УДК 004.912
МРНТИ 20.19.27

ОБРАБОТКА НАУЧНЫХ РЕСУРСОВ ВУЗА В СИСТЕМАХ УПРАВЛЕНИЯ ЗНАНИЯМИ

Г. ЖОМАРТҚЫЗЫ, С.К. КУМАРГАЖАНОВА, Г.В. ПОПОВА

Восточно-Казахстанский государственный технический университет им. Д. Серикбаева

Аннотация: *В данной статье описан онтологически-ориентированный подход текстовой обработки информационных ресурсов вуза, связанных с научной деятельностью. В качестве информационной модели научных знаний используется онтология. Описаны информационная модель научных ресурсов вуза, методы классификации текстов по научным направлениям и тематического аннотирования текстов. Онтологически-ориентированный подход позволяет организовать, структурировать информационные ресурсы вуза, связанные с научной деятельностью, и разработать методы поиска знаний. В качестве языка описания онтологии используется язык OWL DL (англ. Web Ontology Language). При разработке онтологии для описания различных характеристик классов и свойств были составлены OWL аксиомы классов и отношений, установлены ограничения атрибутов. Для классификации научных ресурсов используется kNN-классификация. Задачей классификации в машинном обучении является отнесение объекта к одному из заранее определенных классов на основании его формализованных признаков. Метод kNN (метод k ближайших соседей) – модель векторной классификации. kNN-классификатор относит документ к преобладающему классу ближайших соседей. Параметр k в методе kNN устанавливается предварительно на основании знаний о решаемой задаче классификации. В данной работе метод kNN используется для задачи многозначной классификации. Классификация для классов, которые не являются взаимоисключающими, называется многозначной (англ. Multilabel Classification) классификацией. Классификация документа состоит из следующих действий: лингвистический анализ, извлечение терминов и формирование векторного пространства документа, вычисление k ближайших соседей, ранжирование классов. Для тематического аннотирования текстов используются классы онтологии предметной области. В онтологическом словаре термины сгруппированы по классам предметной области.*

Ключевые слова: *онтология, Text Mining, классификация текстов, метод kNN, лингвистический анализ*

UNIVERSITY'S SCIENTIFIC RESOURCES PROCESSING IN KNOWLEDGE MANAGEMENT SYSTEMS

Abstract*: This paper describes the ontologically-oriented approach of text processing of information resources of the university associated with scientific activities. An ontology is used as an information model of scientific knowledge. The information model of the scientific resources of the university, methods for the classification of texts in scientific fields and thematic annotation of texts are described. The ontologically-oriented approach allows you to organize, structure information resources of the university associated with scientific activities, and develop methods for finding knowledge. The general model of knowledge of the university is described with the help of ontology. OWL DL (Web Ontology Language) is used as the ontology description language. When developing an ontology for describing various characteristics of classes and properties, OWL axioms of classes and relations were compiled, and attributes were established. For the classification of scientific resources used kNN-classification. The task of classification in machine learning is the assignment of an object to one of the predefined classes on the basis of its formalized features. The kNN method (k nearest neighbors method) is a vector classification model. The kNN classifier refers the document to the prevailing class of nearest neighbors. The k parameter in the kNN method is preset based on knowledge of the classification task being solved. In this paper, the kNN method is used for the multivalued classification problem. Classification for classes that are*

not mutually exclusive, are called multi-valued (English Multilabel Classification) classification. Document classification consists of the following actions: linguistic analysis, extraction of terms and formation of the document vector space, calculation of k nearest neighbors, ranking of classes. For subject annotation of texts, domain ontology classes are used. In the ontological dictionary, terms are grouped by domain class.

