УДК 004.912
МРНТИ 20.19.27

МЕТОДЫ ВЫЯВЛЕНИЯ И ВЫБОРА ПРИЗНАКОВ ПРИ ОБРАБОТКЕ НАУЧНЫХ ИНФОРМАЦИОННЫХ РЕСУРСОВ ВУЗА

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

Восточно-Казахстанский государственный технический университет им. Д. Серикбаева

Аннотация: *В данной работе рассматриваются методы выявления и выбора признаков при обработке научных информационных ресурсов вуза. Процедура по обработки неструктурированных информационных ресурсов состоит из нескольких этапов: извлечение терминологических коллокаций, выбор признаков, классификация, тематическое аннотирование, кластеризация документов и аналитический информационный поиск. Методы автоматического извлечения терминологических коллокаций используются для формирования подмножества терминов предметной области. Множество терминологических коллокаций, выделяемое на заданной коллекции научных текстов, характеризует узкую предметную область этой коллекции. Автоматическое извлечение ключевых слов и терминологических коллокаций является основным этапом в задачах обработки естественного языка. Для автоматического извлечения терминологических коллокаций из научных текстов в данной работе рассматривается метод C-value. Установленное ограничение значения C-value позволит рассматривать только термины длиной более одного слова. Полученные таким образом термины-кандидаты формируют список n-грамм (биграммы, триграммы). Основная модификация метода, основанного на статистическом подходе, заключается в предварительном использовании морфологических шаблонов фильтров. Словосочетания, похожие на термины, извлекаются из текста с помощью метода C-value: проводится разделение текста; из текста извлекаются словосочетания, удовлетворяющие установленным условиям; для всех терминов-кандидатов, отобранных по установленному ограничению, создаются записи в базе данных. Методы выбора признаков применяются для сокращения размерности пространства признаков с целью формирования наиболее информативного состава. Выбор признаков способствует повышению эффективности обучения за счет уменьшения размера лексикона и точности классификации благодаря исключению шумовых признаков. Для удаления неинформативных терминов, т.е. для оценки важности терминов, выбран критерий χ2. Корпус документов для обработки собран из статей, опубликованных в журналах по различным направлениям.*

Ключевые слова: *метод C-Value, шаблоны терминов, N-граммная модель, триграммы, биграммы, критерий*

METHODS OF IDENTIFICATION AND SELECTION OF CHARACTERISTICS IN THE PROCESSING OF SCIENTIFIC INFORMATION RESOURCES OF THE UNIVERSITY

Abstract: *This paper discusses methods for identifying and selecting features when processing scientific information resources of a university. The procedure for processing unstructured information resources consists of several stages: the extraction of terminological collocations, the selection of features, classification, thematic annotation, clustering of documents and analytical information retrieval. Methods for automatic extraction of terminological collocations are used to form a subset of domain terms. The set of terminological collocations allocated on a given collection of scientific texts characterizes the narrow subject area of this collection. The automatic extraction of keywords and terminological collocations is the main stage in the tasks of processing natural language. For automatic extraction of terminological collocations from scientific texts in this paper the C-value method is considered. Setting a C-value value limit will only allow for terms longer than one word. The candidate terms thus obtained form a list of n-grams (bigrams, trigrams). The main modification of the method based on the statistical approach is the preliminary use of morphological*

filter patterns. Phrase-like phrases are extracted from the text using the C-value method: the text is divided; phrases that meet the established conditions are extracted from the text; for all candidate terms selected by the established restriction, records are created in the database. Methods of feature selection are used to reduce the dimension of feature space in order to form the most informative composition. The choice of traits contributes to improving the efficiency of learning by reducing the size of the lexicon and the accuracy of classification due to the elimination of noise signs. To remove non-informative terms, i.e. To assess the importance of the terms, the criterion was chosen. The body of documents for processing is assembled from articles published in journals in various fields.

