# Информатика и вычислительная техника

УДК 681.3.07

С.В. Клименко[1]
О.В. Золотарев[2]
М.М. Шарнин[3]

## ИСПОЛЬЗОВАНИЕ ОНТОЛОГИЧЕСКОГО ПОДХОДА ДЛЯ АНАЛИЗА ТЕКСТОВ ЕСТЕСТВЕННОГО ЯЗЫКА

*В данной статье рассматриваются подходы по построению онтологий предметной области на основе анализа текстов естественного языка. Определяются понятия «онтология», «семантический поиск», описывается структура онтологии, анализируются различные аспекты формирования поисковых запросов, рассматриваются особенности извлечения информации из текстов естественного языка.*

***Ключевые слова:*** *онтология, семантический поиск, анализ естественного языка, структура онтологии.*

S.V. Klimenko
O.V. Zolotarev
M.M. Charnine

## ONTOLOGICAL APPROACH FOR THE ANALYSIS OF NATURAL LANGUAGE TEXTS

*This article discusses the approaches for developing onthologies of the subject area based on the analysis of natural language texts. The article defines the concepts of ontology, semantic search, and describes the structure of the ontology, different aspects of the formation of a search query, the features of information extraction from natural language texts.*

***Keywords:*** *ontology, semantic search, natural language analysis, ontology structure.*

### 1. Основные элементы онтологии

Онтология представляет собой структурное описание предметной области, включающее словари, термины, отношения.

Основные элементы (примитивы) онтологии – персоны (примеры), концепции (классы, понятия), свойства (роли или отношения), типы данных (конкретные домены), аксиомы:

– персоны, например Сидоров;

– общие понятия, например сообщества индивидуумов, – человечество;

– свойства определяют взаимосвязи между парами индивидуумов предметной области или между индивидуумом и типом данных, например свойство *isParentOf* может использоваться для определения того, что один человек является родителем другого человека. Свойство *hasAge* может относиться к человеку и выражать его возраст. Свойства также могут определяться посредством неких отношений;

– типы данных представляют конкретные понятия, например числовые данные (действительные, целые, неотрицательные и т.д.), строковые, логические, понятия даты, времени, XML литералы и т.д.;

– аксиомы – некоторые утверждения, кото-

[1] Доктор физико-математических наук, профессор, главный научный сотрудник АНО «Институт физико-технической информатики».

[2] Кандидат технических наук, доцент, заведующий кафедрой информационных систем в экономике и управлении АНО ВО «Российский новый университет».

[3] Кандидат технических наук, ведущий научный сотрудник Федерального исследовательского центра «Информатика и управление» Российской академии наук.

рые могут быть проверены посредством анализа элементов онтологии. Онтология включает множество аксиом, обычно разделенных на три части: утверждения, термины, роли. Например, аксиома может утверждать, что Иван – человек, относится к более общему понятию – человечеству.

Формально модель онтологии можно представить следующим образом:

$$O = < T, A, R, F, TR >, \quad (1)$$

где T – термины прикладной области, описываемой онтологией O; A – аксиомы; R – отношения между терминами заданной прикладной области; F – функции интерпретации, заданные на терминах и/или отношениях онтологии O; TR – триплеты, определяющие выражения типа «субъект – отношение – объект».

Онтология обеспечивает словари для представления и обмена знаниями о некоторой предметной области и множество связей, установленных между терминами в этих словарях.

В основе онтологического анализа лежит описание системы в терминах сущностей, отношений между ними и преобразование сущностей, которое выполняется в процессе решения определенной задачи. Онтологический инжиниринг подразумевает глубокий структурный анализ предметной области. Основным преимуществом онтологического инжиниринга является целостный подход к автоматизации предприятия [1–7].

