой в картограмме знаний, из онтологии задач и методов выбираются и вызываются необходимые процедуры для выполнения операции интеграции знаний.

При определении соответствий, описываемых картограммой знаний, определяются области допустимых значений атрибутов, которые можно получить из источников. Допустимые значения могут быть дискретными и интервальными. Обобщённая модель картограммы знаний представлена на рис. 3.

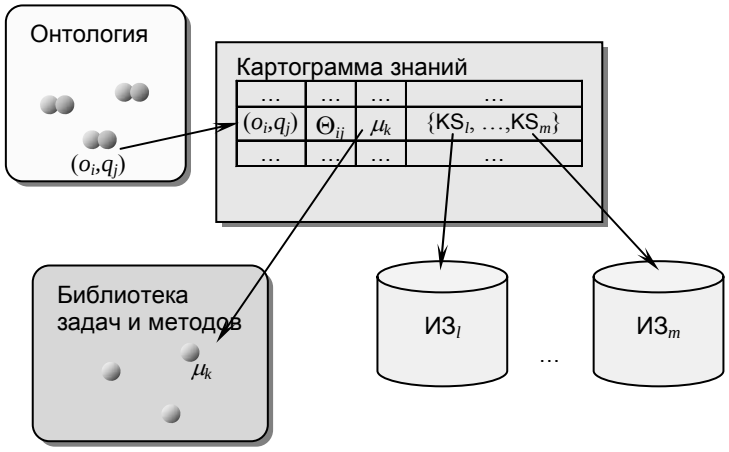

Рисунок 3. Обобщённая схема картограммы знаний

В момент времени t картограмма знаний определяется как множество элементов $KMap_t = \{KMapEl_{lt}\}_{l=1}^{N_{KMap}}$, где $N_{KMap} \in N$. Элемент $KMapEl_t$ является кортежем: $\left\{\left\langle (o_i, q_j), \{KSID_p\}_{p=1}^{m}, \mu_k, \Theta_{ij} \right\rangle\right\}$, где: $\mu_k \in$ онтологии задач и методов; $m \in N$ количество источников знаний, которые должны быть обработаны для получения значения атрибута q_j входящего в связку (o_i, q_j); и Θ_{ij} область возможных значений атрибута, которые могут быть получены из источников знаний $\{KSID_p\}_{p=1}^{m}$. Если μ_k пусто, то m равно 1: $m = 1$. Если рассматривать время t как дискретную величину в последовательности $t_1,...,t_n$, где t_1 время подключения первого источника знаний, t_n время последнего обновления картограммы знаний, то во время сдвига от t_i к t_{i+1} происходит когда (i) новый источник знаний добавляется или удаляется, (ii) содержимое источников знаний изменяется.

Примеры. Пусть дана онтология A, $o_1 \in O$, $q_1 \in Q$, $d \in D, d = R$ (R множество действительных чисел). $KS_1,\ldots, KS_n$ — множество найденных источников знаний. Обозначим $KMap_1$ фрагмент картограммы знаний $KMap$ содержащий только пару (o_1, q_1).

Труды СПИИРАН. 2004. Вып. 2. Т. 1. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2004. Issue 2. V. 1. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

Пример 1. Из источника KS_1 и KS_2 можно получить один и тот же набор значений атрибута q_1, связанного с o_1,: $\{1,2,5\}$ и $\{one, two, five\}$ соответственно. В этом случае, источники знаний KS_1 и KS_2 являются альтернативными относительно пары (o_1, q_1), а фрагмент картограммы имеет вид $KMap_1 = \{\langle (o_1, q_1), KS_1, \varnothing, \{1,2,5\}\rangle, \langle (o_1, q_1), KS_2, ASCII2Num, \{1,2,5\}\rangle\}$, где *ASCII2Num* метод, хранящийся в онтологии задач и методов. В процессе конфигурирования сети, источники знаний будут выбраны в соответствии с заданными пользователем оптимизационными критериями (напр., вызов дополнительного метода может потребовать дополнительного времени обработки запроса).

Пример 2. Из источника KS_1 и KS_2 можно получить один и тот же набор значений атрибута q_1, связанного с o_1,: $\{1,2,5\}$ и $\{1,2,7\}$ соответственно. При подключении ИЗ эксперт принял решение, что для получения полного набора значений необходимо обработать оба источника, а результат объединить. В этом случае фрагмент картограммы имеет вид $KMap_1 = \{\langle (o_1, q_1), \{KS_1, KS_2\}, KSUnion, \{1,2,5,7\}\rangle\}$, где *KSUnion* метод, хранящийся в онтологии задач и методов.