Keywords: *ontology, Text Mining, text classification, Method kNN, linguistic analysis*

БІЛІМДІ БАСҚАРУ ЖҮЙЕЛЕРІНДЕ ЖОҒАРЫ ОҚУ ОРНЫНЫҢ ҒЫЛЫМИ РЕСУРСТАРЫН ӨҢДЕУ

Аңдатпа: *Бұл мақалада университеттің ғылыми қызметімен байланысты ақпараттық ресурстарын мәтіндік өңдеудің онтологиялық бағдарланған тәсілі сипатталған. Онтология ғылыми білімнің ақпараттық моделі ретінде қолданылады. Мақалада жоғары оқу ғылыми ресурстарының ақпараттық моделі, ғылыми мәтіндерді классификациялау, сондай-ақ мәтіндердің тақырыптық аннотациялау әдістері қарастырылған. Онтологиялық бағдарлы тәсілдеме университеттің ғылыми қызметімен байланысты ақпараттық ресурстарын ұйымдастыруға, құрылымдауға және ақпаратты табу әдістерін жасауға мүмкіндік береді. OWL DL (Web Ontology Language) онтологияны сипаттау тілі ретінде пайдаланылады. Класстар мен қасиеттерді сипаттау үшін онтологияны әзірлеу кезінде сыныптар мен қатынастардың OWL аксиомалары құрастырылып, атрибуттар құрылды. Ғылыми ресурстарды классификациялау үшін kNN-классификациясы зерттелген. Машиналық оқытуда классификация міндеті –қасиеттер негізінде объектіні алдын ала анықталған класстардың біріне жатқызу. kNN әдісі (k жақын көршілер әдісі) – векторлық жіктеу моделі. kNN классификаторы құжатты жақын көршілердің басым класына жатқызады. kNN әдісіндегі k параметрі алдын ала анықталады. Осы жұмыста kNN әдісі көпмәнді классификациялау міндеті үшін пайдаланылады. Бір-біріне тәуелсіз болып табылмайтын сыныптарға арналған классификациялау көпмәнді классификациялау (Multilabel Classification) деп аталады. Құжатты классификациялау мынадай әрекеттерді қамтиды: лингвистикалық талдау, терминдерді шығару және құжаттың векторлық кеңістігін қалыптастыру, ең жақын k көршілерді есептеу, реттеу. Мәтіндерді тақырыптық аннотациялау үшін пәндік аймақ онтология класстары құрылған. Онтологиялық сөздікте терминдер домендік класс бойынша топтастырылған.*

Түйінді сөздер: *онтология, мәтінді өңдеу, мәтінді классификациялау, kNN әдісі, лингвистикалық талдау*

1. Введение

Научные знания вуза – важны как для инновационного развития вуза, так и для совершенствования образовательного процесса. Задача сохранения и передачи учёными вуза накопленных знаний выдвигается на передний план, потому исследования, связанные с формированием и пополнением базы научных знаний вуза является актуальными.

В данной работе используются методы Text Mining обработки текстовой информации для построения семантических профилей документов. Онтологический подход позволяет решить следующие задачи: поиск экспертов в интересующей области знаний, выявление основных научных приоритетов вуза, кафедр, преподавателей. Примерами использования семантических профилей документов является: построение профилей преподавателей с возможностью просмотра, например, основных публикаций, направление научных исследований. В данной работе были поставлены следующие основные задачи**:** извлечение терминологических коллокаций, классификация текстов по научным направлениям, тематическое аннотирование научных текстов.

2. Литературный обзор

Онтологический подход аннотирования документов для автоматического выявлению компетенций сотрудников рассматривается в работах [1, 2]. В статье [3] используются Онтолого-ориентированные гибридные мето-

ды TextMining для обработки текстовых заявок. В работе [4] описаны методы формирования профилей специалистов, а также предложены методы вычисления семантического сходства профилей с помощью увеличения активности сетей, полученных из онтологий.

Технология автоматической рубрикации текстов для автоматического концептуального индексирования описаны в литературе [5]. В данной работе описаны методы и проблемы автоматической рубрикации, схема описания рубрик, различные методы машинного обучения для рубрикации текстов. Проблемы классификации и кластеризации документов подробно описаны в работах [6, 7, 8, 9, 10].

