Модель классификации текстовых ресурсов электронного архива на основе онтологии

А.А. Зарубин[1], А.Р. Коваль[1], В.С. Мошкин[2]

[1]Санкт-Петербургский государственный технический университет телекоммуникаций им. проф. М.А. Бонч-Бруевича, Наб. р. Мойки 61, Санкт-Петербург, Россия, 191186
[2]Ульяновский государственный технический университет, Северный венец 32, Ульяновск, Россия, 432027

Аннотация. В данной работе представлена онтологическая модель текстового документа в качестве ресурса электронного архива. В статье также представлен онтологически ориентированный алгоритм классификации технических документов. В заключении представлены результаты экспериментов, подтверждающих эффективность моделей и алгоритмов в решении задачи классификации документов электронного архива.

1. Введение

Задача категоризации текстовых документов для упрощения процесса поиска необходимой информации в крупном электронном архиве организации является как никогда актуальной. В большинстве случаев структуризация архивов производится вручную сотрудниками архива, которые должны обладать необходимыми знаниями в предметной области и учитывать специфику хранимой документации.

Автоматизация процесса категоризации архива электронных текстовых документов должна осуществляться с учетом семантики информации, заложенной в документации, иначе опыт высококвалифицированных специалистов, разрабатывающих данную документацию, будет сложно извлекаем из неструктурированных ресурсов для дальнейшего применения.

В настоящее время исследователями предлагаются различные способы решения подобной задачи.

В [1] алгоритм классификации колоний муравьев используется для классификации данных и применяется для быстрого поиска больших объемов данных интеллектуальных архивов.

В [2] [3] авторы представляют онтологию, предназначенную для моделирования архивного описания коллекций исторических документов. В [4] авторы представляют связанные с Semantic Web аспекты текущей деятельности цифровой библиотеки и представляют их функциональность; они показывают примеры, начиная от общих архитектурных описаний и заканчивая подробным использованием конкретных онтологий. В работе [5] предлагается портал семантического поиска для межкультурных архивов, включающих документы, изображения, аудио и видео.

Одним из возможных решений данной задачи является применение интеллектуальных алгоритмов анализа неструктурированных данных текстовых документов с последующим разбиением архива на классы в соответствии с семантикой предметной области, в которой

работает данная организация. Семантика предметной области в этом случае будет заключена в предметной онтологии, формируемой посредством анализа текстовой документации.

В работах отечественных, а также зарубежных исследователей (Гаврилова Т.А. [6], Загорулько Ю.А. [7], Хорошевский В.Ф., Соловьев В.Д., Лукашевич Н.В., Добров Б.В., Смирнов С.В., Guarino [8], Uschold M. и др.) отмечается актуальность применения онтологического подхода к автоматической структуризации крупных текстовых архивов с использованием онтологического подхода и извлечения семантической основы проектной документации.

2. Модели прикладной онтологии текстовых документов электронного архива

Построение онтологии при классификации документов в электронных архивах необходимо для учета особенностей предметной области работы организации и повышения скорости поиска необходимых документов. Онтология задает семантическую шкалу, позволяющую определить набор документов к одному классу.

Таким образом, прикладную онтологию электронного архива проектной документации можно представить следующим образом:

$$O_{ARC} = \langle T, T_{ORG}, Rel, F \rangle,$$

где T представляет собой множество терминов проектной документации электронного архива; T_{ORG} – это множество терминов проблемной области организации; Rel – это множество отношений онтологии. Множество отношений включает следующие составляющие:

$$Rel = \{R_H, R_{PartOF}, R_{ASS}\},$$

где R_H- отношение иерархии; R_{PartOF}- отношение «часть-целое»; R_{ASS}- отношение ассоциации.

Формально множество терминов проектной документации электронного архива можно представить так:

$$T = (T^{D_1} \cup T^{D_2} \cup \ldots \cup T^{D_k}) \cup T^{ARC},$$

где $T^{D_i}, i = \overline{1,m}$ - представляет собой множество терминов *i-о*й проблемной области; T^{ARC} – это множество терминов проблемной области, полученных из документов электронного архива организации.

Формально функции интерпретации предметной онтологии представляются следующим образом:

$$F = \{F_{T_{ORG}T}, F_{T^{ARC}T^D}\},$$

где $F_{T_{ORG}T}: \{T_{ORG}\} \to \{T\}$ – функция интерпретации, определяющая соответствие между терминами проблемной области организации и терминами проектной документации электронного архива; $F_{T^{ARC}T^D}: \{T^{ARC}\} \to \{T^D\}$ - функция интерпретации, определяющая соответствие между терминами проблемной области, полученными из документов электронного архива организации и терминами проблемной области.

