НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
январь–февраль 2017 Том 17 № 1 ISSN 2226-1494 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
January–February 2017 Vol. 17 No 1 ISSN 2226-1494 http://ntv.ifmo.ru/en

УДК 20.19.27

МЕТОД КОНТРАСТНОГО ИЗВЛЕЧЕНИЯ РЕДКИХ ТЕРМИНОВ ИЗ ТЕКСТОВ НА ЕСТЕСТВЕННОМ ЯЗЫКЕ

И.А. Бессмертный[a], А.Б. Нугуманова[b], М.Е. Мансурова[c], Е.М. Байбурин[b]

[a] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация
[b] Восточно-Казахстанский государственный университет им. С. Аманжолова, Усть-Каменогорск, 070004, Казахстан
[c] Казахский национальный университет имени аль-Фараби, Алматы, 050040, Казахстан
Адрес для переписки: bia@cs.ifmo.ru

Информация о статье
Поступила в редакцию 27.10.16, принята к печати 20.12.16
doi: 10.17586/2226-1494-2017-17-1-81-91
Язык статьи – русский

Ссылка для цитирования: Бессмертный И.А., Нугуманова А.Б., Мансурова М.Е., Байбурин Е.М. Метод контрастного извлечения редких терминов из текстов на естественном языке // Научно-технический вестник информационных технологий, механики и оптики. 2017. Т. 17. № 1. С.81–91. doi: 10.17586/2226-1494-2017-17-1-81-91

Аннотация

Рассмотрена проблема автоматического извлечения терминов предметной области из корпуса документов с привлечением контрастной коллекции. Существующие контрастные методы хорошо справляются с часто используемыми терминами, но работают плохо с редкими терминами, что приводит к обеднению словаря. Среди известных статистических методов оценка точечной взаимной информации хорошо выявляет редкие термины, однако при этом извлекается большое число слов, не относящихся к терминам. Для извлечения редких терминов в работе предложен подход, состоящий в использовании точечной взаимной информации с последующей фильтрацией кандидатов в термины по критерию совместной встречаемости с другими терминами-кандидатами. Для устранения шумов и выявления сильных связей формируется матрица «документы-на-термины», которая подвергается сингулярному разложению. После этого осуществляется переход к матрице «термины-на-термины», отражающей силу связей между словами. Предлагаемый подход апробирован на коллекции документов предметной области «Геология». В качестве контрастной коллекции использованы публикации из разделов «Политика», «Культура», «Экономика» и «Происшествия» на новостных Интернет-сайтах. Результаты эксперимента продемонстрировали работоспособность метода для успешного извлечения редких терминов.

Ключевые слова

контрастное извлечение терминов, терминологичность, взаимная информация, семантические связи, извлечение редких терминов

Благодарности

Работа содержит материалы исследований, частично поддержанных грантом Минобрнауки Республики Казахстан 5033/ГФ4 «Разработка интеллектуальной высокопроизводительной информационно-аналитической поисковой системы обработки слабоструктурированных данных».

METHOD OF RARE TERM CONTRASTIVE EXTRACTION FROM NATURAL LANGUAGE TEXTS

I.A. Bessmertny[a], A.B. Nugumanova[b], M.Ye. Mansurova[c], Ye.M. Baiburin[b]

[a] ITMO University, Saint Petersburg, 197101, Russian Federation
[b] S. Amanzholov East Kazakhstan State University, Ust Kamenogorsk, 070004, Republic of Kazakhstan
[c] Al-Farabi Kazakh National University, Almaty, 050040, Republic of Kazakhstan
Corresponding author: igor_bessmertny@gmail.com, bia@cs.ifmo.ru

Article info
Received 27.10.16, accepted 20.12.16
doi: 10.17586/2226-1494-2017-17-1-81-91
Article in Russian

For citation: Bessmertny I.A., Nugumanova A.B., Mansurova M.Ye., Baiburin Ye.M. Method of rare term contrastive extraction from natural language texts. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2017, vol. 17, no. 1, pp. 81–91. doi: 10.17586/2226-1494-2017-17-1-81-91

Abstract

The paper considers a problem of automatic domain term extraction from documents corpus by means of a contrast collection. Existing contrastive methods successfully extract often used terms but mishandle rare terms. This could yield poorness of the resulting thesaurus. Assessment of point-wise mutual information is one of the known statistical methods of

term extraction and it finds rare terms successfully. Although, it extracts many false terms at that. The proposed approach consists of point-wise mutual information application for rare terms extraction and filtering of candidates by criterion of joint occurrence with the other candidates. We build “documents-by-terms” matrix that is subjected to singular value decomposition to eliminate noise and reveal strong interconnections. Then we pass on to the resulting matrix “terms-by-terms” that reproduces strength of interconnections between words. This approach was approved on a documents collection from “Geology” domain with the use of contrast documents from such topics as “Politics”, “Culture”, “Economics” and “Accidents” on some Internet resources. The experimental results demonstrate operability of this method for rare terms extraction.

Keywords
contrastive term extraction, termhood, mutual information, semantic connections, rare term extraction

Acknowledgements
The paper contains data for study partially financially supported by the Grant 5033/ГФ4 of the Ministry of Education and Science of the Republic of Kazakhstan "The development of intelligent high-performance information and analysis search engine for semistructured data processing"

Введение

Целью настоящей работы является исследование возможности автоматического извлечения редких терминов из текстов предметной области. Редкие термины, имеющие низкую частоту распределения в текстах предметной области, зачастую бывают не менее важны для описания предметной области, чем высокочастотные (ключевые) термины. В некоторых специальных коллекциях документов, например, в описаниях побочных действий лекарственных препаратов, доля редких терминов составляет более 2/3 от общего числа терминов [1], и игнорирование таких слов при формировании словаря предметной области приводит к его существенному обеднению. По этой причине задача извлечения редких терминов имеет высокую актуальность, но при этом в литературе ей уделено недостаточно внимания.

Среди большого количества статистических методов извлечения терминов (см. например, обзор работ в [2, 3]) мы можем назвать только один критерий, позволяющий извлекать редкие термины, это критерий, основанный на вычислении точечной взаимной информации (PMI, point-wise mutual information). Такие хорошо известные статистические критерии извлечения ключевых слов, как хи-квадрат, информационная выгода, мера TF-IDF (term frequency – inverse document frequency), успешно справляются с извлечением из текстов предметной области высокочастотных терминов [4, 5], но при этом не предназначены для извлечения редких терминов, так как частоты последних существенно ниже пороговых значений этих критериев.