Пример 3. Из источника KS_1 и KS_2 можно получить один и тот же набор значений атрибута q_1, связанного с o_1,: $[0:20]$ и $[11.5:25]$ соответственно. При подключении ИЗ эксперт принял решение, что для получения полного набора значений необходимо обработать оба источника, а результат будет находиться в пересечении полученных значений. В этом случае фрагмент картограммы имеет вид $KMap_1 = \{\langle (o_1, q_1), \{KS_1, KS_2\}, KSInter, [11.5:20]\rangle\}$, где *KSInter* метод, хранящийся в онтологии задач и методов.

Построение картограммы знаний

Построению картограммы знаний предшествуют следующие вспомогательные процедуры:

- Построение онтологии приложения. На этапе построения картограммы знаний она используется для выборки пар классов и атрибутов.
- Описание источников знаний в системе (расположение, тип и т.д.).
- Разработка функций приведение слов к каноническому виду (напр., обрезание окончаний и суффиксов, приведение множественного числа к единственному и т.п.) с учётом языка представления слова.
- Подготовка средств поиска незначимых строк в тексте с учётом языка представления текста.
- Разработка интерфейсных модулей извлечению содержимого источников для индексирования относительно элементов онтологии. В рамках исследовательского прототипа эти функции выполняли следующие задачи:
  - Конвертирование содержимого документов в текстовый формат с учётом кодировок хранения информации.
  - Приведение слов к каноническому виду.
  - Изъятие незначимых слов из текста.
  - Сохранение результатов работы для последующей обработки (напр., в текстовом виде).
- Разработка процедур сравнения строк для выявления похожих фрагментов текстов.

Труды СПИИРАН. 2004. Вып. 2. Т. 1. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2004. Issue 2. V. 1. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

После того, как все указанные выше процедуры выполнены, картограмма знаний может быть построена. Возможны два вида формирования картограммы: (i) «обновление на лету», когда специальные средства системы (напр., агенты) производят периодическую проверку источников знаний (доступность, изменение содержимого или структуры) и вносят необходимые изменения в картограмму знаний и (ii) «построение по запросу», когда по событию выполняется процедура, удаляющая содержимое картограммы целиком и строящая её заново. Второй способ требует большего времени на построение картограммы, но прост в реализации.

Обработке запроса пользователя производится в соответствии со следующей последовательностью:

1. Разбивка запроса на отдельные слова, построение списка значимых слов.
2. Выявление незначимых слов и чисел, их изъятие из списка значимых слов.
3. Проверка значимых слов на наличие синтаксических ошибок. В случае обнаружения синтаксических ошибок, пользователю предлагается их исправить или продолжить обработку запроса игнорируя их. Слова, содержащие синтаксические ошибки, удаляются из списка значимых слов.
4. Поиск регулярных выражений (напр., для поиска в запросе единиц измерений атрибутов) и редуцирование списка значимых слов.
5. Поиск ограничений на значения атрибутов (напр., выражения «*РАВНО*», «*<БОЛЬШЕ*» и т.д.) и редуцирование списка значимых слов и редуцирование списка значимых слов.
6. Поиск критериев для решения оптимизационных задач (напр., выражение «*МИНИМИЗИРОВАТЬ*») и редуцирование списка значимых слов.
7. Приведение значимых слов к каноническому виду.
8. Поиск в онтологии приложения элементов (классов и атрибутов), имена которых похожи на значимые слова. При этом сравниваются только канонические формы слов.
9. Оценка соответствия найденных элементов онтологии запросу пользователя и отсечение тех элементов, степень сходства которых меньше порогового значения.
10. Редуцирования списка найденных элементов онтологии по определённым правилам: (напр., если было найдено несколько классов принадлежащих одной ветке таксономии, оставляется только объектный класс — который может иметь сущности — и все атрибуты надклассов). Таким образом, формируется структурная составляющая запроса пользователя.
11. Формирование параметрической составляющей запроса пользователя. Формирование списка ограничения на значения найденных атрибутов и соотнесение критериев оптимизации к атрибутам.
12. Поиск по найденным элементам онтологии информации в картограмме знаний с учётом ограничений на значения атрибутов. В результате поиска формируется список источников, из которых знания будут извлекаться и определяются элементы онтологии задач и методов, которые будут использоваться при интеграции.
13. Извлечение знаний из источников их интеграция и визуализация для представления пользователю.