**2. Сервисы, характерные для онтологии**

Наиболее типичные сервисы в онтологиях спроектированы для того, чтобы проверять следующее:

1) согласованность. Онтология является согласованной, если она имеет модель, которая поддерживает интерпретацию, удовлетворяющую каждой аксиоме онтологии;

2) непротиворечивость. Онтология O является непротиворечивой, если любая модель онтологии удовлетворяет аксиоматике O;

3) реализуемость. Концепция C реализуема в онтологии O, если она не интерпретируется как пустое множество для некоторой модели онтологии O;

4) категоризация. Концепция D включает концепцию C для онтологии O, если C интерпретируется как подмножество D в каждой модели онтологии O;

5) классификация. Для онтологии O должна быть задана иерархия понятий на основе отношения подчиненности [5; 8–13].

**3. Семантический поиск**

Всегда существует некая разница между тем, что пользователь приводит в запросе, и тем, что он получает в результате поиска в Сети. Если в запросе присутствуют термины со множеством значений, это дополнительно усложняет построение ответа на запрос пользователя в силу двусмысленности.

Чаще всего запросы состоят из двух-трех слов. Для снятия неопределенности в поисковых запросах необходимо вводить дополнительные определяющие слова, чтобы обеспечить семантический поиск в Сети. Данный подход используется для категоризации результатов запросов, формируя более релевантный ответ на запрос пользователя.

Поиск может проводиться с использованием ключевых слов или формальных SQL-запросов. Семантический поиск позволяет совместить эти два вида представления, чтобы получить наиболее полный ответ на введенный запрос. При этом нет необходимости погружаться в тонкости синтаксиса запросов к базам данных.

Задача разбора запроса системой сводится к следующему.

1. Определить значение каждого ключевого слова в поисковом запросе.

2. Дать каждому ключевому слову интерпретацию и выразить его в формальном языке.

3. Представить пользователю реальную информацию, характеристики и возможности поисковой системы [14–17].

**4. Раскрытие смысла ключевых слов**

Смысл ключевых слов заложен в контексте, окружении ключевого слова. Смысл – это конкретное значение слова в данном контексте.

Этапы процесса раскрытия смысла ключевых слов.

1. Определение смысла ключевого слова.

Для извлечения смысла ключевых слов могут использоваться различные инструменты WatsOn, DBpedia, другие лексические ресурсы Dynamic Ontology Pool, Lexical Database WordNet.

При поиске используется синонимия. В результате получаем кандидатов на ключевые смысловые слова из онтологий для трактовки входных ключевых слов.

2. Обогащение смысла ключевых слов и удаление избыточности.

При извлечении данных из различных онтологий возможно возникновение избыточности. Инкрементный алгоритм используется для выравнивания различных смыслов ключевых слов.

Вероятная синонимия включает лингвисти-

ческие и структурные характеристики онтологий.

Лингвистическое сходство определяется рассмотрением различных меток каждого термина, структурное сходство вычисляется путем сравнения уровней детализации в структуре описания онтологии. Лингвистические и структурные значения комбинируются, чтобы получить результирующие объединенные характеристики похожести.

Смыслы объединяются, когда оценочная вероятность синонимии превышает определенный предел. В результате получаем множество различных вероятных смыслов для каждого введенного ключевого слова.

3. Избавление от двусмысленности ключевых слов. В результате будут получены семантически определенные ключевые слова, на основе которых может быть сгенерирован семантический запрос. Информация извлекается из разных онтологий, хранилищ данных. Полученная информация должна быть интегрирована. В результате будут получены данные, релевантные запросу.

Процесс избавления от двусмысленности значений смыслов происходит за счет выбора наиболее вероятного смысла для каждого ключевого слова. Смыслы сравниваются посредством комбинирования разных вариантов. Это:

– традиционный поиск в Google или Yahoo;

– пересечение между словами в контексте и словами в выделенном семантическом описании;

– частота использования смысла [18–20].

**5. Извлечение информации из текстов естественного языка**

Извлечение информации из текстов естественного языка требует разработки следующих функций.

• Анализ естественного языка В процессе анализа производится разметка текста, разбор предложений, выстраивается грамматическая иерархия: текст – параграф – предложение – придаточное предложение – фраза – слово – выделение морфем.

• Выделение именованных сущностей.

• Машинное обучение.

• Извлечение графической информации.

• Извлечение географической информации.

• Автоматическая классификация документов.