Авторы работы [5] сравнивали перечисленные критерии (хи-квадрат и информационную выгоду) с PMI с позиции полезности при индексировании документов, и сделали вывод в не в пользу последней. В качестве главного недостатка они указали, что PMI смещает оценку терминологичности в сторону редких слов. На наш взгляд, благодаря такому смещению PMI, продемонстрировавшая низкую результативность при индексировании документов, может оказаться полезной для повышения терминологического охвата тех предметных областей, которые описываются не только высокочастотными терминами, но и редкими, специфическими словами. Другими словами, мы предполагаем, что, используя данный критерий, можно повысить полноту извлечения терминов, но при этом остается открытым вопрос точности такого извлечения.

Дело в том, что, как и любой статистический критерий оценки терминологичности, критерий PMI допускает большое количество ошибок первого рода, т.е. ложных обнаружений. Причина ошибок кроется в наличии так называемых сопряженных слов, т.е. слов, которые «сопровождают» предметную область, но не связаны с ней напрямую и не являются ее терминами. Эта проблема детально рассматривается в работе [6]. В качестве примера, иллюстрирующего проблему, авторы приводят слова «залив» и «Кувейт», которые тесно связаны с предметной областью «Нефть» из-за их частой встречаемости в текстах этой области, хотя на самом деле, эти слова относятся, скорее, к географической терминологии, чем к нефтяной. В связи с этим в фокусе настоящей работы находится снижение количества ложных обнаружений при автоматическом извлечении редких терминов на основе критерия PMI.

Дальнейшее изложение проделанной работы состоит из следующих разделов. Раздел «Связанные работы» содержит обзор родственных работ, посвященных статистическим методам извлечения терминов. В разделе «Постановка задачи» излагается суть проблемы и указываются общие шаги ее решения. В разделе «Предлагаемый подход» указанные шаги уточняются и конкретизируются в виде нового способа извлечения редких терминов, позволяющего исправить существующие недостатки критерия точечной взаимной информации. В разделе «Экспериментальная реализация предлагаемого подхода» описываются эксперименты по извлечению терминов с помощью предложенного способа. В заключении формулируются выводы и приводится план будущих исследований.

Связанные работы

Подробный обзор работ, связанных с проблемой извлечения терминов из текстов предметной области, можно найти, например, в работе [2]. В данном исследовании нас интересуют контрастные методы

извлечения терминов, суть которых заключается в выявлении терминов на основе сравнения их распределений в двух коллекциях: целевой (предметной) и альтернативной. В качестве альтернативной коллекции может использоваться либо контрастная коллекция, т.е. сформированная из текстов другой предметной области, либо общая коллекция, т.е. сформированная из текстов, не относящихся ни к какой предметной области [7].

В числе одной из первых работ, посвященных контрастному извлечению терминов, можно назвать [8]. Ее авторы для оценки терминологичности вводят в обиход новую, интуитивно понятную меру, получившую название «странность» (weirdness). Странность вычисляется для каждого термина-кандидата и представляет собой отношение частоты его употребления в целевой коллекции к частоте употребления в общей коллекции. Поскольку коллекции чаще всего не сбалансированы по размеру, то используются относительные частоты. Для обычных слов формула странности возвращает значения, близкие к 1, а для терминов – значения, намного превышающие 1, так как в этом случае знаменатель формулы близок к 0:

$$Weirdness = \frac{F_{SL}/N_{SL}}{F_{GL}/N_{SL}} = \frac{F_{SL} \cdot N_{GL}}{F_{GL} \cdot N_{SL}}, \quad (1)$$

где F_{SL}, F_{GL} – это частоты употреблений слова в целевой и общей коллекциях соответственно; N_{SL}, N_{GL} – это количества всех слов в целевой и общей коллекциях соответственно.

В своих более поздних работах, например, в [9], авторы представляют модифицированный вариант формулы (1), так как исходная формула, по их словам, проявляет сингулярность, когда знаменатель обращается в 0. Это происходит в тех случаях, когда частота употребления слова в общей коллекции равна 0, что в результате приводит к бесконечности. Модифицированная, сглаженная формула странности отличается от исходной прибавлением 1 к частоте употребления слова в общей коллекции текстов:

$$Weirdness = \frac{F_{SL} \cdot N_{GL}}{(1+F_{GL}) \cdot N_{SL}}. \quad (2)$$

Дальнейшее сравнение работ [8] и [9] показывает, что в первой работе авторы формируют список терминов только на основе высоких значений странности, а во второй они уже используют комбинацию высокой странности с высокой частотой (см. формулу (2)). Тем самым они пытаются избавиться от «странных» слов, случайно оказавшихся в целевой коллекции, т.е. не относящихся к предметной области. Такой подход гарантирует более высокую точность охвата терминов, но при этом, как мы уже отмечали, страдает полнота, поскольку редкие термины выпадают из рассмотрения.

В [10] идея контрастной оценки терминов формируется в виде не одного, а двух утверждений. Во-первых, реже употребляемые в целевой коллекции слова должны иметь более низкую оценку. Во-вторых, чаще употребляемые в целевой коллекции слова должны иметь более высокую оценку, но с оговоркой, что они не встречаются часто в контрастной коллекции или в ограниченном наборе текстов целевой коллекции. Авторы операционализируют эти утверждения в виде метрики, которую называют релевантностью (relevance):

$$Relevance = \frac{1}{\log_2\left(2+\frac{f_{SL} \cdot N_{SL}^t}{f_{GL}}\right)}, \quad (3)$$

где f_{SL} и f_{GL} – это относительные частоты употреблений слова t в целевой и контрастной коллекциях соответственно; N_{SL}^t – относительное число текстов целевой коллекции, в которых встречается данное слово. Приведенная формула (3) хорошо справляется с извлечением репрезентативных терминов, но искусственно занижает оценку редких терминов, что, как мы уже отмечали, негативно влияет на полноту охвата терминов.

В [11] предлагается несколько иной способ оценки терминологичности на базе контрастного подхода. Способ берет за основу известную формулу взвешивания слов TF-IDF, согласно которой вес слова в документе тем выше, чем выше частота его использования в этом документе и чем ниже его разброс по всей коллекции. В новом варианте формулы (см. формулу (4)), который авторы называют «term frequency – inverse domain frequency», оценивается вес слова не в документе, а в целевой коллекции. Согласно новой формуле вес слова тем выше, чем выше относительная частота его использования в целевой коллекции и чем ниже его относительный разброс по всем коллекциям:

$$TF \cdot IDF = TF(t, D) \cdot IDF(t) = \frac{n_{t,D}}{\sum_k n_{k,D}} \cdot \log\left(\frac{|TS|}{|\{d: t \in d\}|}\right), \quad (4)$$

где $n_{t,D}$ – это число вхождений слова t в целевую коллекцию D; $\sum_k n_{k,D}$ – это сумма вхождений всех слов в целевую коллекцию D; $|TS|$ – это количество документов во всех используемых коллекциях; $|\{d: t \in d\}|$ – это количество всех документов, в которые слово t входит хотя бы один раз. Таким образом, авторы считают терминами все слова с высокой концентрацией в пределах узкого подмножества документов. Для определенной части терминов это, безусловно, справедливый подход, но для редких терминов он малопригоден.