Определяющим в онтологии электронного архива является отношение «associate_with», которое определяет, какой предметной области принадлежит тот или иной проектных документ электронного архива, тем самым определяя тематику документа.

Характеристикой веса термина f_i в текстовом документе электронного архива является частота данного *i*-ого термина в конкретном документе. Отсюда актуальными являются следующие закономерности:

- термины, обладающие высокой частотой в конкретном документе, в большинстве своем являются общесистемными;
- термины, обладающие низкой частотой в конкретном документе, не обеспечивают повышение качества поиска документов в архиве;
- наиболее показательными являются термины, обладающие средней частотой встречаемости в документе, но наиболее полно характеризующие конкретный документ относительно рассматриваемой проблемной области [11, 12].

Если частота встречаемости одного термина значительно выше в документе, чем частота его встречаемости во всех анализируемых документах электронного архива, значит данный термин

является семантически значимым. Формально данное правило можно представить следующим образом:

$$s_i = tf_i \cdot log\left(\frac{M}{df(t_i)}\right),$$

где s_i это показатель семантической значимости термина t_i в данном документе; M – это общее число всех документов электронного архива; tf_i – это значение показателя нормализованной частота термина t_i; $df(t_i)$ - это общее число документов, содержащих термин t_i.

Таким образом, онтологическую модель документа электронного архива можно представить в следующем виде:

$$V_j^{doc} = \langle T^{ARC}, T^D \rangle,$$

где T^{ARC}, T^D- это множества терминов проблемной области j-ого документа электронного архива. Отсюда

$$associate\_with(d, T_K) = 1.$$

Данное равенство предполагает, что документ d отображается в пространство терминов $T, T_2, \dots, T_k$. Если t_i^d - это i-й термин документа d, тогда множество терминов документа d можно представить следующим образом:

$$T^d = \{t_1^d, t_2^d, \dots, t_n^d\},$$

где n – общее число терминов в документе d.

3. Модель онтологического индекса

Алгоритм онтологического индексирования текстовых документов электронного архива представлен на рисунке 1:

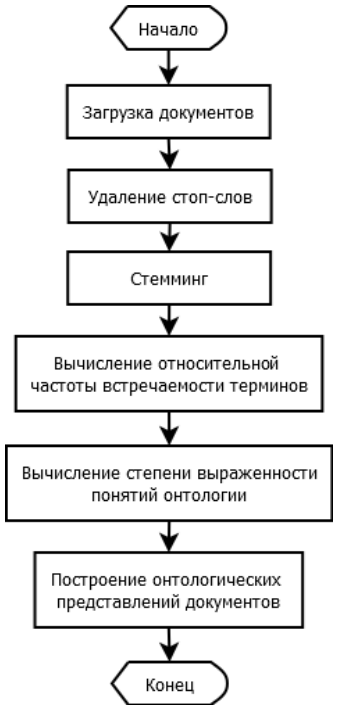

Рисунок 1. Алгоритм онтологического индексирования текстовых документов.

Степень семантической значимости термина онтологии электронного представляет собой значение уровня совпадения контекстного окружения термина с набором терминов документа электронного архива. При этом контекстное окружение составляют термины, семантически близкие к анализируемому понятию проблемной области [14].

Отсюда семантический индекс i-ого документа можно представить следующим образом:

$$\{(t_1^d, s_1), (t_2^d, s_2), \dots, (t_i^d, s_i), \dots, (t_n^d, s_n)\},$$

где l – общее число терминов в i-м документе электронного архива после предобработки текста.

Степень выраженности понятия l_k в i-м документе d будем вычислять по следующей формуле:

$$\mu(l_k) = 1 - \frac{1}{l}\sum|s_k - s_i|,$$

где s_k, s_i это значения показателей частоты термина t_i в описании k-ого термина предметной онтологии в документе d; n - это показатель мощности текстового входа понятия l_k.

Таким образом, после выполнения процесса индексирования документ d представляет собой онтологическое представление - фрагмент онтологии предметной области, в котором для каждого понятия онтологии определена степень выраженности от 0 до 1 (выражение 2).

4. Классификация онтологических представлений документов электронного архива

Для определения классов, по которым будет происходить разбиение документов электронного архива, необходимо определить множество понятий онтологии и лингвистическую метку для определения степени выраженности понятия в классе.

