НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
ноябрь–декабрь 2016 Том 16 № 6 ISSN 2226-1494 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
November–December 2016 Vol. 16 No 6 ISSN 2226-1494 http://ntv.ifmo.ru/en

УДК 20.19.27

СТАТИСТИЧЕСКИЙ МЕТОД ИЗВЛЕЧЕНИЯ ТЕРМИНОВ ИЗ КИТАЙСКИХ ТЕКСТОВ БЕЗ СЕГМЕНТАЦИИ ФРАЗ

И.А. Бессмертный[a], Юй Чуцяо[a], Ма Пенюй[a]

[a] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация
Автор для переписки: bia@cs.ifmo.ru

Информация о статье
Поступила в редакцию 29.07.16, принята к печати 15.10.16
doi: 10.17586/2226-1494-2016-16-6-1096-1102
Язык статьи – русский

Ссылка для цитирования: Бессмертный И.А., Юй Чуцяо, Ма Пенюй. Статистический метод извлечения терминов из китайских текстов без сегментации фраз // Научно-технический вестник информационных технологий, механики и оптики. 2016. Т. 16. № 6. С. 1096–1102.doi: 10.17586/2226-1494-2016-16-6-1096-1102

Аннотация

Работа посвящена проблеме автоматического извлечения знаний из естественно-языковых текстов (*text mining*). Одной из первоочередных задач в рамках данной проблемы является формирование тезауруса предметной области. Существуют достаточно апробированные статистические методы извлечения терминов для алфавитных языков, например, латентный семантический анализ. Применение данных методов для иероглифического письма сопряжено с проблемой, обусловленной отсутствием в таких языках пробелов между словами. Задача сегментации предложений на иероглифических языках обычно решается либо на основе словарей, либо статистическими методами, в частности, с использованием метода взаимной информации. Методы сегментации предложений, как и методы извлечения терминов по отдельности, не обладают 100%-ой точностью и полнотой, а их последовательное применение только увеличивает процент ошибок. **Целью данной работы** является повышение полноты и точности извлечения терминов предметной области из иероглифических текстов. **Предлагаемый метод** состоит в выявлении повторяющихся последовательностей длиной от двух до четырех символов в каждом предложении и соотнесения частот встречаемости этих последовательностей в целевой и контрастной коллекциях документов. В результате проведенного исследования было установлено, что простое ранжирование всех возможных последовательностей символов позволяет удовлетворительно выявлять только наиболее часто используемые термины. Фильтрация последовательностей символов по соотношению их частот в целевой и контрастной коллекциях позволила надежно извлекать часто используемые термины и удовлетворительно – термины с низкой частотой. В работе приведены результаты извлечения терминов предметной области «сетевые технологии» из текста на китайском языке, где в качестве контрастной коллекции использовался набор статей из газеты «Женьминь жибао», в результате чего получены вполне удовлетворительные результаты.

Ключевые слова

обработка естественно-языковых текстов, мешок слов, китайский язык, сегментация слов, извлечение терминов, тезаурус предметной области

STATISTICAL METHOD OF TERM EXTRACTION FROM CHINESE TEXTS WITHOUT PRELIMINARY SEGMENTATION OF PHRASES

I.A. Bessmertny[a], Yu Chuqiao[a], Ma Pengyu[a]

[a] ITMO University, Saint Petersburg, 197101, Russian Federation
Corresponding author: bia@cs.ifmo.ru

Article info
Received 13.04.16, accepted 07.10.16
doi: 10.17586/2226-1494-2016-16-6-1096-1102
Article in Russian

For citation: Bessmertny I.A., Yu Chuqiao, Ma Pengyu. Statistical method of term extraction from Chinese texts without preliminary segmentation of phrases. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2016, vol. 16, no. 6, pp. 1096–1102. doi: 10.17586/2226-1494-2016-16-6-1096-1102