Авторы [12] также предлагают оценивать терминологичность слов на базе формулы TF-IDF. Собственный вариант этой формулы они называют контрастным весом (contrastive weight) и определяют его как меру, которая тем выше, чем выше частота употребления слова в целевой коллекции и чем ниже относительная частота его употребления в контрастных коллекциях:

$$Contrastive\ Weight = TF(t,D) \cdot IDF(t) = \log\left(f_t^{\ D}\right) \cdot \log\left(\frac{F_{TC}}{\sum_j f_t^j}\right), \quad (5)$$

где $f_t^{\ D}$ – частота употребления слова в целевой коллекции; $\sum_j f_t^{\ j}$ – сумма частот всех употреблений слова в контрастных коллекциях; $F_{TC} = \sum_{i,j} f_i^{\ j}$ – сумма частот употреблений всех слов во всех коллекциях, включая целевую. Как отмечают сами авторы, контрастный вес значительно лучше оценивает терминологичность слов, чем чистые частоты, однако общая эффективность метода, определенная с помощью F-меры, по их словам, не бросается в глаза.

В [13] формула (5) подвергается критической оценке. Как отмечают авторы указанной работы, контрастный вес и подобные ему метрики на самом деле оценивают не принадлежность терминов предметной области, а их распространенность. Чтобы исправить указанный недостаток, авторы предлагают оценивать терминологичность на базе двух показателей: меры преобладания слова в целевой коллекции DP (domain prevalence) и меры тяготения слова к целевой коллекции DT (domain tendency). Высокое значение DP указывает на преобладание слова в целевой коллекции по сравнению с другими словами. Высокое значение DT указывает на преобладание слова в целевой коллекции по сравнению с контрастной коллекцией.

Формула для расчета DP по сути является сглаженным вариантом формулы контрастного веса (5):

$$DP(t) = \log_{10}\left(f_t^{\ D} + 10\right) \cdot \log_{10}\left(\frac{F_{TC}}{f_t^{\ D} + f_t^{\ \bar{D}}} + 10\right), \quad (6)$$

где $f_t^{\ D}$ и $f_t^{\ \bar{D}}$ – частоты употреблений данного слова в целевой и контрастной коллекциях соответственно; $F_{TC} = \sum_j f_j^{\ D} + \sum_j f_j^{\ \bar{D}}$– суммы частот употреблений всех терминов-кандидатов в целевой и контрастной коллекциях соответственно.

Формула для расчета DT является сглаженным вариантом формулы странности (1), т.е. штрафует слова, которые часто встречаются в контрастной коллекции:

$$DT(t) = \log_2\left(\frac{f_t^{\ D}+1}{f_t^{\ \bar{D}}+1} + 1\right). \quad (7)$$

Формулы (6) и (7) объединяются в один общий показатель, названный дискриминационным весом DW (discriminative weight). По мнению авторов, этот показатель обладает высокой дифференцирующей способностью:

$$DW(t) = DP(t) \cdot DT(t). \quad (8)$$

Следует отметить, что показатели DT и DP достаточно сильно коррелируют друг с другом. Например, в наших экспериментах значения корреляции этих показателей составили от 0,71 до 0,82. Чтобы понять природу корреляции, мы разделили все термины-кандидаты на 4 непересекающиеся группы в зависимости от значений DT и DP:

1. значения DT и DP ниже среднего;
2. значение DT ниже среднего, а значение DP не ниже среднего;
3. значения DT не ниже среднего, а значения DP ниже среднего;
4. значения DT и DP не ниже среднего.

И экспертные оценки, и оценки на основе формулы (8) показали один и тот же результат: терминами, за небольшим исключением, могут быть признаны только кандидаты групп 3 и 4, что соответствует высоким значениям показателя DT. Данный результат свидетельствует о высокой информативности показателя DT и об избыточности показателя DP.

Использование сразу нескольких показателей для оценки терминологичности отличает не только работу [13]. В [14] для этой цели используются сразу 3 показателя: мера пертинентности DR (domain pertinence), мера согласованности DC (domain consensus) и лексическая когезия LC (lexical cohesion), предназначенная для оценки когезии многословных терминов. В результате итоговая оценка терминологичности слова t в целевой коллекции Di складывается из линейной комбинации трех перечисленных мер:

$$w(t,Di) = \alpha \cdot DR + \beta \cdot DC + \gamma \cdot LC, \quad (9)$$

где α, β, γ – это калибровочные параметры, по умолчанию $\alpha = \beta = \gamma = 1/3$.

Мера пертинентности DR представляет собой меру странности, обобщенную на случай множества контрастных коллекций. Она определяется как отношение частоты слова в целевой коллекции к его наибольшей частоте во всех существующих контрастных коллекциях:

$$DR(t,Di) = \frac{freq(t,Di)}{\max_j(freq(t,Dj))}. \quad (10)$$

Мера согласованности DC позволяет учитывать распределение слов в отдельных документах. Она определяется через нормированные частоты ϕ_k встречаемости слова t в документах целевой коллекции Di и тем выше, чем равномерней распределено слово в этих документах:

$$DC(t,Di) = -\sum_{d_k \in Di} \phi_k \log \phi_k. \quad (11)$$

Вводя меру согласованности, авторы обосновывают ее значимость тем, что термины, которые часто встречаются в большом количестве документов целевой коллекции, должны оцениваться выше, чем термины, которые часто встречаются в ограниченном количестве документов. Данное утверждение является полностью антагонистическим эвристике, использованной в [11]. Этот интересный факт, иллюстрирующий существующую противоречивость в выборе критериев терминологичности, отмечают также авторы [2]. Таким образом, формула (9) представляет собой компромисс в оценке терминологичности на основе странности (формула (10)) и согласованности (формула (11)).

Последняя работа, которую мы хотим отметить в этом ряду, – это [15]. Она также развивает идею штрафов и вознаграждений, заложенную в базовой конструкции формулы TF-IDF, и предлагает новый вариант этой формулы, получивший название «term frequency – disjoint corpora frequency» (DCF). В качестве вознаграждения используется абсолютная частота употребления слова в целевой коллекции, а в качестве штрафа – произведение абсолютных частот употреблений слова в контрастных коллекциях:

$$TF \cdot DCF = \frac{{f_t}^D}{\prod_{g \in G} 1+\log(1+{f_t}^g)}, \quad (12)$$

где ${f_t}^D$ и ${f_t}^g$ – частоты употреблений данного слова t в целевой и контрастной коллекциях соответственно; G – множество всех контрастных коллекций.