Фактически лингвистическая метка определяет смысловое представление для интервала степени выраженности понятия онтологии. Например, лингвистическая метка «Высоко» может соответствовать значению степени выраженности понятия из интервала от 0.7 до 1.0.

Таким образом, на первом шаге алгоритма классификации содержимого электронного архива необходимо задать множество классов G и определить их свойства:

$$G = \{g_1, g_2, \dots, g_i, \dots, g_n\},$$
$$g_i = \{\langle c_1, m\rangle, \dots, \langle c_k, m\rangle\},$$
$$m \in [High, Middle, Low],$$
$$High = [0.7 \dots 1.0],$$
$$Middle = [0.5 \dots 0.7),$$
$$Low = [0 \dots 0.5),$$

где g_n – n-й класс документов (основание классификации); c_k - k-е понятие онтологии; m – лингвистическая метка.

На втором шаге происходит вычисление степени принадлежности документа d каждому классу g_i с помощью следующего выражения:

$$s(g_i) = k - \sum_{i=1}^{k}(1 - \theta_k),$$

где k - количество параметров класса g_i; θ_k –признак соответствия документа d -у свойству класса g_i, которое вычисляется с помощью следующего выражения:

$$\theta_k = \begin{cases} 1, c_k \in d, \mu(c_k) \in m \\ 0 \end{cases}.$$

Таким образом, документ d соответствует признаку θ_k только в том случае, если содержит понятия, характеризующее данный признак и его степень выраженности входит в интервал лингвистической метки.

5. Результаты экспериментов

В рамках данного исследования был проведен ряд экспериментов по оценке качества классификации документов электронного архива ФНПЦ АО «Научно-производственное объединение «Марс» - это организация, осуществляющая проектирование, разработку и сопровождение автоматизированных систем, программных и технических средств для ВМФ РФ.

Для проведения экспериментов были выбраны следующие наборы документов:

- технические задания;
- отчеты о патентных исследованиях;
- спецификации;
- программы и методики тестирования;

- руководства программиста, пользователя, системного администратора и др.

Для проведения экспериментов было выбрано 1037 проектных документов. На рисунке 2 представлены признаки экспертного разбиения документов на классы по определенным признакам.

Рисунок 2. Экспертная классификация документов ФНПЦ АО «НПО Марс».

В рамках проведенных экспериментов были построены онтологический набор индексов документов и классические индексы, которые включают в себя значения «Термин-Частота». В качестве оценочной функции использовалась модель оценки качества классификации из [15]. Результаты проведенных экспериментов представлены на рисунках 3 и 4.

Рисунок 3. Сравнение алгоритмов классификации в соответствие со значениями оценочной функции.

Рисунок 4. Сравнение алгоритмов классификации в соответствие со значениями времени классификации (сек.).

Как видно из результатов проведенных экспериментов, процесс классификации онтологических представлений проходит быстрее (до 27 раз) относительно времени классификации классических индексов. Качество классификации онтологических

представлений по сравнению с результатами классификации классических индексов незначительно хуже только при разбиении по классу документации, при разбиении множества документов по виду документации, разделу документации и тематике работ качество классификации онтологических представлений выше, чем у классических индексов.

8. Заключение

Таким образом, в работе предложена онтологическая модель текстового документа в качестве ресурса электронного архива и онтологически ориентированный алгоритм классификации технических документов. Как видно из результатов экспериментов, формирование онтологического представления каждого отдельного документа в архиве позволяет значительно повысить скорость автоматической классификации документов (до 27 раз) при сохранении или незначительном улучшении качества классификации.

В будущих работах планируется введение элементов нечеткости в онтологическое представление проектных документов.

9. Благодарности

Исследование выполнено в рамках ПНИ по теме «Разработка архитектуры, методов и моделей построения сервера классификации больших слабоструктурированных данных на основе гибридизации семантико-онтологического анализа и машинного обучения» согласно Соглашению о предоставлении субсидий № 05.604.21.0252 в рамках ФЦП «Исследования и разработки по приоритетным направлениям развития научно-технологического комплекса на 2014-2020 годы».

10. Литература

[1] Wang, Y. Improvement of big data retrieval algorithm in the intelligent archives management / Y. Wang, L. Liu, Y. Qiu // 12th IEEE International Conference on Electronic Measurement & Instruments (ICEMI), 2015. – P. 487-491. DOI: 10.1109/ICEMI.2015.7494245.

[2] Pandolfo, L. Towards an Ontology for Describing Archival Resources / L. Pandolfo, L. Pulina, M. Zielinski // Proceedings of the Second Workshop on Humanities in the Semantic Web, 2017. – P. 111-116.