Abstract

Subject of Research. The paper considers the problem of automatic term extraction from natural language texts (text mining). One of the first-priority problems in this topic is creation of domain thesaurus. Some well approved methods of terms extraction exist for alphabetic languages, for instance, the latent semantic analysis. Applying of these methods for

hieroglyphic texts is challenged because of missing blanks between words. The sentences segmentation task in hieroglyphic languages is usually solved by dictionaries or by statistical methods, particularly, by means of a mutual information approach. Methods of sentences segmentation, as methods of terms extraction, separately, do not reach 100 percent accuracy and fullness, and their consistent applying just increases a number of errors. The aim of this work is improving the fullness and accuracy of domain terms extraction from hieroglyphic texts. **Method.**The proposed method lies in detection of repeating two, three or four symbol sequences in each sentence and correlation of occurrence frequencies for these sequences in domain and contrast documents collection. According to research carried out it was stated that a trivial ranging of all possible symbol sequences enables to extract satisfactory only frequently using terms. Filtering of symbol sequences by their ratio of frequencies in the domain and contrast collection gave the possibility to extract reliably frequently used terms and find satisfactory rare domain terms. Some results of terms extraction for the "Network technologies" domain from a Chinese text are presented in this paper. A set of articles from the newspaper "*Rénmín Rìbào*" was used as a contrast collection and some satisfactory results were obtained.

Keywords
text mining, bag of words, Chinese language, words segmentation, terms extraction, domain thesaurus

Введение

Извлечение знаний из естественно-языковых текстов (*text mining*) является актуальной проблемой, решение которой позволит существенно повысить эффективность использования информационных ресурсов, хранящихся в виде текстовых документов [1]. Актуальность проблемы многократно возрастает, если речь идет о документах на иностранных языках, отличных от английского. Если для алфавитных языков существуют методы, в той или иной степени решающие задачу извлечения знаний, то для иероглифических языков удовлетворительных решений на данный момент не существует.

Одной из первоочередных задач в этой области является автоматическое формирование тезауруса предметной области. Наиболее популярный подход основан на модели Bag-of-words (мешок слов) [1, 2], в которой игнорируются связи между словами и предложениями и для которых имеются достаточно апробированные методы извлечения терминов предметной области, среди которых следует упомянуть латентный семантический анализ [3].

Состояние проблемы и текущие исследования

Особенностью иероглифического письма, в частности, китайского языка, является отсутствие пробелов между словами, что порождает проблему сегментации предложений. Таким образом, задача извлечения терминов распадается на задачу сегментации текста на слова и последующего извлечения терминов. Несмотря на то, что существуют правила сегментации, основанные на словарях [4], задача сегментации иероглифических текстов даже с использованием словарей не имеет однозначного решения [5]. Статистические способы сегментации предложений, среди которых следует упомянуть метод взаимной информации [6], позволяют обходиться без словарей, но также не обеспечивают однозначную сегментацию. В упомянутой работе [6] коэффициенты точности и полноты разбиения текста на слова не превышают 0,9.

Методы извлечения китайских терминов из сегментированных текстов чаще всего базируются на хорошо зарекомендовавших себя алгоритмах TF-IDF (Term Frequency-Inverted Document Frequency) [7, 8]. В работе [7] алгоритм TF-IDF модифицируется путем добавления к нему меры информации DI (Distribution Information); при этом точность не превышает 0,68, а полнота – 0,77. С учетом погрешностей сегментации предложений данные цифры следует скорректировать до значений 0,6 и 0,7 соответственно. В работе [8] предлагается для извлечения редких терминов использовать корпус объемом не меньше 3010 слов и игнорировать слова, встречающиеся менее трех раз. Авторы утверждают, что в этом случае при использовании алгоритма TF-IDF вообще не происходит потерь ключевых слов, что представляется сомнительным. Таким образом, известные исследования базируются на разделении фаз сегментации текста и извлечения терминов, и декларируемые результаты демонстрируют большой разброс величин точности и полноты, что свидетельствует о недостаточной разработанности данной темы.

