УДК 004.912:004.8

О.В. Шубкина

Харьковский национальный университет радиоэлектроники, Харьков

МОДЕЛИ СЕМАНТИЧЕСКОГО АННОТИРОВАНИЯ ТЕКСТОВЫХ ДОКУМЕНТОВ С ИСПОЛЬЗОВАНИЕМ ИСКУССТВЕННЫХ НЕЙРОННЫХ СЕТЕЙ СПЕЦИАЛЬНОГО ВИДА

В работе предложены модель семантического аннотирования текстовых документов с учетом бинарных выходов искусственной нейронной сети и вероятностная модель семантического аннотирования текстовых документов. Проведен анализ основных этапов для построения системы семантического аннотирования, предложена ее структурная схема и приведены результаты работы.

Ключевые слова: *семантическое аннотирование, искусственные нейронные сети, структурная схема.*

Введение

В настоящее время широкое развитие получили методы и технологии интеллектуальной обработки текстовой информации. Одним из таких подходов является семантическое аннотирование, заключающееся в присвоении семантических тегов текстовым документам. Как известно, ранее широко использовались такие языки разметки, как Dublin Core [1], а также целый ряд микроформатов, задающих определенный набор тегов. В области информационного поиска метаданные чаще всего рассматривают как набор элементов, описывающих основные свойства текстового документа (включая тематику), которые могут также содержать элементы из уже существующих схем.

Однако вместе с развитием онтологий, которые используются как базы знаний предметной области, подход семантического аннотирования стал все более актуальным. Наиболее известные работы в этой области исследований [2] предполагают получение семантических тегов или метаданных как проекцию текстового документа на онтологию предметной области. Ряд работ [3] включает наполнение онтологии новыми концептами и соответственно выделение новых тегов. Таким образом, становится возможным извлечение знаний из текстовой информации для их последующего использования в качестве машинно-понятного описания ресурса. В зависимости от выбранной стратегии для автоматического получения семантических аннотаций используются подходы, основанные на классификации или кластеризации данных.

Разработанные методы семантического аннотирования можно разделить на две группы [4]: внутреннее аннотирование, заключающееся во встраивании семантической разметки внутрь документа, и внешнее аннотирование, включающее сохранение набора метаданных в RDF-формате [5] в отдельном файле. 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, параграфов, предложений, слов).

В данной работе процесс семантического аннотирования рассматривается как проблема классификации, при этом автоматизация достигается с помощью 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 классу онтологии как экземпляра концепта данного класса.

Постановка задачи исследования

Семантическая аннотация – разметка или набор метаданных рассматриваемого текстового документа – на основе данной онтологии будет определена как $\mathrm{LabelSet} = \left\{ l_i \middle| \exists c_j \in \mathrm{ConceptSet} \wedge l_i = c_j \right\}$, в котором LabelSet – уникальное множество для каждого текстового документа, состоящее из концептов (классов) онтологии предметной области, полученных путем проекции множества текстовых объектов ObjectSet, полученных после предобработки данных и принадлежащих этому документу, на заданную онтологию с использованием методов на основе ИНС. Полученная разметка может быть впоследствии представлена с помощью разработанных моделей семантического аннотирования для формирования RDF-описания и его дальнейшего использования различными программными средствами.

Модель семантического аннотирования текстовых документов с учетом бинарных выходов искусственной нейронной сети

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.

Базовый строительный блок RDF-модели – утверждение, представляющее собой тройку: ресурс, именованное свойство и его значение. В терминологии RDF эти три части утверждения называются соответственно: субъект, предикат и объект.

Ресурсом называют все, что описывается средствами RDF. Это может быть обыкновенная Web-страница или какая-то ее часть, например, отдельный элемент HTML или XML разметки, являющийся частью описываемого документа. Также ресурсом может быть целая коллекция страниц, например, отдельно взятый Web-сайт. И, наконец, в качестве ресурса может выступать нечто, не являющееся доступным непосредственно через Интернет, например, некоторый источник из внутренней текстовой базы данных компании. Одним словом, все, чему можно приписать некоторый URI (универсальный идентификатор) или URI с добавлением внутреннего имени объекта (имени якоря в HTML) может стать ресурсом и быть описано при помощи RDF. Применительно к задаче семантического аннотирования логику формирования метаданных после этапа классификации можно представить в следующем виде. Обозначим искомое множество «ресурс, именованное свойство и его значение» в рассматриваемых терминах. Как было отмечено, на входе классификатора имеем

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

где x(j) – j-й текстовый объект, представленный в виде некоторого набора релевантных признаков, N – мощность исходной выборки текстовых объектов соответственно. Предположим, что каждый обрабатываемый текстовый объект x(j) из множества объектов ObjectSet имеет некоторый уникальный идентификатор ID ($\text{URI} \subset \text{ID}$).