3. Информационная модель научных ресурсов вуза

Онтологически-ориентировнный подход позволяет организовать, структурировать информационные ресурсы вуза, связанные с научной деятельностью, и разработать методы поиска знаний.

Общую модель знаний вуза можно описать как онтлогию $O = \{O_д, O_{ПрО}, S\}$, где $O_д$ – онтология научных деятельности, $O_{ПрО}$ – онтология предметной области (ПрО), S – словарь предметной области.

Онтология научной деятельности описывает основные понятия научной деятельности вуза, такие как организационная структура, субъекты, объекты научных школ и исследований, информационные ресурсы, разделы наук и пр. Словарь онтологии содержит термины по предметным областям научных направлений.

В качестве языка описания онтологии используется язык OWL DL (англ. *Web Ontology Language*) [11]. Разработанная иерархическая таксономия классов и установленные роли между ними охватывают основные элементы управления научной деятельностью вуза. Выразительные возможности DL (англ. *Description Logics*) обеспечивают возможность определения: кардинальности и иерархии ролей, а также обратных и транзитивные ролей.

При разработке онтолгии для описания различных характеристик классов и свойств были составлены OWL аксиомы классов и отношений, установлены ограничения атрибутов, т.е [12]:

A. Классы:

RC_U – исследовательские центры вуза

RC_D – исследовательские центры факультета

RC_I – межвузовские исследовательские центры

D(x)– факультеты

SM_i– сотрудники научных школ

O – организация

SS– научные школы

IR – информационные ресурсы (ИР)

IRSS– информационные ресурсы научных школ

SD– раздел науки

P – персоны

PSS – персоны научных школ

B. Роли:

isWorkOrganization – работает в организации

workSchool – в составе научной школы

personHasIR – имеет публикации

publHasDivis – относится к разделу науки

C. Аксиомы классов и ролей:

$$RC_U \equiv RC_D \cup RC_I$$
$$\mathrm{D(x)} \equiv \{x_1\} \cup \ldots \cup \{x_1\}$$
$$SM_i \equiv P \cap \exists isWorkOrganization.O \cap \exists WorkSchool.SS$$
$$SS_i \subseteq \mathrm{SS}$$
$$PSS_i \equiv P \cap \exists personHasIR.IR$$
$$IRSS_i \equiv \mathrm{IR} \cap \forall \mathrm{publHasDivis}.SD_i$$

4. Классификация текстов по научным направлениям

Для классификации научных ресурсов используется *kNN*- классификация. *Задача классификации* в машинном обучении – это задача отнесения объекта к одному из заранее определенных классов на основании его формализованных признаков.

Метод *kNN* (метод *k* ближайших соседей) – модель векторной классификации. *kNN* классификатор относит документ к преобладающему классу ближайших соседей, где *k* – параметр метода. Параметр *k* в методе kNN часто выбирается на основании опыта или знаний о решаемой задаче классификации.

В данной работе метод *kNN* используется для задачи многозначной классификации. Классификация для классов, которые не являются взаимоисключающими, называются многозначной (англ. *Multilabel Classification*) классификацией [13, 14]. В задаче многозначной классификации происходит обучение j разных классификаторов y_j. Каждый из классификаторов возвращает либо метку класса c_j, либо метку класса $\bar{c}_j$ т.е., $\bar{c}_j$, т.е. $\gamma_j(d) \epsilon \{c_j, \bar{c}_j\}$, где d – тестируемый документ. Документ может принадлежать нескольким классам одновременно, одному классу или не принадлежать ни одному классу. Классы условно не зависят друг от друга.

Решение задачи многозначной классификации с помощью линейных классификаторов можно описать следующим образом:

– строится классификатор для каждого класса, при этом обучающее множество состоит из набора документов, принадлежащих классу;

– имея тестовый документ, применяем к нему каждый классификатор по отдельности; решение одного классификатора не влияет на решение другого.