• Семантический анализ. Выделение терминов из контекста, их реальных значений, анализ отношений между словами, поиск синонимов, антонимов, гиперонимов (слово с более широким значением, выражающее общее, родовое понятие, название класса), гипонимов (понятие, выражающее частную сущность по отношению к другому, более общему понятию), исключение двусмысленных понятий. Данная задача решается с использованием средств обработки естественного языка и обращений к онтологиям.

• Автоматическое расширение поисковых запросов. Обеспечивается поиск с использованием ключевых слов с учетом семантического контекста ключевых слов и отношений между ними. Процедура заключается в трех шагах.

1. Определение слов, относящихся к той же сфере деятельности, что и ключевое слово в запросе.

2. Поиск слов с похожим значением – синонимов, анализ предыдущих запросов для исключения двусмысленности.

3. Очищение запросов с именованными сущностями.

• Интеграция. Подготовка отчетов на основе шаблонов для представления результатов анализа. Для подготовки отчета используются базы данных, документы, базы данных триплетов, объединенные данные из Интернета [21].

В статье были рассмотрены подходы к формированию структуры онтологий для поиска и анализа информации в Сети. Использование семантических запросов в Интернете в существенной степени расширяет возможности поиска и извлечения релевантной информации. Это особенно актуально по причине наличия большого количества спама в сети Интернет. Объединение возможностей поиска по ключевым словам и на основе информации из онтологий позволит в существенной степени сократить затраты на нахождение необходимых сведений в Интернете.

**Литература**

1. Zolotarev, O.V., Charnine, M.M., Matskevich, A.G., Kuznetsov, K.I. Business Intelligence Processing on the Base of Unstructured Information Analysis from Different Sources Including Mass Media and Internet // Proceedings of the 2015 International Conference on Artificial Intelligence (ICAI 2015). – Vol. I, WORLDCOMP'15, July 27–30, 2015. – Las Vegas Nevada, USA. – V. I. – Pp. 295–299.

2. Galina, I.V., Charnine, M.M., Somin, N.V., Nikolaev, V.G., Morozova, Y.I., Zolotarev, O.V. Method for Generating Subject Area Associative Portraits: Different Examples // Proceedings of the 2015 International Conference on Artificial Intel-

ligence (ICAI 2015). – Vol. I, WORLDCOMP'15, July 27–30, 2015. – Las Vegas Nevada, USA. – V. I. – Pp. 288–294.

3. Zolotarev, O., Charnine, M., Matskevich, A. A Conceptual Business Process Structuring by Extracting Knowledge from Natural Language Texts // Proceedings of the 2014 International Conference on Artificial Intelligence (ICAI 2014). – Vol. I, WORLDCOMP'14, July 21–24, 2014. – Las Vegas Nevada, USA. CSREA Press. – Pp. 82–87.

4. Золотарев О.В., Шарнин М.М., Клименко С.В. Семантический подход к анализу террористической активности в сети Интернет на основе методов тематического моделирования // Вестник Российского нового университета. Серия «Сложные системы: модели, анализ и управление». – 2016. – Выпуск 3. – С. 64–71.

5. Bobed, Carlos, Yus, Roberto, Bobillo, Fernando, Ilarri, Sergio, Bernad, Jorge, Mena, Eduardo, Trillo-Lado, Raquel and Ángel Luis Garrido. Emerging Semantic-Based Applications // Springer International Publishing AG Switzerland is part of Springer Science+Business Media. Semantic Web. 2016. – P. 39–85.

6. Золотарёв О.В., Шарнин М.М., Клименко С.В., Кузнецов К.И. Система PullEnty – извлечение информации из текстов естественного языка и автоматизированное построение информационных систем // Ситуационные центры и информационно-аналитические системы класса 4i для задач мониторинга и безопасности (SCVRT2015-16) : труды Международной научной конференции : в 2-х томах. – Пущино, 2016. – С. 28–35.

7. Клименко С.В., Шарнин М.М., Хакимов А.Х., Золотарев О.В., Мацкевич А.Г. Методы оценки качества и влияния (IMPACT) научных статей для повышения объективности индекса научного цитирования // Вестник Российского нового университета. Серия «Сложные системы: модели, анализ и управление». – 2016. – Выпуск 3. – С. 51–59.