Свойством, относительно которого строится классификатор, является определение принадлежности входного объекта некоторому текущему концепту ConceptSet из онтологии Ont. При этом

$$\text{ConceptSet} = \left(\text{c}_1, \text{c}_2, \ldots, \text{c}_\text{i}, \ldots \text{c}_{\text{N}_1}\right),$$

где c_i – i-й концепт из Ont, N_1 – количество концептов (классов) онтологии. Значение этого свойства, полученное на выходе *y* ИНС, является значением LabelSet для каждого текстового документа, $\text{LabelSet} = \left\{\text{l}_\text{i} \middle| \exists \text{c}_\text{j} \in \text{ConceptSet} \wedge \text{l}_\text{i} = \text{c}_\text{j}\right\}$. В таком случае искомый триплет для представления семантической разметки можно получить из значений множеств $\text{ID}, \text{LabelSet}, \text{y}$. Тогда модель семантического аннотирования текстовых документов с учетом бинарных выходов искусственной нейронной сети:

$$\text{metadata}_\text{i} = \left\{ \begin{matrix} \text{id}, \text{l}_\text{j}, \text{value}_\text{k} \middle| \exists \text{id} \in \text{ID} \wedge \\ \wedge \exists \text{l}_\text{j} \in \text{LabelSet} \wedge \exists \text{bin\_value}_\text{k} \in \text{y} \end{matrix} \right\}, \quad (1)$$

где id – уникальный идентификатор для текущего текстового объекта, $_\text{j}$ – j-й элемент множества LabelSet, $\text{bin}\_\text{value}_\text{k}$ – бинарное значение, полученное после обработки информации на выходе ИНС, которое соответствует 1 в случае, если текстовый объект принадлежит текущему концепту онтологии, и 0 – в противоположном случае.

Таким образом, получаем представление текстовых документов в машинно-понятной форме путем создания их семантических аннотаций на основе онтологии, отражающей структурные знания.

Предложенная логическая модель описания знаний, полученных из текстовых источников с использованием аппарата на основе ИНС [6], и представление их в RDF-формате является заключительным этапом для формирования семантических аннотаций и включения их в текущую базу знаний. Такой подход позволит избежать неоднозначности (неопределенности) информационного поиска, а также повысит возможность взаимодействия и интеграции информации из гетерогенных источников.

Вероятностная модель семантического аннотирования текстовых документов

Процесс семантического аннотирования в данном исследовании рассматривается как проблема классификации, при этом автоматизация достигается с помощью применения методов интеллектуального анализа данных. В свою очередь классификация образов предусматривает отнесение образов или экземпляров к заранее определенному набору классов или категорий.

Что касается классификации текстовой информации, то в общем случае классы отражают темы, которые рассматриваются в документах, представленных в виде образов или экземпляров. Классификация документов – ключевой и хорошо зарекомендовавший себя инструмент для организации, систематизации, упорядочивания больших объемов текстовой информации. Развитие Web и огромный рост цифрового контента во внутрикорпоративных сетях, базах данных, архивах способствует увеличению предъявляемых требований к категоризации [7]. Несмотря на скорость и сложность этого процесса, классификация вручную является экономически неэффективной, поэтому создание автоматических методов и инструментариев является актуальным.

В большинстве случаев для классификации текстовой информации используются различные статистические методы или методы машинного обучения, потому что на практике значительно легче вывести правила категоризации на основе примеров документов, чем устанавливать их с помощью эксперта предметной области.

Потенциальный недостаток традиционных методов классификации заключается в том, что они рассматривают одноуровневую неразветвленную структуру классов. Наибольшее число исследований было проведено именно в этой области [8]. Однако такой подход не отображает взаимосвязи между категориями, во многих случаях на практике представленными в виде иерархий или таксономий.

Наиболее известными примерами могут служить такие системы, как Всемирная организация интеллектуальной собственности (около 69 тыс. групп патентов), Gene Ontology (около 17 тыс. термов для описания сочетания генов), Web-каталоги, разработанные компанией Yahoo [9], проект Open Directory (около 590 тыс. категорий для Web-страниц). Категории обычно организованы в иерархии или таксономии, наиболее часто с помощью представления соподчиненных концептов и установления связей типа «is-a». В данном случае иерархия является одним из ценных источников знаний о предметной области.

