УДК 004.8: 004.9

П. А. Ломов, А. В. Маслобоев

ТЕХНОЛОГИЯ ФОРМИРОВАНИЯ ВАРИАНТОВ РАСШИРЕНИЯ ПОИСКОВЫХ ЗАПРОСОВ НА ОСНОВЕ ОБЩЕСИСТЕМНОГО ТЕЗАУРУСА

Рассматривается проблема использования общесистемного расширяемого тезауруса как основы высокоуровневого пользовательского интерфейса информационной системы распределенного семантического поиска. Представлена технология формирования поискового запроса на основе дополнения его наиболее существенными, а также близкими по смыслу понятиями и свойствами.

Ключевые слова: *информационная технология, онтология, тезаурус, семантическая интеграция информации, формирование запроса.*

Введение. Важным аспектом создания информационных систем распределенного семантического поиска является разработка высокоуровневого интерфейса, обеспечивающего эффективный диалог пользователя с системой. Основная задача интерфейса заключается в обеспечении поддержки пользователя в процессе формирования и исполнения поисковых запросов. Это предполагает оперирование знакомым пользователю терминологическим аппаратом, а также представление в интерактивном режиме понятий и их атрибутов, которые целесообразно использовать в поисковом запросе. Результативность решения данной задачи требует в первую очередь достаточно тесной интеграции интерфейса с понятийной системой предметной области.

На сегодняшний день существуют несколько подходов, направленных на создание высокоуровневого пользовательского интерфейса для реляционных баз данных [1, 2]. Среди их недостатков выделяют: отсутствие правил логического вывода, позволяющих получить новое знание, вытекающее из хранимых фактов; невозможность использования метаинформации и запросов, дающих приближенный результат; слабая адаптация системы под конкретного пользователя на основе истории взаимодействия с ним и др.

Среди современных концепций можно также выделить подходы, использующие в своей основе динамическую фасетную классификацию понятий [3, 4]. Это позволяет использовать простые правила уточнения запроса на каждом шаге его формирования, но так как сама фасетная классификация достаточно слабо отражает разнообразие связей между объектами предметной области, то возможности такого уточнения достаточно ограничены.

Наиболее перспективными являются подходы, направленные на создание пользовательского интерфейса на основе формальной онтологической модели, содержащей понятия, в которых отображаются элементы данных, хранящихся в различных гетерогенных источниках [5]. В этом случае онтология используется как система связанных между собой понятий предметной области, в которой, как правило, пользователь ориентируется. Это позволяет пользователю оперировать известными ему терминами, что облегчает навигацию по понятийной системе и упрощает формирование сложных запросов, включающих несколько объектов поиска. Наряду с этим пользователь имеет возможность ознакомиться с предметной областью без совершения запросов для более ясного понимания того или иного термина, определяемого его иерархическим положением и взаимосвязями с другими терминами. Однако такие подходы часто опираются на единую онтологию, задающую общую семантику понятий, описываемых в различных информационных источниках, что соответствует централизованному подходу к интеграции информации, имеющему некоторые недостатки [6]. Для их устранения было предложено использовать разновидность гибридного подхода к семантической интеграции информации с использованием расширяемого общесистемного тезауруса. Однако тезаурус

обладает гораздо меньшей выразительностью по сравнению с единой онтологией, что снижает его возможности в отношении поддержки пользователя при поиске информации.

Решению данной проблемы и посвящена настоящая статья, в которой рассматриваются вопросы, связанные с использованием расширяемого тезауруса как основы высокоуровневого интерфейса информационной системы распределенного семантического поиска. Рассматривается возможность включения элементов онтологии DOLCE и метасвойств в расширяемый тезаурус, что позволит повысить уровень его выразительности. Описывается процесс формирования запроса на основе правил определения возможных вариантов его построения.

Расширения тезауруса с использованием элементов онтологии DOLCE и метасвойств OntoClean. В общем случае тезаурус можно определить как словарь терминов, связанных определенным набором семантических отношений (синонимия, гипонимия, родовая связь). Однако в данной статье понятие тезауруса не в полной мере соответствует общепринятому: тезаурус позиционируется как каноническая модель для осуществления семантической интеграции информационных ресурсов на основе их онтологических моделей [7].