Авторы доказывают на серии экспериментов, что формула (12) является лучшей по точности извлечения терминов по сравнению с оценками, предложенными, например, в [11] и ряде других работ. Они оправдывают использование произведения в знаменателе формулы тем фактом, что штраф должен расти в геометрической прогрессии за каждое употребление слова в очередной контрастной коллекции. По мнению авторов, терминологичность слов, которые употребляются небольшое число раз в большом количестве контрастных коллекций, должна оцениваться ниже, чем слов, которые используются много раз, но в небольшом количестве контрастных коллекций. В случае, когда имеется одна контрастная коллекция, результаты формулы смещаются в сторону высокочастотных терминов.

Таким образом, в данном обзоре мы рассмотрели 7 наиболее интересных контрастных методов извлечения терминов (табл. 1). Все эти методы являются эвристическими, т.е. основаны на предположениях относительно характера распределения терминов в целевой и контрастной коллекциях. Сравнительный анализ этих утверждений показывает, что имеют место как совпадения позиций разных авторов, так и серьезные расхождения, что свидетельствует о наличии нерешенных проблем в данной области.

Авторы и ссылка на источник	Название метрики или индикатора	Год
Ahmad et al [8, 9]	Weirdness	1999, 2005
Peñas A. et al [10]	Relevance	2001
Kim et al [11]	Term frequency-inverse domain frequency	2009
Basili et al [12]	Contrastive weight	2001
Wong et al [13]	Domain prevalence, Domain tendency	2007
Sclano et al [14]	Domain pertinence, Domain consensus	2007
Lopes et al [15]	Term frequency-disjoint corpora frequency	2016

Таблица 1. Наиболее значимые способы операционализации контрастного подхода к извлечению терминов

Постановка задачи

В работе [16] отмечается, что методы извлечения терминов, основанные на эвристических предположениях, часто критикуют за отсутствие теоретической строгости. По словам авторов, такая критика становится очевидной, когда ставятся простые, но важные вопросы о способах операционализации тех или иных эвристик, например, «Почему в метрике используются разные основания логарифмов?» или «Почему объединение двух весов в метрике производится с помощью сложения, а не умножения?».

Между тем, в области автоматического извлечения терминов разработан ряд статистических критериев, опирающихся на строгие математические основания теории информации и теории вероятностей. К числу таких критериев относится взаимная информация (MI – mutual information) [4]. Ее понятие восходит к более общему фундаментальному понятию теории информации – информационной дивергенции, также известной как относительная энтропия или расстояние Кульбака–Лейблера. Информационная дивергенция – это несимметричная мера удаленности друг от друга двух дискретных распределений $P = \{p_i\}$ и $Q = \{q_i\}$:

$$D(P||Q) = \sum_i p_i \log\left(\frac{p_i}{q_i}\right). \quad (13)$$

Здесь в формуле (13) и далее в других формулах под основанием логарифма имеется в виду стандартное значение 2. Как правило, одно из сравниваемых распределений является «истинным» (наблю-

даемым), а второе – ожидаемым (проверяемым). Исходя из этого, информационную дивергенцию можно трактовать как меру того, насколько «истинное» распределение расходится с ожидаемым, приближенным.

MI представляет собой частный случай информационной дивергенции, при котором распределение P представляет собой совместное распределение двух случайных дискретных величин X и Y, а распределение Q – произведение маргинальных распределений этих случайных величин [17]:

$$MI(X,Y) = D(P(X,Y)||\ P(X) \times P(Y)\} = \sum_{i,j} p(x_i, y_j) \cdot \log\left(\frac{p(x_i,y_j)}{p(x_i)p(y_j)}\right). \quad (14)$$

Если случайные величины X и Y независимы, то вероятность их совместного распределения равна произведению вероятностей их маргинальных распределений $p(x_i, y_j) = p(x_i)p(y_j)$, тогда $\log\left(\frac{p(x_i,y_j)}{p(x_i)p(y_j)}\right) = \log 1 = 0$. Отсюда взаимная информация этих величин равна 0. Интуитивно это можно объяснить так: если две случайные величины независимы, то появление одной из них не дает никакой информации относительно появления другой. Соответственно, взаимную информацию можно трактовать как меру корреляции этих величин.

С понятием взаимной информации, определенной для двух случайных величин X, Y, тесно связано понятие точечной взаимной информации (PMI), определенной для конкретной пары исходов (x, y) этих случайных величин:

$$PMI(x,y) = \log\frac{p(x,y)}{p(x)\cdot p(y)}. \quad (15)$$

Сравнив между собой формулы (14) и (15), можно заметить, что взаимная информация MI – это средневзвешенная оценка значений PMI по всем парам исходов случайных величин X и Y, где в качестве весов используются вероятности этих пар исходов:

$$MI(X,Y) = \sum_{i,j} p(x_i, y_j) \cdot PMI(x_i, y_j). \quad (16)$$

Полученные формулы (15)–(16) хорошо адаптируются к задаче контрастного извлечения терминов. Прежде чем сформулировать постановку задачи с использованием взаимной информации, введем обозначения, которые используют авторы [4] (табл. 2). На основе этой таблицы они оценивают вероятности распределения двух случайных величин: X – присутствие слова в документе; Y – отношение документа к предметной области (табл. 3–5).

Количество документов	Целевой коллекции	Контрастной коллекции	Итого
Содержащих данное слово	A	B	$A + B$
Не содержащих данное слово	C	D	$C + D$
Итого	$A + C$	$B + D$	N=A+B+C+D

Таблица 2. Таблица сопряженности, описывающая распределение слов в коллекциях

Исход	Вероятность
t: Документ содержит данное слово	$p(t)$=(A+B)/N
$\bar{t}$: Документ не содержит данное слово	$p(t)$=(C+D)/N
Σ:	$p(t)+p(\bar{t})$ =1

Таблица 3. Маргинальное распределение случайной величины X

Исход	Вероятность
d: Документ относится к предметной области	$p(d)$=(A+C)/N
$\bar{d}$: Документ не относится к предметной области	$p(\bar{d})$=(B+D)/N
Σ:	$p(d)+p(\bar{d})$ =1

