УДК 004.8:004.912

Н.В. Рябова, О.В. Шубкина

Харьковский национальный университет радиоэлектроники, Харьков

ОБОБЩЕННАЯ МОДЕЛЬ СЕМАНТИЧЕСКОГО АННОТИРОВАНИЯ ТЕКСТОВЫХ ДОКУМЕНТОВ В СИСТЕМАХ УПРАВЛЕНИЯ ЗНАНИЯМИ

В работе рассмотрены вопросы интеллектуального анализа документов в системах управления знаниями. Проведен анализ основных особенностей семантического аннотирования текстовых документов, введены его формальное представление и основные характеристики.

Ключевые слова: *семантическое аннотирование, управление знаниями, классификация.*

Введение

Менеджмент знаний, несмотря на междисциплинарный характер данного направления исследований, в настоящее время наиболее активно развивается в искусственном интеллекте, предлагая прогрессивные решения, основанные на парадигме Semantic Web [1], по созданию систем корпоративной памяти (КП) предприятий, порталов знаний и др. [2]. Основное предназначение КП состоит в накапливании, систематизации, управлении и совместном использовании профессиональными группами сотрудников всей необходимой и полезной информации в целях достижения конкурентного преимущества организации, эффективного и своевременного решения текущих задач, исключении дублирования, противоречивости и потери знаний, накопленных в процессе жизнедеятельности организации [3, 4]. Поскольку большая часть информации, как в корпоративных Intranet-сетях, так и в Web-пространстве, содержится в текстовом виде [5], технологии интеллектуальной обработки текстов помогают решать многие задачи на основе извлечения знаний из текстовых коллекций, их структурирования и анализа.

Одним из таких подходов является семантическое аннотирование текстовых документов, которое заключается в создании мета описания документа на основе существующего корпуса текстов и онтологии, отображающей структуру знаний компании.

Существует набор стандартных решений, которые разработаны для описания метаданных и формирования семантических аннотаций, как например, стандарт Dublin Core [6]. Однако набор заданных тегов для описания текстовых документов не отражает информацию, которая может являться актуальной для текущей онтологии предметной области, а зачастую несет лишь общие сведения. Стоит отметить также, что создание семантических аннотаций вручную занимает достаточно много времени и требует больших денежных затрат. Это привело к разработке методов полуавтоматического построения семантических аннотаций, которые в свою очередь имеют ряд недостатков, например, использование шаблонов заполнения или априори заданных правил. Поэтому актуальной задачей в системах управления знаниями является разработка моделей и методов семантического аннотирования текстовых документов.

Анализ проблемы и постановка задач исследования. В современных системах управления знаниями (СУЗ) используется синтез нескольких типов информации: онтологий, текстовых документов, внешних источников данных (баз данных), баз знаний. Любую предметную область можно описать через объекты, которые в ней находятся, а также связи между ними. Каждый объект характеризуется некоторым набором атрибутов и их значений, которые формируют набор экземпляров концептов.

Ввиду того, что большая часть информации в корпоративных системах хранится в текстовом виде (в виде электронных документов, рассылок новостей), каждому экземпляру концептов онтологии, отражающей структурные знания, можно поставить в соответствие какой-либо текстовый документ или какую-то его часть в зависимости от заданных условий. Такой процесс формирования метаданных называется семантическим аннотированием [7] и состоит из трех основных компонент: онтологии, корпуса текстов и способа построения классификатора для извлечения знаний. Метаданные могут ранжироваться, начиная от ссылок на автора документа до упоминаний обо всех компаниях и событиях, указанных в тексте [8].

Семантическая аннотация, получаемая в процессе работы, должна раскрывать семантическое значение каждого аннотируемого документа и должна представлять собой набор метаданных, выделенных из текстового документа в результате его проекции на онтологию. Благодаря такому описанию документа любая программа или программный агент, использующие эту же онтологию, может понять также смысл документа. Исходя из вышесказанного, в работе предлагается обобщенная схема семантического аннотирования текстовых документов, а также его формальное представление и основные характеристики.

Описание процесса семантического аннотирования текстовых документов

Процесс семантического аннотирования включает следующие этапы (рис. 1):

– выделение необходимых текстовых объектов с учетом выбранного уровня семантического аннотирования;

– формирование вектора признаков для каждого объекта;

– построение классификатора на основе онтологии и полученного набора данных;

– выделение метаданных с использованием классификатора;

– преобразование метаданных каждого документа в формат RDF [9].