Для формального определения тезауруса введем следующие обозначения: C — множество всех концептов (классов), содержащихся в онтологиях исходных интегрируемых ресурсов; A — множество всех свойств понятий онтологий. Заметим, что в онтологиях, как правило, встречаются также экземпляры — нижнеуровневые компоненты, представляющие конкретные объекты реального мира. Экземпляры распределяются по классам (концептам) онтологии: например, „314“ будет являться экземпляром класса „Число“. Так как обычно понятия предметной области выражаются в виде концептов, а не экземпляров, то последние в тезаурус не включаются.

Каждый концепт из множества C представляется соответствующим ему элементом тезауруса типа „Объект“, обозначаемым как $O_i \in O$, где O — множество всех объектов тезауруса: $O_i = <N_i, A_i>$, где N_i — описание объекта, соответствующее какому-либо описанию представляемого им концепта, которым может быть название или определение на естественном языке; A_i — множество онтологий, в которых представлен концепт, соответствующий объекту O_i. Каждый атрибут из множества A_i представлен в тезаурусе соответствующим элементом типа „Свойство“, обозначаемым как $P_i \in P$, где P — множество всех свойств тезауруса: $P_i = <N_i, A_i>$, где N_i — символьное имя свойства или описание P_i, соответствующее наименованию или описанию атрибута из онтологии; A_i — множество онтологий, в которых представлен атрибут, соответствующий свойству P_i.

На множествах элементов тезауруса заданы различные отношения, среди которых основными являются: $\mathrm{HP} \subseteq (O \times O)$ — отношение гипонимии между объектами, $\mathrm{PR} \subseteq O \times P$ — отношение принадлежности свойства объекту тезауруса, $\mathrm{VL} \subseteq (P \times V) \cup (P \times O)$ — отношение принадлежности свойству некоторого значения — литерала V или объекта O, $\mathrm{SYN} \subseteq (O \times O)$ — отношение синонимии, $\mathrm{ASC} \subseteq (O \times O)$ — отношение ассоциации.

Одной из основополагающих задач тезауруса — как интеграционной модели — является определение отношений между понятиями различных онтологий, однако для этого часто необходима дополнительная семантика, которую нельзя отразить, используя существующий набор элементов тезауруса. Также малое количество основных видов связей между элементами снижает выразительные возможности тезауруса, необходимые для интерактивного формирования запросов.

Данную проблему можно решить путем определения некоего общего основания, способного отразить фундаментальные смысловые особенности понятий любой предметной области или задачи. Способом реализации такого подхода является использование онтологий верхнего уровня, включающих абстрактные понятия и отношения, которые могут быть детализированы в онтологиях, представляющих более узкие предметные области. Поэтому

целесообразно рассматривать онтологии верхнего уровня как некое общее основание, полагаясь на которое, можно выполнять анализ предметной области или задачи для выделения значимых понятий и отношений, а также осуществлять последующий синтез понятийной системы. Включение элементов таких онтологий в расширяемый тезаурус позволит сохранить исходную семантику терминов при их добавлении в тезаурус, с большей точностью определить и обосновать смысловую близость понятий или ее отсутствие, формально определить базовые понятия предметной области и их основные атрибуты, а также упростить разработку и обеспечить правильность создания онтологических спецификаций исходных информационных ресурсов.

В качестве онтологии верхнего уровня (или метаонтологии) авторами настоящей статьи предлагается использовать онтологию базовых категорий естественного языка и здравого смысла (Descriptive Ontology for Linguistic and Cognitive Engineering — DOLCE) [8]. Выбор именно этой онтологии объясняется ее ориентацией на различные социальные субъекты, объекты и процессы, такие как организации, коллективы, планы и нормы.

Для отображения основных понятий онтологии DOLCE необходимо определить их принадлежность к типам объектов тезауруса, учитывая исходную семантику для обеспечения непротиворечивости при отображении. Необходимо также отметить следующее: несмотря на то, что данная онтология верхнего уровня создана в результате исследования метасвойств, общих для понятий различных предметных областей, сами метасвойства непосредственно в ней не представлены. С этой целью, наряду с концептами DOLCE, можно включить в тезаурус новые элементы, соответствующие метасвойствам, определенным в проекте OntoClean [9], и впоследствии использовать их для аннотирования понятий или отношений между ними.