Keywords: *C-Value method, term patterns, N-gram model, trigrams, bigrams, chi-squared test*

ЖОҒАРЫ ОҚУ ОРНЫНЫҢ ҒЫЛЫМИ АҚПАРАТТЫҚ РЕСУРСТАРЫН ӨҢДЕУДЕ ҚАСИЕТТЕРДІ БӨЛІП АЛУ ЖӘНЕ ТАҢДАУ ӘДІСТЕРІ

Аңдатпа: *Бұл мақалада жоғары оқу орнының ғылыми ақпараттық ресурстарын өңдеуде қасиеттерді бөліп алу және таңдау әдістері қарастырылған. Құрылымдық емес ақпараттық ресурстарды өңдеу процедурасы бірнеше кезеңдерден тұрады: терминологиялық коллокацияларды бөліп алу, қасиеттерді іріктеу, классификация, тақырыптық талдау, құжаттарды кластерлеу және ақпаратты аналитикалық іздеу. Терминологиялық жинақтарды автоматты түрде бөліп алу әдістері пәндік аймақ терминдерінің жиынын қалыптастыру үшін пайдаланылады. Берілген ғылыми мәтіндер жинағына бөлінген терминологиялық коллокациялар жиынтығы осы коллекциялар бойынша белгілі бір пәндік аймақты ғана қамтиды. Кілттік сөздерді автоматты түрде алу және терминологиялық коллокацияларды іріктеу табиғи тілдерді өңдеу мәселесінде негізгі кезеңі болып есептеледі. Осы мақалада ғылыми мәтіндерден терминологиялық коллокацияларды автоматты түрде алу үшін C-value әдісі ұсынылады. C-value шамасын орнату тек бір сөзден көп терминдерді қарастыруға мүмкіндік береді. Осылайша, термин-кандидаттар n-граммдардың тізімін құрастырады (bigrams, trigrams). Статистикалық тәсілге негізделген әдістегі негізгі модификациясы – фильтрлердің морфологиялық шаблондары. Терминдер тәрізді сөз тіркестері сөйлемдерден C-value әдісінің негізінде шығарылады: мәтін бөлінеді; белгіленген шарттарға сай сөз тіркесі мәтіннен алынады; белгіленген шектеулермен таңдалатын барлық термин-кандидаттар үшін жазбалар дерекқорда сақталады. Қасиеттерді таңдау әдістері қасиеттер кеңістігін азайту және сапалы қасиеттер құрамын қалыптастыру мақсатында пайдаланылады. Қасиеттерді таңдау лексикон көлемін азайту негізінде оқытудың тиімділігін арттыруға және шулы қасиеттерді жоюға байланысты классификация дәлдігін қамтамасыз ету арқылы ықпал етеді. Ақпаратсыз терминдерді шығару үшін, яғни, терминдердің маңыздылығын бағалау үшін χ^2 критерийі қолданылған. Өңдеуге арналған құжаттар жинағы әртүрлі саладағы журналдарда жарияланған мақалалардан жиналған.*

Түйінді сөздер: *C-Value әдісі, терминдер үлгілері, N-грамм үлгісі, триграммалар, биграммалар, χ2 критерийі*

1. Введение

В данной работе рассматриваются методы выявления и выбора признаков при обработке неструктурированных информационных ресурсов вуза при формирования онтологического словаря по научным направлениям для классификации документов. Методы автоматического извлечения терминологических коллокаций используются для формирования подмножества терминов предметной области [1, 2]. В онтологии предметной области используются отношения вида тема-подтема.

2. Онтологически-ориентировнный подход текстовой обработки информационных ресурсов вуза

Процедуру обработки информационных ресурсов вуза с целью формирования научных школ и научных направлений вуза иллюстрирует рисунок 1. Ниже приведены основные этапы обработки информационных ресурсов:

1. Извлечение терминологических коллокаций. В качестве метода для выявления коллокации используется метод C-value [2, 3, 4].

2. Выбор признаков. В качестве метода для оценки важности терминов, выбран критерий [5].

3. Классификация текстов по научным направлениям. Для классификации текстов используется метод k ближайших соседей (kNN) [5, 6, 7, 8].

4. Тематическое аннотирование научных текстов.

В данной работе описывается начальный этап – извлечение терминов процедуры обработки научных информационных ресурсов вуза

Рис. 1 – Этапы обработки научных информационных ресурсов вуза

3. Автоматическое извлечение терминологических коллокаций научных текстов

Коллокации понимаются как неслучайное сочетание двух и более лексических единиц, характерных для большинства научных текстов. Множество терминологических коллокаций, выделяемое на заданной коллекции научных текстов, характеризует узкую предметную область (темы и подтемы) этой коллекции.