В первую очередь следует отметить, что создание семантических аннотаций текстовых документов начинается с определения уровня аннотирования и построения соответствующего пространства признаков (может рассматриваться уровень документов, предложений, слов). Самая низкая ступень представляет собой уровень слов и начинается с выделения термов – лингвистических (именованных) сущностей, представляющих собой именные группы, полученные на основе лексического анализа корпуса текстов предметной области. Для идентификации термов лучше использовать стандартные приложения, например, такие как GATE framework [10], по-

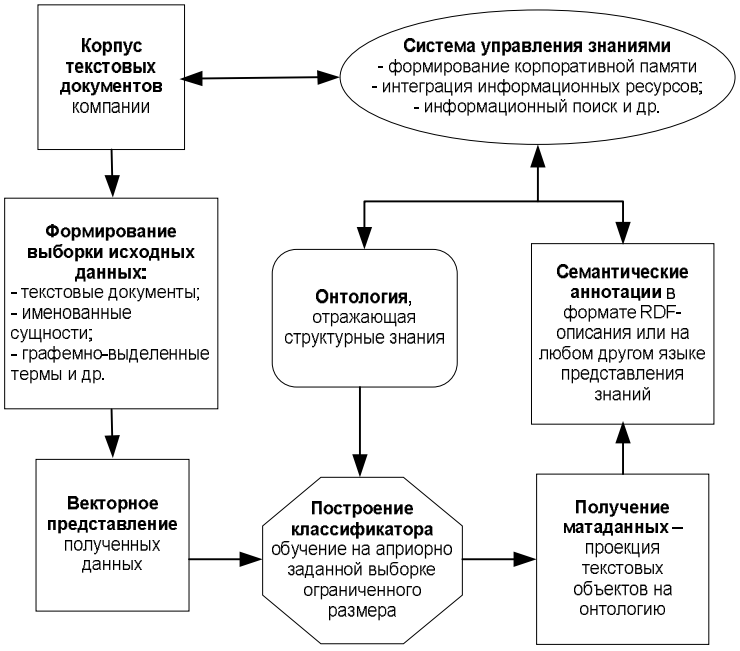

Рис. 1. Обобщенная схема семантического аннотирования текстовых документов

зволяющий производить морфологический поиск, выделение предложений, анализ частей речи и выявление конструкций именных групп с помощью регулярных выражений для различных частей речи и компонентов слова. Примером выделенных именованных сущностей являются персоны, организации, географические объекты и прочие объекты, обозначаемые в тексте с использованием имен собственных, различных специальных символов (кавычек и прочее) [11].

Кроме того, в качестве текстовых объектов из документов могут извлекаться события [12] или графемно-выделенные слова [13]. Например, концепт «университет», «кафедра», экземпляры этих концептов – «ХНУРЭ» и «кафедра искусственного интеллекта» соответственно.

Вторым этапом семантического аннотирования является векторное представление исходных данных. В общем случае атрибуты (признаки) вектора кодируют контекстную информацию контента и формируют распределенное векторное пространство. В частности, в качестве контекстных признаков могут выступать частоты появления слова в различных документах корпуса [14]. Важен сам принцип формирования набора подходящих признаков для описания исходных текстовых данных, которые в процессе распознавания выделяются из корпуса текстов. Следует отметить, что способ построения векторного пространства зависит от выбранного уровня аннотирования. В данной задаче, как показано в [15], для обобщения концептом онтологии может выступать непосредственно текстовый документ, следовательно, наиболее предпочтительным алгоритмом для получения вектора признаков текстовых данных является статистическая оценка TFIDF, описанная в [16]. Далее для удобства будем называть исходные данные текстовыми объектами, которые получены после предобработки и соответствующего кодирования необходимых признаков. Таким образом, получаем

$$\text{ObjectSet} = \left(x(1), x(2), \ldots, x(j), \ldots, x(N)\right),$$

где $x(j)$ – j-й текстовый объект, представленный в виде некоторого набора релевантных признаков большой размерности; N – мощность исходной выборки текстовых объектов соответственно.

Процесс семантического аннотирования можно рассматривать как проблему классификации, при этом автоматизация может быть достигнута путем применения методов интеллектуального анализа данных. Главной задачей в таком случае является построение классификатора на основе онтологии Ont и исходной выборки данных ObjectSet для получения проекции текстового объекта $x(j)$ на онтологию предметной области. Под проекцией в данном случае следует понимать отнесение $x(j)$ к определенному классу онтологии как экземпляра концепта данного класса.