Рассмотрим основные концепты DOLCE — абстрактный (abstract), длящийся (endurant) и постоянный (perdurant) объекты, которые являются подклассами концепта „Частность“ (patricular), обозначающего любую сущность. Исходя из этого их можно представить как объекты тезауруса — прямые гипонимы[1] объекта „Сущность“ — o, являющегося, в свою очередь, гипернимом[2] всех объектов тезауруса:

$$\text{prd}=<\text{"Perdurant"}, A_{\text{prd}}>, \text{prdHP}o,$$

$$\text{abs}=<\text{"Abstract"}, A_{\text{abs}}>, \text{absHP}o,$$

$$\text{end}=<\text{"Endurant"}, A_{\text{end}}>, \text{endHP}o.$$

Для определения метасвойств предлагается включить в тезаурус функции, отображающие множества его элементов на некоторые множества значений. Для этого определим следующие функции, соответствующие метасвойствам OntoClean.

— $\text{RG}: O \cup C \to \{\text{"rigid"},\text{"non-rigid"},\text{"antirigid"}\}$ — функция, соответствующая свойству „Стойкость“ (Rigid); область ее значений составляют литералы, указывающие на свойственные этому понятию стойкость — „*rigid*“, антистойкость (изменчивость) — „*antirigid*“ и отсутствие стойкости — „*non-rigid*“. Данная функция применима к понятиям в онтологии и тезаурусе, а не к свойствам, так как в OntoClean наличие стойкости означает неизменность значимого свойства (essential property) у экземпляров определенного класса (концепта) в онтологии, однако само таковое свойство, как правило, непосредственно не указывается. Таким образом, функция „Стойкость“ обозначает неизменность членства экземпляров в данном классе. Исходя из этого значение „antirigid“ можно использовать для определения классов, соответствующих понятиям-ролям, например „Несовершеннолетний“, экземпляры которых со временем покинут такие классы.

[1] Гипоним — понятие, в отношении к другому понятию выражающее подвид, более конкретное понятие: например, понятие „Оптика“ — гипоним понятия „Наука“.

[2] Гиперним (гипероним) — понятие, в отношении к другому понятию выражающее более общую сущность: например, понятие „Животное“ — гипероним понятия „Собака“.

— $\mathrm{UN}: C \cup O \to \{\text{"un","non-unity"}\}$ — функция, определяющая свойство „Единство“ (Unity); ее значениями являются литералы „*un*“ и „*non-unity*“, обозначающие соответственно наличие и отсутствие данного метасвойства у понятия. Аналогично функции „Стойкость“ данная функция применяется к понятиям онтологии, так как атрибут понятия, через который оно проявляется, как правило, явно не указывается.

— $\mathrm{ID}: C \cup O \cup A \cup P \to \{\text{"id","non-id"}\}$ — функция, определяющая свойство „Идентифицируемость“ (Identity); ее значениями являются литералы „*id*“ и „*non-id*“, обозначающие соответственно наличие и отсутствие возможности идентификации экземпляра некоторого класса по некоторому свойству.

— $\mathrm{DP}: C \cup O \cup A \cup P \to \{\text{"dep","non-dep"}\}$ — функция, определяющая свойство „Зависимость“ (Dependence); ее значениями являются литералы „*dep*“ и „*non-dep*“, обозначающие соответственно наличие и отсутствие зависимости существования одной сущности от существования другой. Например, справка о регистрации гражданина зависит от наличия ордера на жилое помещение или иного документа, на основании которого она может быть выдана.

Наличие метасвойств идентифицируемости и/или зависимости у понятий может определяться по принадлежащим им свойствам, через которые проявляется идентифицируемость или зависимость. Для указания данных факторов введем следующие отношения: $\mathrm{IDF} \subseteq (P \times O)$ — отношение между объектом и идентифицирующим его свойством; $\mathrm{DPF} \subseteq (P \times O)$ — отношение между объектом и характеризующим его свойством, через которое проявляется зависимость от другого объекта; $\mathrm{ESF} \subseteq (P \times O)$ — отношение между объектом и его существенным свойством; $\mathrm{PRC} \subseteq (O \times O)$ — отношение партисипации (participation): участие объекта в процессе или событии; $\mathrm{PRT} \subseteq (O \times O)$ — отношение партономии (partOf) между частью и целым.