Таблица 4. Маргинальное распределение случайной величины Y

Исход	Вероятность
$t \wedge d$: Документ содержит данное слово и относится к предметной области	$p(t \wedge d) = A/N$
$\bar{t} \wedge d$: Документ не содержит данное слово и относится к предметной области	$p(\bar{t} \wedge d) = C/N$
$t \wedge \bar{d}$: Документ содержит данное слово и не относится к предметной области	$p(t \wedge \bar{d}) = B/N$
$\bar{t} \wedge \bar{d}$: Документ не содержит данное слово и не относится к предметной области	$p(\bar{t} \wedge \bar{d}) = D/N$
Σ:	1

Таблица 5. Совместное распределение случайных величин X, Y

Для каждого из четырех возможных исходов, приведенных в табл. 5, на основе формулы (15) выводится своя формула PMI (см. формулы (17)–(20)). Главный интерес, безусловно, представляет формула для исхода $t \wedge d$:

$$PMI(t \wedge d) = \log \frac{p(t \wedge d)}{p(t)p(d)} = \log \frac{A/N}{((A+B)/N)\times((A+C)/N)} = \log \frac{A \times N}{(A+B)\times(A+C)}. \quad (17)$$

Указанная формула позволяет оценить количество информации, которое несет факт присутствия данного слова в документе предметной области. Высокое количество информации говорит о том, что слово является хорошим индикатором предметной области. Формулы PMI для остальных трех исходов имеют следующий вид:

$$PMI(\bar{t} \wedge d) = \log \frac{p(\bar{t} \wedge d)}{p(\bar{t})p(d)} = \log \frac{C \times N}{(C+D)\times(A+C)}; \quad (18)$$

$$PMI(t \wedge \bar{d}) = \log \frac{p(t \wedge \bar{d})}{p(t)p(\bar{d})} = \log \frac{B \times N}{(A+B)\times(B+D)}; \quad (19)$$

$$PMI(\bar{t} \wedge \bar{d}) = \log \frac{p(\bar{t} \wedge \bar{d})}{p(\bar{t})p(\bar{d})} = \log \frac{D \times N}{(C+D)\times(B+D)}. \quad (20)$$

Аналогичным образом, на основе формулы (16) и найденных значений PMI выводится формула для MI:

$$MI = \frac{A}{N}\log \frac{A \times N}{(A+B)(A+C)} + \frac{C}{N}\log \frac{C \times N}{(C+D)(A+C)} + \frac{B}{N}\log \frac{B \times N}{(A+B)(B+D)} + \frac{D}{N}\log \frac{D \times N}{(C+D)(B+D)}. \quad (21)$$

Анализируя формулу (21), можно заметить, что она является симметричной по отношению к обеим коллекциям: целевой и контрастной. Это означает, что формула назначает высокие оценки как словам, тесно связанным с предметной областью, так и словам, тесно связанным с контрастной областью. В связи с этим при использовании этой формулы необходимо дополнительно проверять направление связи: если $A > B$, то слово связано с предметной областью, если $B > A$, то слово связано с контрастной областью.

Таким образом, обе формулы взаимной информации (17) и (21) оценивают количество информации, которое несет данное слово о предметной области. Тем не менее, между этими оценками существует принципиальная разница, лучше всего иллюстрируемая известным французским выражением «briller par son absence» (блистать отсутствием). Иначе говоря, MI, в отличие от PMI, оценивает связь между словом и предметной областью, учитывая не только вероятность его присутствия в целевой коллекции (исход $t \wedge d$), но и вероятность его отсутствия (исход $\bar{t} \wedge d$), а также вероятности его присутствия и отсутствия в контрастной коллекции (исходы $t \wedge \bar{d}$ и $\bar{t} \wedge \bar{d}$ соответственно).

В [4] отмечается, что в силу указанных особенностей PMI смещена в сторону редких терминов, в то время как средневзвешенная взаимная информация нормализует это смещение за счет использования весов. Рассмотрим, как именно происходит смещение оценок, т.е. определим, при каких условиях формула (17) достигает максимума. В приведенной формуле величины N и $A + C$ (общее число документов и число документов целевой коллекции соответственно) являются константами, так как не зависят от распределения слов. В этом случае их можно не учитывать, а рассматривать функцию $\log\left(\frac{A}{A+B}\right) = \log\left(\frac{1}{1+B/A}\right)$. Поскольку речь идет о логарифме по основанию 2 (возрастающей функции), то функция стремится к максимуму, когда B/A стремится к минимуму. Выражение B/A достигает минимума только при $B = 0$, т.е. когда слово ни разу не встречается в контрастной коллекции. При этом не имеет никакого значения, чему равно A, т.е. неважно, 50 или 10 раз встречается слово в целевой коллекции, оценка будет одинаковой. В остальных случаях, когда B отлично от 0, значение A также не влияет на оценку терминологичности, имеет значение только соотношение B/A. Например, если первое слово встретилось 50 раз в целевой коллекции и 10 раз в контрастной, а второе слово встретилось 10 раз в целевой и 1 раз в контрастной коллекции, то выше будет оценка терминологичности второго слова, так как $B_1/A_1 = 10/50 > B_2/A_2 = 1/10$.

Теперь мы можем сформулировать постановку задачи и заодно алгоритм ее решения.

1. Даны две коллекции текстов: предметная (целевая) и контрастная.
2. Тексты коллекции необходимо подвергнуть токенизации (разбиению на слова), лемматизации (приведению слов в нормальные формы) и исключению стоп-слов.
3. Из полученных слов необходимо сформировать словарь коллекции, и для каждого слова определить количество его вхождений в тексты предметной и контрастной коллекций.
4. На основе этих вхождений для каждого слова необходимо вычислить значения *A*, *B*, *C*, *D* (табл. 2).
5. Найденные значения необходимо подставить в формулу (17) и, таким образом, для каждого слова определить точечную взаимную информацию.
6. Список слов необходимо отсортировать по убыванию значений PMI и выделить из него топ первых *N* слов (*N* определяется экспериментально). Это список терминов-кандидатов (список вероятных редких терминов).
7. Дальнейшая задача заключается в том, чтобы отсеять из этого списка лжетермины и оставить только действительно редкие термины, относящиеся к предметной области.

Дадим постановку и этой задачи.

1. Дан список терминов-кандидатов (список редких слов).
2. Для каждого слова из этого списка необходимо определить такой критерий, который позволял бы более точно оценить его терминологичность, т.е. силу его связи с предметной областью. Таким образом, на данном шаге отпадает необходимость использования контрастной коллекции. Силу связи слова с предметной областью нужно оценивать на основе его близости к другим терминами предметной области (назовем это внутренней связью).
3. Оценив степень силу внутренней связи для каждого из слов-кандидатов в списке, следует отсортировать этот список по убыванию силы внутренней связи и отбросить слова с низкими значениями. Пороговое значение следует определить эмпирически.