Для автоматического извлечения терминологических коллокаций из научных текстов были изучены методы, такие как ме-

тод взаимной информации MI (англ. Mutual Information) и метод C-value [1, 2, 3]. В данной работе рассматривается Метод C-value.

Метод C-value ранжирует термины на основе их частоты и вложенности:

$$C-value(a)=\begin{cases}\log_2|a|\times f(a), \text{если строка } a \text{ не вложена в другие подстроки}\\ \log_2|a|\times\left(f(a)-\frac{1}{P(T_a)}\times\sum_{b\in T_a}f(Ta)\right), \text{в противном случае;}\end{cases} \quad (1)$$

где:

a – кандидат в термины;

$|a|$ – длина словосочетания a;

$f(a)$ – частота словосочетания a;

T_a – множество словосочетаний, содержащих a;

$P(T_a)$ – общее количество более длинных словосочетаний, содержащих a.

$f(T_a)$ – суммарная частота $P(T_a)$.

Основная модификация метода, основанного на статистическом подходе, заключается в предварительном использовании морфологических шаблонов фильтров, следующего вида (используются следующие сокращения: сущ. – существительное, прил. – прилагательное, р.п. – родительный падеж) [3]:

[сущ.+прил.(р.п.)+сущ.(р.п.)]
[прил.+прил.+сущ.]
[прил.+сущ.+сущ.(р.п.)]
[сущ.+сущ.(р.п.)+сущ.(р.п.)]
[прил.+сущ.]
[прич.+сущ.]
[сущ.+сущ.(р.п.)]
[сущ.+сущ.)]

Для получения списка доминантных терминов документа необходимо решить следующие задачи:

– получение списка всех терминов, употребляющихся в документе;

– выделение из этого списка терминов, которые являются доминантными для данного документа.

Словосочетания, похожие на термины, извлекаются из текста с помощью метода C-value:

– проводится разделение текста на строки, учитывается пунктуация, рассматриваются любые последовательности слов в тексте, не разделённые знаками препинания;

– из текста извлекаются словосочетания, удовлетворяющие следующим условиям:

- проводится морфологический анализ текста, для каждого слова определяется часть речи и морфологические характеристики, каждое слово приводится в нормальную форму, максимальная последовательность слов – триграммы;
- отбираются только те словосочетания, которые удовлетворяют шаблонам, удаляются стоп-слова, не несущие самостоятельной смысловой нагрузки;

– для всех терминов-кандидатов, для которых значение C-value больше 1, создаются записи в базе данных.

Ограничение значения C-value позволит рассматривать только термины длиной более одного слова, т. к. для термина длиной в одно слово значение C-value всегда равно нулю.

Полученные таким образом термины-кандидаты формируют список n-грамм (биграммы, триграммы). Корпус документов для обработки собран из статей, опубликованных в журнале «Физика твёрдого тела» по различным направлениям, учредителями которого являются: Российская Академия Наук, отделение Общей Физики и Астрономии РАН, физико-технический институт им. А.Ф.Иоффе РАН [9].

Для каждого термина-кандидата вычислено значение меры C-value по формуле (1). В таблице 1 представлены результаты работы разработанного модуля по извлечению терминов для раздела науки «Физика твёрдого тела», входящего в состав созданной онтологии предметной области.

4. Выбор признаков для классификации

Методы выбора признаков применяются для сокращения размерности пространства признаков с целью формирования наиболее информативного пространства [5, 10]. Выбор признаков способствует повышению:

– эффективности обучения за счет уменьшения размера лексикона;

– точности классификации благодаря исключению шумовых признаков.

Для каждого класса вычисляется мера полезности каждого термина из лексикона и выбирается терминов, имеющих наибольшее значение. Все другие термины отбрасываются и в классификации не участвуют. В данной работе были рассмотрены три меры полезности: взаимная информация, критерий, информационная выгода (англ. Expected Mutual Information).