Анализ остальных концептов онтологии DOLCE и их использование для расширения тезауруса здесь не рассматриваются. Подробно данный вопрос изложен в работе [10].

Применение тезауруса для формулировки запроса. Основной задачей формулировки пользовательского запроса является как можно более полное и точное определение объекта поиска с помощью элементов тезауруса. Процедура формулировки запроса начинается с определения объекта поиска путем его выбора из множества объектов тезауруса. При расширении запроса пользователю будут представлены как объекты, непосредственно связанные с начальным, так и имеющие с ним опосредованные отношения. Это позволит отобразить различные контексты, а также наиболее значимые условия проведения поиска.

Технология расширения запроса основывается на следующих основных правилах.

Правило транзитивной идентификации. При наличии у некоторого объекта t идентифицирующего свойства (IDF) или свойства зависимости (DPF) — p^t, значением которого является другой объект — b, также имеющий идентифицирующее свойство или свойство зависимости — p^b, между свойством p^b и объектом t формируется динамическое отношение принадлежности свойства объекту. Таким образом, пользователь имеет возможность сразу задать ограничение на свойство p^b без перехода к обзору свойств объекта b. В случае же наличия более длинной цепочки объектов, идентифицирующих друг друга, все их идентифицирующие свойства также представляются пользователю. Формально правило транзитивной идентификации имеет следующий вид:

$$(t\mathrm{PR}p^t) \wedge (p^t \mathrm{VL} b) \wedge (b\mathrm{PR}p^b) \wedge (p^t \mathrm{IDF}t \vee p^t \mathrm{DPF}t) \wedge (p^b \mathrm{IDF}b \vee p^b \mathrm{DPF}b) \to t\mathrm{PR}p^b.$$

Правило транзитивности свойств синонима. При наличии у объекта t свойства p_i^t, значением которого является объект b, имеющий, в свою очередь, синоним — объект f, между t и свойствами f — p_m^f, обладающими метасвойствами, устанавливаются динамические

отношения принадлежности. Формально правило транзитивности свойств синонима можно записать в следующем виде:

$$(t\mathrm{PR}p_i^t) \wedge (p_i^t \mathrm{VL}b) \wedge (b\mathrm{SYN}f) \wedge (f\mathrm{PR}p_m^f) \wedge (p_m^f \mathrm{IDF}f \vee p_m^f \mathrm{DPF}f \vee p_m^f \mathrm{ESF}f) \to t\mathrm{PR}p_m^f .$$

Правило партисипативной/партономической идентификации. Данное правило может быть использовано в следующих случаях:

1) объект t, имеющий метасвойство единства, может быть идентифицирован посредством указания значений идентифицирующих свойств p_i^b его объектов-частей (PRT) или объектов-участников (PRC) — b:

$$(\mathrm{UN}(t) = \text{"un"}) \wedge ((t\mathrm{PRT}b) \vee (t\mathrm{PRC}b)) \wedge (b\mathrm{PR}p_i^b) \wedge (p_i^b \mathrm{IDF}b) \wedge \to t\mathrm{PR}p_i^b ;$$

2) объект-часть или объект-участник t может быть идентифицирован через объект b, имеющий метасвойство единства, частью которого он является или в котором участвует, посредством указания значений идентифицирующих свойств p_i^f других объектов-частей (PRT) или объектов-участников (PRC) — f:

$$(\mathrm{UN}(b) = \text{"un"}) \wedge ((b\mathrm{PRT}t) \vee (b\mathrm{PRC}t)) \wedge ((b\mathrm{PRT}f) \vee (b\mathrm{PRC}f)) \wedge (f\mathrm{PR}p_i^f) \wedge (p_i^f \mathrm{IDF}f) \to t\mathrm{PR}p_i^f .$$

Перечисленные правила приводят к неявному расширению запроса путем использования метаинформации для представления наиболее важных свойств объекта поиска, а также вариантов запроса по связанным с ним объектам.