Для текущей онтологии, отражающей структурные знания, Ont набор концептов (классов) определяется как $\text{ConceptSet} = \left(c_1, c_2, \ldots, c_i, \ldots c_{N_1}\right)$, где c_i – i-й концепт из Ont, N_1 – количество концептов (классов) онтологии. При этом следует обратить внимание, что для уменьшения неоднозначности и повышения эффективности дальнейшей интеграции различных информационных источников необходимо найти класс c_i, максимально соответствующий данному текстовому объекту $x(j)$. Однако возможность принадлежности одного объекта к нескольким классам не исключается.

Стоит отметить также, что в настоящее время в современных разработках с использованием интеллектуальных технологий хорошо зарекомендовали себя искусственные нейронные сети (ИНС) [17,18], поэтому данное формальное описание семантического аннотирования, предоставляет возможность использовать нейросетевой подход для получения семантических аннотаций текстовых документов. Кроме того, для формирования метаданных текстовых документов применение ИНС дает возможность построить классификатор, независимый от уровней семантического аннотирования. В таком случае можно использовать различные способы представления исходного корпуса текстов в векторном пространстве признаков, что не отражается на качестве обработки данных.

Итак, в рассматриваемой задаче мощность N_2 априорно заданной обучающей выборки определяется как некоторая часть ObjectSet, такая, что $N_2 < N$. Для обучения на вход ИНС подается вектор признаков $x = \left(x_1, x_2, \ldots, x_n\right)^T$, размерности $(n \times 1)$. Для построения классификатора в качестве обучающего сигнала, кодирующего множество классов ConceptSet используется сигнал $d_j(k)$, принимающий значение 1, если входной объект x_j принадлежит заданному классу, и -1 – в противном случае. Необходимо принять во внимание, что объем выборки ограничен. Кроме того, исходные данные представляют собой набор релевантных признаков большой размерности, что влияет на выбор конструкции ИНС и предполагает ряд сложностей при использовании стандартных средств, не ориентированных на обработку таких данных.

Значения выхода нейронной сети y формируют множество $\text{LabelSet} = \left\{ l_i \middle| \exists c_j \in \text{ConceptSet} \wedge l_i = c_j \right\}$ – уникальное множество для каждого текстового документа TD, состоящее из концептов (классов) онтологии предметной области, полученных путем проекции множества текстовых объектов ObjectSet, принадлежащих этому документу, на заданную онтологию Ont. Таким образом, получаем семантическую разметку (четвертый этап согласно предложенной обобщенной схеме).

Для описания полученных метаданных удобно использовать RDF или любой другой язык описания

знаний, которые в последнее время активно развиваются в рамках парадигмы Semantic Web, что должно обеспечить компьютерную обработку информации с учетом ее семантики для дальнейшего использования интеллектуальными агентами в системах управления знаниями [19]. Формирование семантических аннотаций на основе множества LabelSet и представление их в RDF-формате будет заключительным этапом для получения семантических аннотаций и включение их в текущую базу знаний. Такой подход позволит избежать неоднозначности (неопределенности) информационного поиска, а также повысить возможность взаимодействия и интеграции информации из гетерогенных источников, накопленных организациями в процессе становления и развития.

Выводы

В работе рассмотрена возможность применения интеллектуального анализа текстовой информации в системах управления знаниями. Проведен анализ особенностей семантического аннотирования текстовых документов, предложены его формальное представление и основные характеристики, что даст возможность получения семантических аннотаций текстовых документов путем классификации данных информационных ресурсов в соответствии с онтологией предметной области, отражающей структурные знания.

Список литературы

1. Хорошевский В.Ф. Пространства знаний в сети Интернет и Semantic Web / В.Ф. Хорошевский // Искусственный интеллект и принятие решений. – 2008. – 1. – С. 80-97.

2. Luger G. Artificial Intelligence: Structures and Strategies for Complex Problem Solving, 5th edition / G. Luger. – UK: Pearson Education. – 2006. – 889 p.

3. Mertensson M. A critical review of knowledge management as a management tool / M. Mertensson // Journal of Knowledge Management. – 2000. – V. 4, No 3. – P. 234-240.