В связи с этим введем модель семантического аннотирования текстовых документов с использованием вероятностного подхода. Ключевым свойством модели должно стать значение вероятностной принадлежности текстового объекта x(j) к каждому классу $probab\_value$, которое вводится как одно из свойств логического описания триплета метаданных.

Такой подход обеспечит возможность оценки отнесения текстового объекта x(j) к различным концептам онтологии, которая, как было отмечено ранее, может быть задана иерархией таксономией. Полученная разметка впоследствии может быть представлена в виде RDF-триплета для дальнейшего использования различными программными средствами.

Таким образом, введем в рассмотрение модель семантического аннотирования текстовых документов на основе вероятностного подхода. Будем использовать искомое множество «ресурс, именованное свойство и его значение» в рассматриваемых терминах, а также введем значение свойства вероятностной принадлежности $probab\_value$. Для определения модели семантического аннотирования с использованием вероятностного подхода введем значение вероятности принадлежности обрабатываемого текстового объекта $probab\_value$ текущему концепту из ConceptSet онтологии предметной области Ont.

Учитывая новое свойство, можем записать вероятностную модель семантического аннотирования текстовых документов в следующем виде:

$$\text{metadata}_i = \left\{ \begin{array}{l} \text{id}, l_j, probab\_value_k \,\middle|\, \exists \text{id} \in \text{ID} \wedge \\ \wedge \exists l_j \in \text{LabelSet} \wedge \\ \wedge \exists probab\_value_k \in y \end{array} \right\}, \quad (2)$$

где id – уникальный идентификатор для текущего текстового объекта, I_j – j-й элемент множества LabelSet, $probab\_value_k$ – значение вероятности принадлежности обрабатываемого текстового объекта концепту онтологии, полученного на выходе работы ИНС [10].

Стоит отметить, что задав некоторое пороговое значение ThresholdValue, которое определяется на основе разницы принадлежности объекта x(j) из множества объектов ObjectSet к каждому классу из множества CoceptSet, появляется возможность выявить текстовые объекты x(j), которые относятся к нескольким из потенциально возможных концептов онтологии Ont. Как показывают исследования [11], такие случаи не стоит исключать, потому что в компаниях с увеличением объема текстовой информации, доступной через внутренние хранилища данных или Интернет, возникает проблема обработки, поиска, хранения такого рода информации и преобразования ее в формат, удобный для использования другими программными средствами. Основной характеристикой полученной модели является то, что с помощью введенного описания становится возможным включение в RDF-триплет значения вероятности принадлежности. Это позволит оценить взаимосвязи, как текстовых документов, так и концептов онтологии, а также делать вывод новых знаний благодаря полученным значениям $probab\_value_k$. Таким образом, получаем представление текстовых документов в машинно-понятной форме на основе онтологии, отражающей структурные знания.

Предложенная модель семантического аннотирования текстовых документов с использованием вероятностного подхода и ее представление в RDF-формате является ключевым этапом для формирования семантических аннотаций и включения их в текущую базу знаний.

Практическая реализация моделей в системе семантического аннотирования текстовых документов

В соответствии с методами семантического аннотирования и моделями, описанными в [6, 10], рассмотрим структурную схему системы автоматического семантического аннотирования текстовых документов. Ниже приведена схема последовательности этапов, реализующих способы семантического аннотирования.

Этап 1. Ввод и анализ корпуса текстовых документов.

Этап 2. Предварительная обработка и векторное представление текстовой информации.

Этап 3. Построение классификатора, использующего нейросетевой подход для обработки текстовых данных и знаний онтологии предметной области.

Этап 4. Работа системы на тестирующей выборке.

Этап 5. Формирование метаданных для семантических аннотаций.

В соответствии с функциями, реализуемыми на каждом из указанных этапов, разработана структурная схема работы системы автоматического семантического аннотирования текстовых документов (рис. 1). Как следует из структурной схемы (рис. 1), на этапе 1 осуществляется ввод корпуса текстовых документов и выбор уровня семантического аннотирования (уровень параграфов, фрагментов или уровень документа в целом). На этапе 2 осуществляется предварительная обработка текстовых документов согласно [12] для получения векторного пространства признаков, а затем представлению каждого текстового объекта в векторном виде.

Этап 3 (рис. 1) предназначен для обучения системы на основе входных данных, которые представлены в некотором векторном виде Система продолжает процесс обучения до тех пора, пока не достигнет выбранной точности. Если положительный результат достигнут, то необходимо перейти к выполнению этапа 4 и работы с тестовыми данными, с помощью чего получаем набор информации для формирования метаданных на этапе 5. Модели семантического аннотирования позволяют получить семантические аннотации для выбранного уровня аннотирования и выбранного способа формирования векторного пространства признаков.