Таблица 1 – Терминологические n-граммы (биграммы, триграммы) для области знаний физики твёрдого тела

Триграммы		Биграммы	
Термины	**C-value**	**Термины**	**C-value**
парциальный давление кислород	42,79	пластический деформация	162,50
междоузельный атом кремний	38,04	твердый тело	139,00
собственный междоузельный атом	33,28	размер зерно	107,67
образование вакансионный микропора	33,28	точечный дефект	76,31
собственный точечный дефект	31,70	комнатный температура	73,00
коэффициент деформационный упрочнение	23,77	граница зерно	70,60
амплитуда колебательный деформация	23,77	атом кислород	59,56
бездислокационный монокристалл кремний	23,77	пластический течение	57,60
скорость свободный поверхность	22,19	вектор бюргерс	56,09
бинарный твердый раствор	22,19	плотность дислокация	54,20
динамик точечный дефект	22,19	модуль юнга	52,83
поперечный размер кристалл	22,19	скорость деформация	49,83
степень пластический деформация	19,02	деформационный упрочнение	48,56
упругий планарный мезодефект	19,02	дисклинационный диполь	40,40
аморфный ковалентный материал	17,43	кристаллический решетка	40,25
фактор размерный несоответствие	17,43	междоузельный атом	32,00
скорость пластический деформация	15,85	межатомный связь	31,25

Источник: собственная разработка с использованием корпуса документов журнала «Физика твёрдого тела».

Для удаления неинформативных терминов, т.е. для оценки важности терминов, выбран критерий x_2. Критерий x_2 используется для проверки независимости двух случайных событий, где события считаются независимыми. При выборе признаков двумя событиями являются появление термина и появление класса.

Экспериментальная оценка метода показала, что он извлекает ключевые термины с высокой точностью и полнотой [55]. Онтологический словарь формируется с помощью этих терминов, и используются затем для классификации тематического аннотирования текста.

5. Заключение и перспективы последующих разработок

В работе предлагаются методы выявления и выбора признаков при обработке неструктурированных информационных ресурсов вуза для классификации научных электронных документов. Формирование словаря осуществляется с помощью методов автоматического извлечения и выбора списка ключевых терминологических коллокаций из корпуса научных статей. Онтологический словарь предметной области собирается в иерархическую структуру тем научных областей. Седеющим этапом исследования

является классификация научных текстов по научных направлениям и описание информационной модели научных ресурсов вуза.

Работа выполнялась в рамках гранта №0213РК00305 «Разработка онтологической базы знаний е-университета».

ЛИТЕРАТУРА

1. Pivovarova L.M.,Yagunova E.V. (2010). Extraction and classification of terminological collocations on the material of linguistic scientific texts (preliminary observations). In Proceedings of Symposium: "Terminology and knowledge" Russia, Moscow, url: http://webground.su/data/lit/pivovarova_yagunova/Izvlechenie_i_klassifikatsiya_terminoligicheskih_kollokatsyi.pdf.
2. Sedova Y.A., Kvyatkovskaya I.Y. (2011). Intelligent analysis of corps of scientific information. In Bulletin of the Astrakhan State Technical University. Series: Management, Computing and Informatics, Vol. 1, Russia, P. 128-136.
3. Braslavsky P. Sokolov, E. A (2008). Comparison of five methods for extraction of terms of arbitrary length. In Proceedings of International Conference "Dialogue" - Computational Linguistics and Intelligent Technologies, Vol. 7(14). Russia, P. 67-74.
4. Min J., Josh C.D., Buzhou T., Hongxin C., Hua X. (2012). Extracting semantic lexicons from discharge summaries using machine learning and the C-Value method. Procedding of the AMIA Symposium, P. 409-416.
5. Manning Ch.D., Raghavan P., Schütze H. (2009). Introduction to Information Retrieval.
6. 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.
7. 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 AP.lications 39, P.: 1503–1509.
8. 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 AP.lications 39, P. 2813–2821.
9. Science journal "Solid State Physics", url: http://journals.ioffe.ru/ftt/ (date accessed - September 2013).
10. Altınçay H., Erenel Z. (2010). Analytical evaluation of term weighting schemes for text categorization. In Proceedings of the Pattern Recognition Letters, 1, P. 1310–1323.