ОНТОЛОГИЯ КАК СРЕДСТВО УПРАВЛЕНИЯ ЗНАНИЯМИ В УСЛОВИЯХ МНОГОЯЗЫЧНОСТИ

Д. Ю. Постаногов

ВВЕДЕНИЕ

Деятельность современных организаций обуславливается постоянно растущими объемами перерабатываемой информации. Как правило, документооборот при этом организован посредством использования статических общедоступных или локальных систем хранения документов, в следствие чего современные предприятия сталкиваются со следующими проблемами: затрудненный доступ к знаниям, дублирование информации, наличие противоречий в различных источниках, возможность потери важной информации.

Одной из наиболее развитых современных концепций, направленных на преодоление подобных затруднений, является концепция управления знаниями (Knowledge Management), согласно которой предприятия сознательно организуют и анализируют знания сотрудников и рабочих групп, а также предоставляют им доступ к этим знаниям. В свете данной концепции перед современными организациями стоят следующие наиболее важные задачи управления знаниями: организация информационных потоков; извлечение знаний из различных информационных источников; представление знаний в унифицированном виде; обеспечение эффективного доступа к релевантным источникам знаний; реферирование документов с целью ускорения анализа источников знаний; тематическая группировка документов и других источников знаний; визуализация извлеченных знаний с возможностью навигации.

Более того, в силу все большего развития и распространения транснациональных компаний и необходимости поддержки координированных взаимодействий сотрудников и рабочих групп, имеющих различные языки коммуникации, многоязычность является актуальным условием при решении перечисленных задач управления знаниями.

АКТУАЛЬНОСТЬ ВОПРОСА НАПОЛНЕНИЯ БАЗ ЗНАНИЙ

Реализация системы управления знаниями затрагивает ряд принципиальных вопросов, среди которых основными являются: (1) выбор модели представления знаний в системе; (2) организация эффективного наполнения базы знаний требуемого объема; (3) разработка эффективного ин-

струментария, оперирующего базой знаний системы для решения задач управления знаниями.

С конца 1990-х годов World Wide Web Consortium (W3C) при участии большого числа исследователей развивает идею использования семантических сетей в виде стандартных средств работы с метаданными, оформляемых в концепции Semantic Web [1]: Resource Description Framework (RDF) и Web Ontology Language (OWL). Ключевым моментом концепции является обязательное сопровождение различных информационных ресурсов метаданными, представленными в соответствии с предложенными спецификациями, с целью упрощения их машинной интерпретации и переработки. За счет аннотирования ресурсов и проектирования метаданных на общие или локальные онтологии предполагается решение проблемы неструктурированности и смысловой неоднозначности текстов, выраженных на естественном языке (ЕЯ). При этом разрабатывается методология, алгоритмы и программное обеспечение, реализующее инструментарий для решения различных задач управления знаниями в сети и на предприятиях, входом которого является именно стандартизированная аннотация информационных ресурсов, составляющая в конечном итоге базу формализованных знаний.

Таким образом, вопросы представления знаний (1) и программного инструментария для работы с ними (3) в настоящий момент в должной мере прорабатываются исследователями в сфере управления знаний. Однако, вопрос наполнения баз знаний достаточного объема (2) остается открытым, т.к. согласно концепции Semantic Web решение задачи предоставления входной информации в аннотированном для машинной переработки виде возлагается на самих поставщиков информационных ресурсов, а следовательно актуальными остаются следующие проблемы:

- высокие издержки либо невозможность занесения всех информационных ресурсов больших предприятий в базу знаний;
- несогласованность содержания баз знаний на различных уровнях;
- необходимость обновления метаданных в соответствии с изменениями содержания информационных ресурсов;
- необходимость ручного перевода базы знаний на другие языки в условиях многоязычности.

ОНТОЛОГИЧЕСКИЙ ПРОЦЕССОР GOLDFIRE

В данной работе в качестве средства эффективного решения поставленных задач рассматривается технология автоматической переработки текста, в частности, онтологический процессор *Goldfire* [2], по сути обеспечивающий автоматическое извлечение знаний из текстовых доку-

ментов и их представление в виде онтологий предметных областей [3]. При этом многоуровневая лексическая, морфологическая, синтаксическая, семантическая и онтологическая обработка текстов осуществляется следующими этапами работы преформатора, лингвистического и онтологического процессоров:

- конвертирование документов (извлечение текстовых данных из документов различных форматов) и преформатирование текста;
- распознавание границ слов и предложений;
- определение частей речи и других грамматических значений слов;
- построение дерева синтаксического разбора предложения;
- выделение синтаксических отношений в предложении;
- распознавание и канонизация смысловых единиц, их структуры и свойств в контексте предложения;
- снятие семантической неоднозначности слов и фраз с определением места в иерархическом родовидовом тезаурусе онтологии;
- распознавание анафорических связей;
- распознавание и выделение семантических отношений между концептами вида:
  - Субъект – Акция – Объект – Обстоятельства [4] (SAO);
  - Причина – Следствие [5] (Cause-Effect);
- выделение набора наиболее информативных концептов из текста документа в виде тематической аннотации;
- составление реферата документа в виде множества его наиболее информативных предложений;
- категоризация документов на основе выделенной темы.

Унифицированный подход к обработке текстовой информации с выделением семантических отношений единой структуры вне зависимости от языка позволяет реализовать эффективные схемы управления знаниями в условиях многоязычности.

В частности, подсистема семантического межъязыкового поиска реализуется посредством применения общих иерархических тезаурусов онтологий, разработанных для различных языков (английский, французский, немецкий, японский) в структуре, аналогичной WordNet [6]. Анализ ЕЯ-запроса пользователя при этом включает перечисленные выше этапы обработки текста, в том числе частичное или полное снятие смысловой неоднозначности концептов с определением места в тезаурусе, что позволяет с использованием многоязычных онтологий переводить поисковой образ на другие поддерживаемые системой языки и производить высокоэффективный семантический поиск в соответствующих частях многоязычной базы знаний.

ЗАКЛЮЧЕНИЕ

Разработанный процессор онтологий обладает рядом важных преимуществ, а именно:

- Обеспечивается выделение концептов, их свойств, структуры и онтологических отношений с другими концептами из текста в универсальной форме вне зависимости от языка документа.
- Реализован высокоэффективный межъязыковой поиск с учетом синонимов в пределах смысла концептов без привлечения систем машинного перевода, уровень качества которых в настоящий момент не достаточен для решения подобного рода задач.
- Поддержка любого нового языка в подсистеме межъязыкового поиска не требует реализации механизмов попарного перевода между всеми поддерживаемыми языками. Для этого достаточно реализовать приведенную схему для нового языка.

Описанная в контексте реализации системы *Goldfire* технология позволяет осуществить эффективное наполнение баз знаний, а также их использование в условиях многоязычной среды.

Литература

1. Интернет адрес: http://www.w3.org
2. Интернет адрес: www.invention-machine.com
3. *Совпель И. В.* Система автоматического извлечения знаний из текста и ее приложения // Ж. «Искусственный интеллект». 2004. № 3. С.668–679.
4. *Tsourikov V. M., Batchilo L. S., Sovpel I. V.* Document semantic analysis/selection with knowledge creativity capability utilizing subject-action-object (SAO) structures // US Patent 6. 167. 370. 2000.
5. *Todhunter J., Sovpel I., Pastanohau Dz., Vorontsov A.* Semantic processor for recognition of cause-effect relations in natural language documents // US Patent Application 20060041424. 2006.
6. WordNet: a Lexical Database for English. Princeton University. Princeton, NJ, 2001.