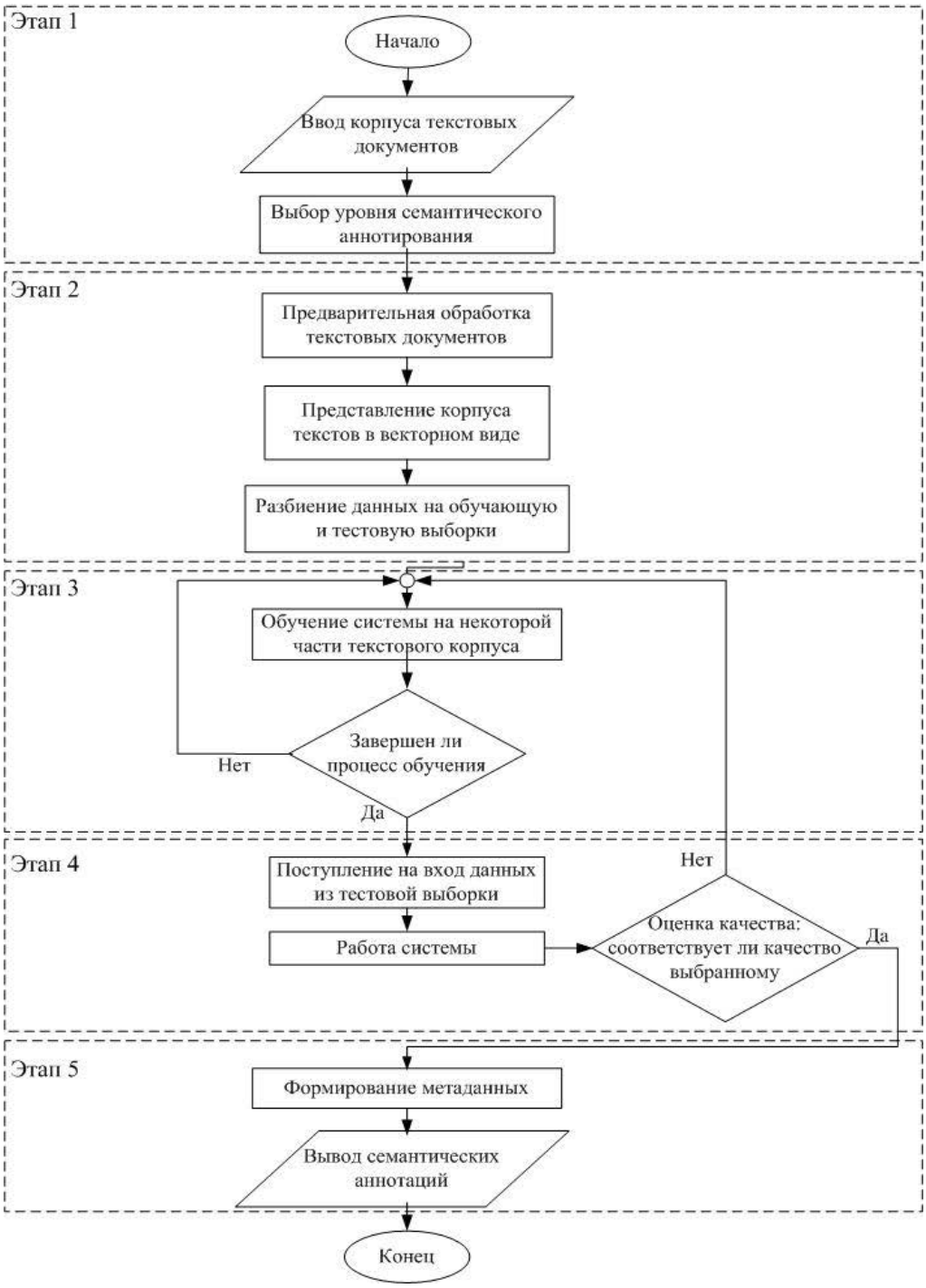

Рис. 1. Структурная схема системы семантического аннотирования текстовых документов

Представленная структурная схема и детализация этапов семантического аннотирования текстовых документов с использование разработанных моделей и методов позволили разработать программное обеспечение, реализующее методы автоматического формирования семантических аннотаций текстовых документов.

Экспериментальными исследованиями было установлено, что предложенные модели семантического аннотирования текстовых документов с использованием искусственных нейронных сетей специального вида позволяют получить описание текстовых документов в машинно-понятной форме на основе текущей онтологии. Данный подход базируется на методах, обладающих высокими показателями точности и быстротой работы, поэтому становится возможным повысить качество извлечения знаний из текстовых источников при ограниченной выборке (в среднем 6-8 %).