Эксперименты показали, что большая часть времени уходит на шаги 8, 11, 12 и 13. Поэтому задача правильной организации картограммы знаний сыграло значимую роль в процессе ускорения обработки запроса пользователя.

Труды СПИИРАН. 2004. Вып. 2. Т. 1. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2004. Issue 2. V. 1. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

Заключение

В данной статье предложена модель картограммы знаний, разработанная в рамках подхода «СИЗ». Предложена формальная модель картограммы, основанная на использованном формализме описания знаний при помощи онтологий и представления онтологий в нотации объектно-ориентированных сетей ограничений. Консолидация информации о связях между описанием проблемной области, содержимым источников знаний и ссылках на задачи и методы позволяют решить задачу индексирования знания и дают возможность в будущем реализовать быстрый поиск знаний для их интеграции.

Дальнейшая работа в данной области будет ориентирована на реализацию данной модели для реального практического приложения, разработке методов онлайнового мониторинга содержимого картограммы и оценке её эффективности при обработке запросов пользователей.

Литература

[1] *Смирнов А. В.*, *Левашова Т. В.*, *Пашкин М. П.*, *Шилов Н. Г.* Онтолого–ориентированный многоагентный подход к построению систем интеграции знаний из распределённых источников // Информационные технологии и вычислительные системы. 2002. № 1. С. 62–82.

[2] *Смирнов А. В.*, *Пашкин М. П.*, *Шилов Н. Г.*, *Левашова Т. В.* Подход к конфигурированию сети источников знаний для логистики знаний // Известия ТРТУ. Тематический выпуск «Интеллектуальные САПР»: Материалы Международной научно-технической конференции «Интеллектуальные САПР». Таганрог: ТРТУ, 2003. № 3 (31). С. 28–32.

[3] *Smirnov A.*, *Pashkin M.*, *Chilov N.*, *Levashova T.* KSNet-Approach to Knowledge Fusion from Distributed Sources // Computing and Informatics. 2003. V. 22. P. 105–142.

[4] *Smirnov A.*, *Pashkin M.*, *Chilov N.*, *Levashova T.*, *Krizhanovsky A.* In: R. Meersman, Z. Tari, D.C. Schmidt et al. (Eds.): Ontology-Driven Knowledge Logistics Approach as Constraint Satisfaction Problem. On the Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE. Proceedings of the International Conference on Ontologies, Databases and Applications of Semantics Information Reuse and Integration (ODBASE'2003). November 3-7, 2003. Catania, Sicily, Italy. Springer. *Lecture Notes in Computer Science* 2888. P. 535–652.

[5] *Смирнов А. В.*, *Пашкин М. П.*, *Шилов Н. Г.*, *Левашова Т. В.* Управление онтологиями // Известия РАН. Теория и системы управления. 2003. Ч. 1, № 4. С. 132–146; Ч. 2. № 5. с. 89–101.

[6] *Vail E.F.* Knowledge Mapping: Getting Started with Knowledge Management // Information Systems Management. Fall, 1999. P. 16–23.

[7] *Park J.Y.*, *Gennari J.H.*, *Musen M.A.* Mappings for Reuse in Knowledge-based Systems: SMI Technical Report 97-0697, 1997.

[8] *Gordon J.L.* Creating Knowledge Maps by Exploiting Dependent Relationships // Knowledge-Based Systems. 2000. V. 13. P. 71–79.

[9] *Пашкин М.П.* Ранжирование альтернативных источников знаний на основе технологии групповой поддержки принятия решений // Труды СПИИРАН / Под ред. Р.М. Юсупова, СПб.: СПИИРАН, 2002. Вып. 1, т. 3. С. 22–30.

[10] *Oussalah M.* Study of Some Algebraical Properties of Adaptive Combination Rules // Fuzzy Sets and Systems. 2000. V. 114. P. 391–409.

[11] *Hunter A.* Merging Potentially Inconsistent Items of Structured Text // Data & Knowledge Engineering. 2000. P. 305–332.