8. Хакимова А.Ф., Шарнин М.М., Клименко С.В., Золотарев О.В., Родина И.В. Мера подобия текстов как инструмент оценки интертекстуальности при анализе больших коллекций документов // Вестник Российского нового университета. Серия «Сложные системы: модели, анализ и управление». – 2016. – Выпуск 4. – С. 62–71.

9. Цимбалов А.В., Золотарев О.В. Метод шинглов // Вестник Российского нового университета. Серия «Сложные системы: модели, анализ и управление». – 2016. – Выпуск 4. – С. 72–79.

10. Золотарев О.В., Козеренко Е.Б., Шарнин М.М. Проведение аналитической разведки на основе анализа неструктурированной информации из различных источников, включая Интернет и средства массовой информации // Вестник Российского нового университета. – 2015. – Выпуск 9. – С. 49–54.

11. Шарнин М.М., Шагаев И., Протасов В.И., Родина И.В., Золотарев О.В., Попова О.А. Использование веб-семантики для совершенствования образовательных программ вузов // Вестник МГГУ им. М.А. Шолохова. Филологические науки. – 2015. – № 2. – С. 97–112.

12. Шарнин М.М., Золотарев О.В., Сомин Н.В. Извлечение и обработка знаний из неструктурированных текстов деловой сферы и социальных сетей // Труды IV Международной научно-практической конференции «Социальный компьютинг: основы, технологии развития, социально-гуманитарные эффекты», Москва, МПГУ, 2014. – 22–24 октября.

13. Михеев М.Ю., Сомин Н.В., Галина И.В., Золотарев О.В., Козеренко Е.Б., Морозова Ю.И., Шарнин М.М. Фальштексты: классификация и методы опознания текстовых имитаций и документов с подменой авторства // Информатика и ее применения. – 2014. – Т. 8. – Выпуск 4.

14. Золотарев О.В., Шарнин М.М. Методы извлечения знаний из текстов естественного языка и построение моделей бизнес-процессов на основе выделения процессов, объектов, их связей и характеристик // Труды XIX Международной конференции СРТ2014. Ларнака, Кипр, 12–18 мая 2014. – М. : Изд-во Института физико-технической информатики (ИФТИ), 2015. – С. 92–98.

15. Золотарев О.В. Козеренко Е.Б., Шарнин М.М. Принципы построения моделей бизнес-процессов предметной области на основе обработки текстов естественного языка // Вестник Российского нового университета. – 2014. – Выпуск 4. – С. 82–88.

16. Золотарев О.В. Методы выделения процессов, объектов, отношений из текстов естественного языка // Проблемы безопасности российского общества. – Смоленск : Свиток, 2014. – № 3–4. – С. 276–283.

17. Золотарев О.В. Инновационные решения в формировании функциональной структуры предметной области // Вестник Российского нового университета. – 2013. – Выпуск 4. – С. 82–84.

18. Золотарев О.В. Методы и инструменты моделирования предметной области // Цивилизация знаний: проблемы социальных комму-

никаций : труды Тринадцатой Международной научной конференции, г. Москва, 20–21 апреля 2012 г. – М. : РосНОУ, 2012.

19. Золотарев О.В. Новые подходы в построении функциональной структуры предметной области // Сборник трудов по материалам конференции «20 лет постсоветской России: кризисные явления и механизмы модернизации». – Екатеринбург : Гуманитарный университет, 2011.

20. Золотарев О.В. Формализация знаний о предметной области на основе анализа естественно-языковых структур // Цивилизация знаний: проблема человека в науке XXI века : труды Двенадцатой Международной научной конференции, г. Москва, 22–23 апреля 2011 г. – М. : РосНОУ, 2011.

21. Золотарев О.В. Средства анализа информации в системах, основанных на семантических сетях // Цивилизация знаний: проблемы модернизации России : труды Одиннадцатой Международной научной конференции, Москва, 23–24 апреля 2010 г. – М. : РосНОУ, 2010.