Выводы

В работе предложены модель семантического аннотирования текстовых документов с учетом бинарных выходов искусственной нейронной сети и вероятностная модель семантического аннотирования текстовых документов, которые являются развитием RDF-структур, при этом обладают определенно заданными свойствами. Это позволяет формировать RDF-триплеты из полученных после обработки информации данных для включения их в текущую базу знаний или дальнейшего использования программными средствами. Также, в статье предложена структурная схема работы системы семантического аннотирования текстовых документов.

Список литературы

1. Dublin Core: Metadata Initiative [Электронный ресурс]. – Режим доступа: www.dublincore.org – 26.03.2011 г. – Загл. с экрана.

2. Uren, V. Semantic annotation for knowledge management: Require-ments and a survey of the state of the art [Текст] / 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.

3. Xu, G. Web Mining and Social Networking: Techniques and Applications [Текст] / G. Xu, Ya. Zhang, L. Li. – Springer: Web Information Systems Engineering and Internet Technologies Book Series. – 2011. – V. 6. – 210 p.

4. Al-Khalifa, H. The evolution of metadata from standards to semantics in E-learning applications [Текст] / H. Al-Khalifa, H. Davis // Proceedings of the 17th conference on Hypertext and hypermedia table of contents. – Odense, Denmark. – 2006. – P. 69–72.

5. Resource Description Framework: Overview. [Электронный ресурс]. – Режим доступа: http://www.w3.org/RDF/ – 26.03.2011 г. – Загл. с экрана.

6. Бодянский, Е.В. Семантическое аннотирование текстовых документов на основе иерархической радиально-базисной нейронной сети [Текст] / Е.В. Бодянский, О.В. Шубкина // Восточно-Европейский журнал передовых технологий. – 2010. – Вып. 6/3 (48). – С. 72–77.

7. Шубкина, О.В. Интеллектуальный анализ текстов в системах менеджмента знаний [Текст] // Научно-технический журнал «Бионика интеллекта». – 2009. – № 1(70) . – С. 142-146.

8. Tenenboim, L. Ontology-based classification of news in an electronic newspaper [Текст] / L. Tenenboim, B. Shapira, P. Shoval // Int. Book Series "Information Science and Computing". – Sofia: ITHEA, 2009. – Vol. 3. – P. 89-97.

9. Labrou, Ya. Yahoo! as an ontology: using Yahoo! categories to describe documents [Текст] / Ya Labrou, T. Finin // Proceedings of the eighth international conference on Information and knowledge management. – ACM New York. – NY, USA, 1999. – P. 180-187.

10. Бодянский, Е.В. Семантическое аннотирование текстовых документов с использованием модифицированной вероятностной нейронной сети [Текст] / Е.В. Бодянский, О.В. Шубкина // Системные технологии. Региональный межвузовский сборник научных трудов. – Вып. 4 (75). – Днепропетровск, 2011. – С. 48-55.

11. Sebastiani, F. Machine learning in automated text categorization [Текст] / F. Sebastiani // ACM Computing Surveys. – 2002. – Vol. 34(1). – P. 1-47.

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

Поступила в редколлегию 1.03.2011

Рецензент: д-р техн. наук, проф. В.А. Филатов, Харьковский национальный университет радиоэлектроники, Харьков.

МОДЕЛІ СЕМАНТИЧНОГО АНОТУВАННЯ ТЕКСТОВИХ ДОКУМЕНТІВ З ВИКОРИСТАННЯМ ШТУЧНИХ НЕЙРОННИХ МЕРЕЖ СПЕЦІАЛЬНОГО ВИДУ

О.В. Шубкіна

В роботі запропоновані моделі семантичного анотування текстових документів з урахуванням бінарних виходів штучної нейронної мережі та ймовірнісна модель семантичного анотування текстових документів. Проведено аналіз основних етапів для побудови системи семантичного анотування, запропонована її структурна схема та наведено результати роботи.

Ключові слова: *семантичне анотування, штучні нейронні мережі, структурна схема.*

SEMANTIC ANNOTATION OF TEXT DOCUMENTS MODELS SPECIAL FORM ARTIFICIAL NEURAL NETWORKS

O.V. Shubkina

Semantic annotation of text documents model taking into account the binary outputs of artificial neural network is proposed; probabilistic semantic annotation of text documents model is proposed. Analysis of the main stages for the semantic annotation system is carried out, structural schematic model is proposed, and results of the work are considered.

Key words: *semantic annotation, artificial neural network.*