Каждый из документов d в этой задаче представляется в виде вектора, $v(d) = (w_1, w_2, \ldots, w_N)$ в N-мерном пространстве, каждое измерение в котором представляет собой описание одного из признаков документа [6, 7, 15]. Вес каждого признака (термина) вычисляется следующим образом:

$$w_{t,d} = tf_{t,d} \times log \frac{|D|}{df_d} \qquad (1)$$

где:

D – множество документов класса c_j.

$|D|$ – общее количество документов в корпусе.

$tf_{t,d}$ – частота термина в документе.

df_d – документная частота, представляет собой количество документов в коллекции, содержащих термин t.

Для вычисления k ближайших соседей в качестве основной меры близости выбрана косинусная мера [6, 8, 9].

$$cos(v(d'), v(d)) = \frac{\sum_{i=1}^{n} v_i(d') \times v_i(d)}{\sqrt{\sum_{i=1}^{n} (v_i(d'))^2} \times \sqrt{\sum_{i=1}^{n} (v_i(d))^2}}$$

где:

$v(d')$ – векторное пространство документа d' из обучающего множества документов.

$v(d)$ – векторное пространство тестируемого документа d.

В многозначном класссификаторе *kNN* ранги классов вычисляются следующим образом:

$$rank(c, d) = \sum_{d' = S_k} I_c(d') \times \cos(v(d'), v(d))$$

где:

S_k – множество k ближайших соседей документа d'.

$I_c(d') = 1$ тогда и только тогда, когда документ d' принадлежит классу с. $I_c(d') = 0$ – в противном случае.

Таким образом, приписанные классы документов ранжируются согласно формуле (1).

Схема работы многозначного классификатора представлена на рисунке 1.

Рис. 1 – Схема работы классификатора

Классификации научных ресурсов состоит из нескольких шагов:

– обучение классификатора (нормировка терминов векторного пространства документов);

– выбор очередного документа;

– классификация документа.

Классификация документа состоит из следующих действий: лингвистический анализ, извлечение терминов и формирование векторного пространства документа, вычисление k ближайших соседей, ранжирование классов.

5. Тематическое аннотирование научных текстов

Для тематического аннотирование текстов используются классы онтологии предметной области. В онтологическом словаре термины сгруппированы по классам предметной области (ПрО) [1, 2].

$$O = < K, T >, \quad (4)$$

где:

K – множество всех классов онтологии, с = $|C|$;

T – множество всех терминов онтологии.

Для определения релевантности документа d теме i осуществляется поиск всех терминов, определяется число вхождений каждого термина документа, вычисляется релевантность документа теме . Рассмотрим данный процесс подробней.

Вначале производится автоматическое аннотирование документа, т.е. поиск в нем всех терминов из T. В результате получается аннотация документа:

$$V^d = (v_1^d, \dots, v_j^d, \dots, v_{|T|}^d), \quad (5)$$

где:

v_j^d – число вхождений j-го термина в документ d.

Уровень релевантности R_i^d документа d теме вычисляется по следующей формуле:

$$R_i^d = \alpha \frac{\sum_{j=1}^{|J_i|} v_j^d}{|J_i|} + \beta \frac{|J_i^d|}{|J_i|}, \quad (6)$$

где:

a, β – коэфициенты, $a + \beta = 1$,

J_i – множество всех терминов по теме i;

J_i^d – множество терминов темы i в документе d, $J_i^d \subseteq J_i$.

Таким образом, вычисляя релевантность документа по каждой теме получается профиль документа d в виде вектора:

$$PD(d) = (R_{1,\ldots}^d R_c^d) \quad (7)$$

Завершающим шагом тематического аннотирования текстов является формирование семантического профиля документа посредством создания индивидов класса информационные ресурсы онтологии деятельности Од.