Предлагаемый подход

Предлагаемая идея для оценки связи слова с предметной областью не нова. В [18] отмечается, что традиционно значимость каждого термина-кандидата определяется на основе анализа того, как он связан с другими словами текста или коллекции.

Считается, что термин-кандидат имеет большой вес, если он связан либо с большим количеством других терминов-кандидатов, либо с терминами-кандидатами, которые сами имеют большой вес. В данной работе мы опираемся на эту идею и, как и авторы [19], в качестве инструмента измерения связей между словами рассматриваем матрицу совместной встречаемости.

Одной из разновидностей матрицы совместной встречаемости терминов является матрица «документы-на-термины», описывающая распределение слов (терминов) в текстах (документах обучающей коллекции). Как следует из названия матрицы, ее строками являются документы, столбцами – термины, а элементами – частоты употребления терминов в документах. Вклад нашей работы (и соответственно, отличие от [19]) заключается в способе построения и обработки этой матрицы.

Сначала мы формируем матрицу «документы-на-термины», строками которой являются только документы целевой коллекции, а столбцами – только выделенные репрезентативные слова (для их поиска можно использовать критерий хи-квадрат или информационную выгоду, т.е. средневзвешенную взаимную информацию) и сами термины-кандидаты. Затем мы подвергаем эту матрицу операции сингулярного разложения, чтобы избавить ее от шума и разреженности, поскольку, как отмечается в [20], «подобные матрицы имеют склонность быть разреженными и зашумленными, особенно если обучающая коллекция относительно мала в размерах».

Согласно теореме Эккарта–Янга, сингулярное разложение позволяет снизить шум и разреженность исходной матрицы, заменяя ее матрицей той же размерности, но меньшего ранга, в которой сохранена только самая значимая информация [21]. Более формально эта теорема звучит следующим образом.

Теорема. (Эккарт–Янг). Пусть дана матрица $\mathbf{A}$ размерности $m \times n$, для которой известно сингулярное разложение $\mathbf{A} = \mathbf{USV}^T$ и которую требуется аппроксимировать матрицей $\mathbf{A}_k$ с заданным рангом $k < r = \mathrm{rank}(\mathbf{A})$. Если в матрице $\mathbf{S}$ оставить k наибольших сингулярных значений, а остальные заменить нулями, то разложение

$$\mathbf{A}_k = \mathbf{US}_k\mathbf{V}^T \tag{22}$$

даст наилучшее приближение исходной матрицы $\mathbf{A}$ ранга k в смысле нормы Фробениуса. Если при этом элементы матрицы $\mathbf{S}$ отсортированы по убыванию $\mathbf{s}_1 \geq \mathbf{s}_2 \geq \mathbf{s}_N \geq 0$, то формула (22) может быть записано в другой форме:

$$\mathbf{A}_k = \mathbf{U}_k\mathbf{S}_k\mathbf{V}_k^T, \tag{23}$$

где $\mathbf{U}_k$ и $\mathbf{V}_k$ – это матрицы, полученные выделением первых k столбцов из матриц $\mathbf{U}$ и $\mathbf{V}$ соответственно. Сингулярное разложение, представленное равенством (23), называется экономным, поскольку в случае, когда k намного меньше, чем m и n, оно позволяет произвести существенное сжатие исходной информации.

После этого мы переходим к формированию матрицы семантических связей. Поскольку в матрице «документы-на-термины» каждое слово представляет собой вектор-столбец, то семантическую связь между любыми двумя словами можно трактовать как близость между соответствующими этим словам векторами, используя при этом любые известные метрики [20]. В данной работе используем косинусную меру:

$$r_{ij} = \cos\left(\overline{\mathbf{T}}_i, \overline{\mathbf{T}}_j\right) = \frac{\overline{\mathbf{T}}_i \cdot \overline{\mathbf{T}}_j}{|\overline{\mathbf{T}}_i| \cdot |\overline{\mathbf{T}}_j|}, \tag{24}$$

где $\overline{\mathbf{T}}_i, \overline{\mathbf{T}}_j$ – это вектор-столбцы матрицы «документы-на-термины», соответствующие i-му и j-му словам соответственно (i, j пробегают весь список слов), r_{ij} – это значение близости, элемент матрицы семантических связей. Определение косинуса в первом квадранте декартовых координат позволяет утверждать, что максимально возможное значение близости между словами равно 1, а минимально возможное – 0.

Поскольку нас интересуют только самые сильные и устойчивые связи, то мы принимаем во внимание только значения близости, полученные по формуле (24), которые выше определенного порога. Таким образом, для каждого термина-кандидата мы оцениваем количество сильных связей с другими словами, и

если это количество превышает средний показатель по всем рассматриваемым словам, то мы считаем данное слово термином.

Экспериментальная реализация предлагаемого подхода

Для проведения экспериментов мы использовали учебник по общей геологии [22]. Мы разбили главы учебника на отдельные документы и осуществили необходимые операции по предобработке целевой коллекции: токенизацию, лемматизацию, морфологический анализ. В качестве контрастной коллекции мы взяли сборную коллекцию новостей из разделов «Политика», «Культура», «Экономика» и «Происшествия» одного из новостных интернет-сайтов. Мы провели те же операции по предобработке и для контрастной коллекции. Общий словарь двух коллекций составил 14207 слов. Коллекции были ограничены по размеру, но, тем не менее, даже на таких данных мы получили довольно интересные результаты.

Используя критерии MI и PMI, мы извлекли репрезентативные и редкие слова предметной области «Геология». В класс репрезентативных слов мы отобрали все слова, для которых значение средневзвешенной взаимной информации MI было выше критического значения 0,0192, рассчитанного нами на основе таблицы Стьюдента. Всего в класс репрезентативных слов было отобрано 1421 слово. Наивысшее значение средневзвешенной взаимной информации 0,5518 было получено для слова «порода».

В класс специфических и редких слов мы отобрали все слова, для которых значение PMI было выше критического значения 0,9449, также рассчитанного нами на основе таблицы Стьюдента. Чтобы классы репрезентативных и специфических слов не пересекались, мы исключили из класса специфических слов все слова, у которых показатель MI был выше 0,0192. Всего в класс специфических слов вошло 6489 слов.