4. Tiwana A. The knowledge management toolkit: orchestrating IT, strategy, and knowledge platforms / A. Tiwana. – Prentice Hall PTR. – 2002. – 388 p.

5. Feldman R. The text mining handbook: advanced approaches in analyzing unstructured data / R. Feldman, J. Sanger. – Cambridge University Press. – 2007. – 410 p.

6. Dublin Core: Metadata Initiative. [Элект. ресурс]. – Режим доступа к ресурсу: http://www.dublincore.org.

7. Semantic annotation for knowledge management: Requirements and a survey of the state of the art / V. Uren, Ph. Cimiano, J. Iria, S. Handschuh, M. Vargas-Vera, E. Motta, F. Ciravegna // Web Semantics: Science, Services and Agents on the World Wide Web. – 2006. – Vol. 4, No. 1. – P. 14-28.

8. KIM – a semantic platform for information extraction and retrieval / B. Popov, A. Kiryakov, A. Kirilov, D. Manov, D. Ognyanoff, M. Goranov // Journal Natural Language Engineering. – 2004. – Vol. 10, No. 4. – P. 375-392.

9. Resource Description Framework: Overview. [Электронный ресурс]. – Режим доступа к ресурсу: http://www.w3.org/RDF/.

10. GATE: an Architecture for Development of Robust HLT Applications / H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan // Recent Advanced in Language Processing. – 2002. – P. 168-175.

11. Extracting named entities using Support Vector Machines / Yu-Ch. Wu, T.-K. Fan, Y.-Sh. Lee, Sh.-J. Yen // Lecture Notes in Computer Science. – 2006. – P. 91-103.

12. Terrorism Event Classification using Fuzzy Inference Systems / U. Inyaem, Ph. Meesad, Ch. Haruechaiyasak, D. Tran // International Journal of Computer Science and Information Security. – 2010. – Vol. 7, No. 3. – P. 247-256.

13. Habib M. Addressing scalability issues of named entity recognition using Multi-class Support Vector Machines / M. Habib // World Academy of Science, Engineering and Technology. – 2008. – 37. – P. 69-78.

14. Chifu E. Şt. A Neural Model for Unsupervised Named Entity Classification / E. Şt. Chifu, V. R. Chifu // The IEEE International Conference on Intelligent Agents, Web Technologies and Internet Commerce. – 2008. – P.1073-1078.

15. Todorov K. Mining concept similarities for heterogeneous ontologies / K. Todorov, P. Geibel, K.-U. Kühnberger // Lecture Notes in Computer Science. – 2010. – 6171. – P. 86-100.

16. Salton G. Term-weighting approaches in automatic text retrieval / G. Salton, C. Buckley // Information Processing & Management. – 1988. – 24(5). – P. 513-523.

17. Rojas R. Neural Networks. A Systematic Introduction. / R. Rojas. – Berlin: Springer – Verlag, 1996. – 502 p.

18. Бодянский Е.В. Искусственные нейронные сети: архитектуры, обучение, применения / Е.В. Бодянский, О.Г. Руденко. – Х.: ТЕЛЕТЕХ, 2004. – 372 с.

19. Fensel D. Ontology-based knowledge management / D. Fensel // Computer. – 2002. – Vol. 35, No. 11. – P. 56-59.

Поступила в редколлегию 2.12.2010

Рецензент: д-р техн. наук, проф. Е.И. Кучеренко, Харьковский национальный университет радиоэлектроники, Харьков.

УЗАГАЛЬНЕНА МОДЕЛЬ СЕМАНТИЧНОГО АНОТУВАННЯ ТЕКСТОВИХ ДОКУМЕНТІВ У СИСТЕМАХ УПРАВЛІННЯ ЗНАННЯМИ

Н.В. Рябова, О.В. Шубкіна

У роботі розглянуті питання інтелектуального аналізу документів в системах управління знаннями. Проведено аналіз основних особливостей семантичного анотування текстових документів, введені його формальне подання та основні характеристики.

Ключові слова: *семантичне анотування, управління знаннями, класифікація.*

GENERALIZED SEMANTIC ANNOTATION MODEL FOR TEXT DOCUMENT IN THE KNOWLEDGE MANAGEMENT SYSTEMS

N.V. Ryabova, O.V. Shubkina

Text mining in the knowledge management systems is considered. The analysis of the main features for the text document semantic annotation is entered; its formal presentation and basic characteristics are explained.

Keywords: *semantic annotating, management, classification, knowledges.*