Заключение. В ходе проведенных исследований по проблеме использования расширяемого тезауруса в качестве основы высокоуровневого интерфейса информационной системы распределенного семантического поиска получены следующие результаты:

— рассмотрено дополнение расширяемого тезауруса элементами онтологии DOLCE и метасвойствами, что позволит повысить выразительные возможности тезауруса и использовать его в качестве основы высокоуровневого пользовательского интерфейса;

— разработана технология формирования поискового запроса на основе правил определения возможных вариантов его расширения, включающих наиболее важные условия поиска, а также близкие по смыслу понятия и свойства.

Полученные результаты могут найти применение при решении практических задач, связанных с семантической интеграцией информации, построением онтологических моделей предметных областей, семантическим поиском и разработкой пользовательского интерфейса.

Статья подготовлена по результатам работ, выполненных при поддержке Российского фонда фундаментальных исследований, проект № 08-07-00301-а.

СПИСОК ЛИТЕРАТУРЫ

1. *Benzi F., Maio D., Rizzi S.* VISIONARY: a viewpoint-based visual language for query in relational databases // J. of Visual Languages and Computing. 1999. Vol. 10 (2). P. 117—145.

2. *Catarci T., Francesca M., Levialdi S., Batini C.* Visual query systems for databases: A survey // J. of Visual Languages and Computing. 1997. Vol. 8 (2). P. 215—260.

3. *Обухова О. Л., Бирюкова Т. К., Гершкович М. М., Соловьев И. В., Чочиа А. П.* Метод динамического создания связей между информационными объектами базы знаний // Тр. XI Всерос. науч. конф. „Электронные библиотеки: перспективные методы и технологии, электронные коллекции“. Петрозаводск: КарНЦ РАН, 2009. С. 39—45.

4. *Песков Д. Н., Каберник В. В., Михеев А. Н., Афонцев С. А.* Динамическая фасетная классификация (MGI-классификация) и ее применение к задачам управления знаниями в вузе [Электронный ресурс]: <http://www.mgimonics.com/index.php?option=com_content&task=view>.

5. *Catarci T., Dongilli P., DiMascio T., Franconi E., Santucci G., Tessaris S.* An ontology-based visual tool for query formulation support // Proc. of the 16th European Conf. on Articial Intelligence. 2004.

6. *Ломов П. А. Шишаев М. Г.* Интеграция данных на основе онтологий для обеспечения информационной поддержки управленческих решений // Тр. Ин-та системного анализа РАН. 2008. Т. 39. С. 159—173.

7. *Ломов П. А., Шишаев М. Г.* Разработка метода семантической интеграции информации в сфере государственного и муниципального управления // Тр. XI Всерос. науч. конф. „Электронные библиотеки: перспективные методы и технологии, электронные коллекции“. Петрозаводск: КарНЦ РАН, 2009. С. 78 — 86.

8. *Masolo C., Borgo S., Gangemi A., Guarino N., Oltramari A., Schneider L.* DOLCE: A Descriptive Ontology for Linguistic and Cognitive Engineering // DOLCE Documentation [Электронный ресурс]: <http://www.loa-cnr.it/DOLCE.html>.

9. *Guarino N., Welty C.* An overview of OntoClean // Handbook on Ontologies; Eds.: *S. Staab, R. Studer*. Berlin: Springer, 2004. P. 151—172.

10. *Ломов П. А., Шишаев М. Г., Диковицкий В. В.* Онтологическая модель государственного и муниципального управления для проведения семантической интеграции информации в области государственного и муниципального управления // Материалы VIII Всерос. школы-семинара „Прикладные проблемы управления макросистемами“, 29 марта — 2 апр. 2010 г., Апатиты: Тр. Ин-та системного анализа РАН. 2010. Т. 59. С. 118—132.

Сведения об авторах

Павел Андреевич Ломов — аспирант; Институт информатики и математического моделирования технологических процессов Кольского научного центра РАН, Апатиты; E-mail: lomov@iimm.kolasc.net.ru

Андрей Владимирович Маслобоев — канд. техн. наук, доцент; Институт информатики и математического моделирования технологических процессов Кольского научного центра РАН, Апатиты; E-mail: masloboev@iimm.kolasc.net.ru

Рекомендована Институтом

Поступила в редакцию
07.02.12 г.