Номер	Слово	Значение PMI	Количество связей	Термин
1	**бомба**	1	136	да
2	бомбардировка	1	4	нет
3	**брекчия**	1	134	да
4	буря	1	10	нет
5	**бухта**	1	85	да
6	вектор	1	24	нет
7	**гематит**	1	45	да
8	**гидрогематит**	1	54	да
9	изобилие	1	2	нет
10	**калишпат**	1	66	да
11	кальмар	1	9	нет
12	камешек	1	29	нет
13	камыш	1	13	нет
14	курорт	1	8	нет
15	курортный	1	27	нет
16	**абразия**	1	82	да

Таблица 6. Таблица специфических терминов предметной области «Геология»

Таким образом, как мы уже отмечали, среди слов обоих классов встречались как термины, так и слова, похожие на термины, но терминами не являющиеся. Решение о терминологичности редких слов принималось на основе оценки количества сильных связей: если количество сильных связей слова было выше порогового значения 30, то слово считалось термином (табл. 6). К сожалению, мы не нашли способа оценить это значение не эмпирически, и скорее всего с этим будут связаны наши будущие работы. Для расчета сильных связей мы сформировали матрицу «документы-на-термины», в которую вошли все документы целевой коллекции и слова из обоих классов. Таким образом, размерность матрицы составила 87× 7522. Полученная матрица была подвергнута сингулярному разложению с параметром $k = 30$, после чего были выполнен переход к матрице «термины-на-термины», определяющей связи между словами.

Интересно отметить, что когда наш метод отобрал в качестве специфического геологического термина слово «бомба», у которого имелось 136 сильных связей с другими словами, это нас озадачило и заставило сомневаться в достоверности нашего метода. Однако, обратившись к Википедии, мы выяснили, что слово бомба – это действительно геологический термин, означающий комок или обрывок лавы, выброшенный во время извержения вулкана из жерла и получивший при выжимании, во время полета и застывания на воздухе, специфическую форму.

Таким образом, из 6489 терминов-кандидатов было отобрано 1226 истинных терминов. Это еще раз подтверждает утверждение авторов работы [4], что концептуальный аппарат предметной области формируется не столько из высокочастотных репрезентативных слов, сколько из слов редких и специфических.

Заключение

Полученные в ходе эксперимента позитивные результаты подтверждают преимущества раздельного отбора редких и репрезентативных терминов и показывают перспективность предлагаемого подхода. Следующим шагом развития нашего подхода является сравнение точности и полноты наших результатов с результатами, полученными другими методами, в частности, с подходом, описанным в [15]. Предварительные эксперименты по сравнению двух подходов были уже проведены и показали, что у каждого из методов есть свои сильные и слабые стороны. Однако требуется детальный анализ выявленных особенностей, чему и будет посвящена наша следующая работа.

Литература

1. Weeber M., Vos R., Baayen R.H. Extracting the lowest-frequency words: pitfalls and possibilities // Computational Linguistics. 2000. V. 26. N 3. P. 301–317. doi: 10.1162/089120100561719
2. Astrakhantsev N.A., Fedorenko D.G., Turdakov D.Y. Methods for automatic term recognition in domain-specific text collections: a survey // Programming and Computer Software. 2015. V. 41. N 6. P. 336–349. doi: 10.1134/s036176881506002x
3. Heylen K., De Hertog D. Automatic term extraction / In: Handbook of Terminology. Amsterdam, 2014. V. 1.
4. Yang Y., Pedersen J.O. A comparative study on feature selection in text categorization // Proc. 14th Int. Conf. on Machine Learning (ICML). 1997. V. 97. P. 412–420.
5. Браславский П.И., Соколов Е.А. Сравнение четырех методов автоматического извлечения двухсловных терминов из текста // Компьютерная лингвистика и интеллектуальные технологии. Сборник трудов Международной конференции Диалог'2006. Москва, 2006. С. 88–94.
6. Kim S.N., Cavedon L. Classifying domain-specific terms using a dictionary // Proc. Australasian Language Technology Association Workshop 2011. 2011. P. 57.
7. Conrado M.S., Pardo T.A.S., Rezende S.O. A machine learning approach to automatic term extraction using a rich feature set // Proc. NAACL HLT Student Research Workshop. Atlanta, USA, 2013. P. 16–23.
8. Ahmad K., Gillam L., Tostevin L. University of surrey participation in TREC8: weirdness indexing for logical document extrapolation and retrieval (WILDER) // Proc. 8th Text Retrieval Conference TREC. Gaithersburg, USA, 1999. P. 717.
9. Gillam L., Tariq M., Ahmad K. Terminology and the construction of ontology // Terminology. 2005. V. 11. N 1. P. 55–81.
10. Penas A., Verdejo F., Gonzalo J. Corpus-based terminology extraction applied to information access // Proceedings of Corpus Linguistics. 2001. V. 2001. P. 458–465.
11. Kim S.N., Baldwin T., Kan M.-Y. An unsupervised approach to domain-specific term extraction // Proc. Australasian Language Technology Association Workshop. 2009. P. 94–98.
12. Basili R. A contrastive approach to term extraction // Proc. 4th Terminological and Artificial Intelligence Conference (TIA2001). Nancy, France, 2001.
13. Wong W., Liu W., Bennamoun M. Determining termhood for learning domain ontologies using domain prevalence and tendency // Proc. 6th Australasian Conference on Data Mining and Analytics. Gold Coast, Australia, 2007. V. 70. P. 47–54.
14. Sclano F., Velardi P. Termextractor: a web application to learn the shared terminology of emergent web communities / In : Enterprise Interoperability II. Springer, 2007. P. 287–290. doi: 10.1007/978-1-84628-858-6_32
15. Lopes L., Fernandes P., Vieira R. Estimating term domain relevance through term frequency, disjoint corpora frequency-tf-dcf // Knowledge-Based Systems. 2016. V. 97. P. 237–249. doi: 10.1016/j.knosys.2015.12.015
16. Wong W., Liu W., Bennamoun M. Determining termhood for learning domain ontologies in a probabilistic framework // Proc. 6th Australasian Conference on Data Mining and Analytics. Gold Coast, Australia, 2007. V. 70. P. 55–63.