[3] Pandolfo, L. A framework for automatic population of ontology-based digital libraries / L. Pandolfo, L. Pulina, G. Adorni // Advances in Artificial Intelligence, 2016. – P. 406-417.

[4] Kruk, S.R. Semantic digital libraries / S.R. Kruk, B. McDaniel – Springer, 2009.

[5] Yan, Z. Semantic Search on Cross-Media Cultural Archives / Z. Yan, F. Scharffe, Y. Ding // Advances in Intelligent Web Mastering. Advances in Soft Computing. – 2007. – Vol. 43. – P. 375-380.

[6] Zagorulko, Yu.A. Semantic approach to the analysis of documents based on the ontology of the subject area / Yu.A. Zagorulko, I.S. Kononenko, E.A. Sidorova [Electronic resource]. – Access mode: http://www.dialog-21.ru/digests/dialog2006/materials/html/SidorovaE.htm.

[7] Gavrilova, T.A. Knowledge Base of Intelligent Systems / T.A. Gavrilova, V.F. Khoroshevsky – St. Petersburg: Peter, 2000.

[8] Schneider, T. Ontology for Big Systems: The Ontology Summit / T. Schneider, A. Hashemi, M. Bennett, M. Brady, C. Casanave, H. Graves, M. Grüninger, N. Guarino, A. Levenchuk, E. Lucier, L. Obrst, S. Ray, R. Sriram, A. Vizedom, M. West, T. Whetzel, P. Yim // Communiqué. Applied Ontology. – 2012. – Vol. 7. – P. 357-371. DOI: 10.3233/AO-2012-0111.

[9] Serrano-Guerrero, J. Physical and Semantic Relations to Build Ontologies for Representing Documents / J. Serrano-Guerrero, J.A. Olivas, J. de la Mata, P. Garces // Fuzzy logic, Soft Computing and Computational Intelligence (Eleventh International Fuzzy Systems Association World Congress IFSA) – Beijing, China. – 2005. – Vol. I. – P. 503-508.

[10] Zagoruyko, N.G. Applied methods of data and knowledge analysis – Novosibirsk: IM SB RAS, 1999.

[11] Yarushkina, N. Development of a knowledge base based on context analysis of external information resources / N. Yarushkina, V. Moshkin, A. Filippov // Proceedings of the

International conference Information Technology and Nanotechnology. Session Data Science – Samara, 2018. – P. 328-337.
[12] Namestnikov, A. An Ontology-Based Model of Technical Documentation Fuzzy Structuring / A. Namestnikov, A. Filippov, V. Avvakumova // 2nd International Workshop on Soft Computing Applications and Knowledge Discovery, 2016.
[13] Афанасьева, Т.В. Онтологический и нечеткий анализ слабоструктурированных информационных ресурсов / Т.В. Афанасьева, В.С. Мошкин, А.М. Наместников, И.А. Тимина, Н.Г. Ярушкина – Ульяновск : УлГТУ, 2016. – 130 с.
[14] Filippov, A. Approach to Translation of RDF/OWL-Ontology to the Graphic Knowledge Base of Intelligent Systems / A. Filippov, V. Moshkin, A. Namestnikov, G. Guskov, M. Samokhvalov / Proceedings of the II International Scientific and Practical Conference "Fuzzy Technologies in the Industry – FTI" – Ulyanovsk, 2018. – P. 44-49.
[15] Radionova, Yu.A. A method for constructing an evaluation function that determines the effectiveness of automatic clustering algorithms // Automation of control processes, 2009. – Vol. 15. – P. 23-28.

Ontology-based classification model of text resources of an electronic archive

A.A. Zarubin[1], A.R. Koval[1], V.S. Moshkin[2]

[1]The Bonch-Bruevich Saint - Petersburg State University of Telecommunication, Moika street 61, Saint-Petersburg, Russia, 191186
[2]Ulyanovsk State Technical University, Severny Venetz street 32, Ulyanovsk, Russia, 432027

Abstract. A modern design organization has a significant electronic archive of documents in an unstructured form. Solving the problem of using the experience of previous projects to solve new problems can be based on the use of intelligent methods and algorithms for analyzing text documents of an organization in order to build a classification system for electronic archives. This work presents an ontological model of a text document as an electronic archive resource. The paper also presents an ontologically-oriented classification algorithm for technical documents. In conclusion, the results of experiments confirming the effectiveness of models and algorithms in solving the problem of classifying a document archive are presented.