6. Заключение и перспективы последующих разработок

В работе описан онтологически-ориентировнный подход текстовой обработки информационных ресурсов вуза, связанных с научной деятельностью. Следующим этапом исследования является разработка информационно-поисковой системы извлечения смысловой информации из неструктурированных информационных ресурсов. Несмотря на актуальность проблемы, на сегодняшний день существует не так уж много систем семантического поиска. В существующих системах отсутствует функции классификации информационных ресурсов. Работа выполнялась в рамках гранта №0213РК00305 «Разработка онтологической базы знаний е-университета».

ЛИТЕРАТУРА

1. Ceci F., Pietrobon R., Gonçalves, A.L. (2012) Turning Text into Research Networks: Information Retrieval and Computational Ontologies in the Creation of Scientific Databases. PLoS ONE, Vol. 7(1), P. 1-9.
2. Kryukov K.V., Kuznetsov O.P., Suhoverov V.S. (2013). On the notion of a formal competency researchers. In Proceedings of III International Scientific and Technical Conference– OSTIS-2013, Russia, Moscow, P. 143-146.
3. Ma J., Xu W., Sun Y., Turban E., Wang Sh., and Liu O. (2012). An Ontology-Based Text-Mining Method to Cluster Proposals for Research Project Selection. IEEE Transactions on Systems, Man, and Cybernetics — Part A: Systems and Humans, Vol. 42(3), P. 784-790.
4. Thiagarajan R., Manjunath G., Stumptner M. (2008). Finding Experts By Semantic Matching of User Profiles. HP Laboratories, URL: http://www.hpl.hp.com/techreports/2008/HPL-2008-172.pdf
5. Lukashevich N.V. (2011). Thesauri in information retrieval tasks. Moscow University Publishing House, ISBN 978-5-211-05926-0, Russia, P. 415.
6. Manning Ch.D., Raghavan P., Schütze H. (2009). Introduction to Information Retrieval.
7. Bolshakov E.I., Klyshinsky E.S., Lande D.V., Noskov, A.A, Peskov O.V., Yagunova E.V. (2011). Automatic processing of natural language text and computational linguistics. MIEM Publishing House, ISBN 978–5–94506–294–8, Russia, P. 272.
8. Du M., Chen X. (2013). Accelerated k-nearest neighbors algorithm based on principal component analysis for text categorization. In Journal of Zhejiang University-Science C-Computers & Electronics, Vol. 14(6), P. 407-416.
9. Shengyi Jiang, Guansong Pang, Meiling Wu, Limin Kuang. (2012). An improved K-nearest-neighbor algorithm for text categorization. In Proceedings of the Expert Systems with Applications 39, pp: 1503–1509.
10. Jiang J., Tsai Sh., Lee Sh. (2012). FSKNN: Multi-label text categorization based on fuzzy similarity and k nearest neighbors. In Proceedings of the Expert Systems with Applications 39, P. 2813–2821.

11. Allemang D., Hendler J.(2011). Semantic Web for the Working Ontologist. ISBN-13: 978-0-12-373556-0, USA.
12. Guarino N. (2009). The Ontological Level: Revisiting 30 Years of Knowledge Representation. Conceptual Modeling: Foundations and Applications, P. 52-67.
13. Malarvizhi P., Ramachandra V.P. (2013). Multilabel classification of documents with MAPREDUCE. In International Journal of Engineering and Technology (IJET), ISSN : 0975-4024, P.1260-1267.
14. Cherman E.A., Monard M.C., Metz J. (2011). Multi-label Problem Transformation Methods: a Case Study. In Electronic Journal CLEI, Vol. 14(1), P. 4-13.
15. Liu Y., Loh Han T., Sun A. (2009). Imbalanced text classification: A term weighting approach. In Proceedings of the Expert Systems with Applications, Vol. 36, P. 690–701.