References

1. Weeber M., Vos R., Baayen R.H. Extracting the lowest-frequency words: pitfalls and possibilities. *Computational Linguistics*, 2000, vol. 26, no. 3, pp. 301–317. doi: 10.1162/089120100561719
2. Astrakhantsev N.A., Fedorenko D.G., Turdakov D.Y. Methods for automatic term recognition in domain-specific text collections: a survey. *Programming and Computer Software*, 2015, vol. 41, no. 6, pp. 336–349. doi: 10.1134/s036176881506002x
3. Heylen K., De Hertog D. Automatic term extraction. In *Handbook of Terminology*. Amsterdam, 2014, vol. 1.
4. Yang Y., Pedersen J.O. A comparative study on feature selection in text categorization. *Proc. 14th Int. Conf. on Machine Learning ICML*, 1997, vol. 97, pp. 412–420.
5. Braslavskii P.I., Sokolov E.A. Comparison of four methods for automatic two-word term extraction. *Computational Linguistics and Intellectual Technologies. Proc. Int. Conf. Dialog 2006*. Moscow, 2006, pp. 88–94. (In Russian)
6. Kim S.N., Cavedon L. Classifying domain-specific terms using a dictionary. *Proc. Australasian Language Technology Association Workshop 2011*, 2011, p. 57.
7. Conrado M.S., Pardo T.A.S., Rezende S.O. A machine learning approach to automatic term extraction using a rich feature set. *Proc. NAACL HLT Student Research Workshop*. Atlanta, USA, 2013, pp. 16–23.
8. Ahmad K., Gillam L., Tostevin L. University of surrey participation in TREC8: weirdness indexing for logical document extrapolation and retrieval (WILDER). *Proc. 8th Text Retrieval Conference TREC*. Gaithersburg, USA, 1999, p. 717.
9. Gillam L., Tariq M., Ahmad K. Terminology and the construction of ontology. *Terminology*, 2005, vol. 11, no. 1, pp. 55–81.
10. Penas A., Verdejo F., Gonzalo J. Corpus-based terminology extraction applied to information access. *Proceedings of Corpus Linguistics*, 2001, vol. 2001, pp. 458–465.
11. Kim S.N., Baldwin T., Kan M.-Y. An unsupervised approach to domain-specific term extraction. *Proc. Australasian Language Technology Association Workshop*, 2009, pp. 94–98.
12. Basili R. A contrastive approach to term extraction. *Proc. 4th Terminological and Artificial Intelligence Conference TIA2001*. Nancy, France, 2001.
13. Wong W., Liu W., Bennamoun M. Determining termhood for learning domain ontologies using domain prevalence and tendency. *Proc. 6th Australasian Conference on Data Mining and Analytics*. Gold Coast, Australia, 2007, vol. 70, pp. 47–54.
14. Sclano F., Velardi P. Termextractor: a web application to learn the shared terminology of emergent web communities. In *Enterprise Interoperability II*. Springer, 2007, pp. 287–290. doi: 10.1007/978-1-84628-858-6_32
15. Lopes L., Fernandes P., Vieira R. Estimating term domain relevance through term frequency, disjoint corpora frequency-tf-dcf. *Knowledge-Based Systems*, 2016, vol. 97, pp. 237–249. doi: 10.1016/j.knosys.2015.12.015
16. Wong W., Liu W., Bennamoun M. Determining termhood for learning domain ontologies in a probabilistic framework. *Proc. 6th Australasian Conference on Data Mining and Analytics*. Gold Coast, Australia, 2007, vol. 70, pp. 55–63.
17. Prelov V. Mutual information of several random variables

17. Prelov V. Mutual information of several random variables and its estimation via variation // Problems of Information Transmission. 2009. V. 45. N 4. P. 295–308. doi: 10.1134/s0032946009040012
18. Hasan K.S., Ng V. Automatic keyphrase extraction: a survey of the state of the art // Proc. 52nd Annual Meeting of the Association for Computational Linguistics. 2014. V. 1. P. 1262–1273. doi: 10.3115/v1/p14-1119
19. Matsuo Y., Ishizuka M. Keyword extraction from a single document using word co-occurrence statistical information // International Journal on Artificial Intelligence Tools. 2004. V. 13. N 1. P. 157–169. doi: 10.1142/s0218213004001466
20. Slonim N., Tishby N. The power of word clusters for text classification // Proc. 23rd European Colloquium on Information Retrieval Research. 2001. V. 1.
21. Eckart C., Young G. The approximation of one matrix by another of lower rank // Psychometrika. 1936. V. 1. N 3. P. 211–218. doi: 10.1007/bf02288367
22. Общая геология / Под ред. А.К. Соколовского. М.: КДУ, 2006. Т. 1. 448 с.

and its estimation via variation. *Problems of Information Transmission*, 2009, vol. 45, no. 4, pp. 295–308. doi: 10.1134/s0032946009040012
18. Hasan K.S., Ng V. Automatic keyphrase extraction: a survey of the state of the art. *Proc. 52nd Annual Meeting of the Association for Computational Linguistics*, 2014, vol. 1, pp. 1262–1273. doi: 10.3115/v1/p14-1119
19. Matsuo Y., Ishizuka M. Keyword extraction from a single document using word co-occurrence statistical information. *International Journal on Artificial Intelligence Tools*, 2004, vol. 13, no. 1, pp. 157–169. doi: 10.1142/s0218213004001466
20. Slonim N., Tishby N. The power of word clusters for text classification. *Proc. 23rd European Colloquium on Information Retrieval Research*, 2001, vol. 1.
21. Eckart C., Young G. The approximation of one matrix by another of lower rank. *Psychometrika*, 1936, vol. 1, no. 3, pp. 211–218. doi: 10.1007/bf02288367
22. *General Geology*. Ed. A.K. Sokolovskyi. Moscow, KDU Publ., 2006, vol. 1, 448 p. (In Russian)

Авторы

Бессмертный Игорь Александрович – доктор технических наук, профессор, доцент, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, bia@cs.ifmo.ru

Нугуманова Алия Багдатовна – кандидат наук, старший преподаватель, Восточно-Казахстанский государственный университет им. С. Аманжолова, Усть-Каменогорск, 070004, Казахстан, yalisha@yandex.kz

Мансурова Мадина Есимхановна – кандидат физико-математических наук, доцент, доцент, Казахский национальный университет имени аль-Фараби, Алматы, 050040, Казахстан, mansurova01@mail.ru

Байбурин Ержан Мухаметкалиевич – инженер-программист, Восточно-Казахстанский государственный университет им. С. Аманжолова, Усть-Каменогорск, 070004, Казахстан, ebaiburin@gmail.com

Authors

Igor A. Bessmertny – D.Sc., Professor, Associate Professor, ITMO University, Saint Petersburg, 197101, Russian Federation, bia@cs.ifmo.ru

Aliya B. Nugumanova – PhD, Senior lecturer, S. Amanzholov East Kazakhstan State University, Ust Kamenogorsk, 070004, the Republic of Kazakhstan, yalisha@yandex.kz

Madina Ye. Mansurova – PhD, Associate professor, Associate professor, Al-Farabi Kazakh National University, Almaty, 050040, Republic of Kazakhstan, mansurova01@mail.ru

Yerzhan M. Baiburin – software engineer, S. Amanzholov East Kazakhstan State University, Ust Kamenogorsk, 070004, the Republic of Kazakhstan, ebaiburin@gmail.com