В настоящей работе предложено отказаться от фазы сегментации предложений и использовать в качестве кандидатов в термины все возможные последовательности символов между знаками препинания. Основная идея предлагаемого подхода состоит в выявлении часто используемых последовательностей символов и их фильтрации в зависимости от частот встречаемости в целевой и контрастной коллекции.

Особенности иероглифических текстов

Минимальной лексической единицей иероглифических языков, среди которых и китайский язык, является иероглиф. Отличие иероглифов от слов алфавитных языков состоит в том, что иероглиф обозначает достаточно широкое понятие, несущее в себе десятки смыслов. Объединение двух и более иероглифов сужает и конкретизирует передаваемый ими смысл. Например, если иероглиф 出 означает движение изнутри наружу, а 口 – рот или любое отверстие, то 出口 означает выход или выезд. Несмотря на наличие

более точных иероглифов 門 и 门, обозначающих дверь и ворота соответственно, сочетание 出口 используется для обозначения как выхода из здания, так и для выезда с парковки. Таким образом, в отличие от букв алфавита, каждый иероглиф несет в себе смысл, и практически любое сочетание иероглифов можно интерпретировать каким-либо образом. Это серьезно усложняет поставленную задачу. С другой стороны, как показано выше, имеет место унификация сочетаний иероглифов, что позволяет выявлять статистически значимые последовательности.

Китайский язык отличает жесткий порядок слов и чрезвычайно простая грамматика: отсутствуют склонения, спряжения, времена и числа. Следовательно, этап лемматизации при обработке китайских текстов не требуется. Вопросительные предложения имеют тот же порядок слов, что и повествовательные, но с использованием вопросительных слов в конце предложения. Все это делает китайский язык достаточно привлекательным в качестве объекта автоматического извлечения знаний.

Наконец, иероглифические языки отличает отсутствие пробелов между словами, что объясняется тем обстоятельством, что для носителя языка это не является проблемой. Так же, как мы не читаем по буквам, китайцы не интерпретируют отдельные иероглифы, а распознают их устойчивые сочетания. Следовательно, аналогичный антропоморфный подход может быть применен и для автоматической обработки иероглифических текстов.

Частотный анализ последовательностей символов

Представим текст в виде последовательности символов вида *abcdefghijk*, расположенной между терминальными символами, в качестве который выступают не только знаки препинания, но и любые символы, отличные от иероглифов. Считая, что термины предметной области могут состоять из двух, трех или четырех символов, возможны следующие интерпретации указанной последовательности: четырехсимвольные – *abcd*, *bcde*, *cdef*, *defg*, *efgh*, *fghi*, *ghij*, *hijk*, трехсимвольные – *abc*, *bcd*, *cde*, *def*, *efg*, *fgh*, *ghi*, *hij*, *ijk*, двухсимвольные – *ab*, *bc*, *cd*, *de*, *ef*, *fg*, *gh*, *hi*, *ij*, *jk*. Часть из них является терминами предметной области, часть – общетехническими терминами, часть – общеупотребительными словами, остальные – бессмысленные сочетания.

В качестве предметной области выберем компьютерные сетевые технологии, а эталонный текст возьмем из книги «Основы компьютерных сетей», глава 3 – Локальные вычислительные сети (Local Area Networks – LAN) (http://ebook.qq.com/hvread.html?bid=637747&cid=3). Объем данного текста – 19 тыс. символов, из которых были получены 10978, 12563 и 14383 цепочки по 4, 3 и 2 символа соответственно. После извлечения всех цепочек символов и ранжирования по их встречаемости в тексте N получим результаты, фрагмент которых приведен в табл. 1.

В табл. 1 показаны наиболее часто встречающиеся четырехсимвольные сочетания иероглифов, среди которых есть термины предметной области («беспроводной доступ в Интернет», «радиосигнал», «зона Френеля» и др.), а также их производные (притяжательные обороты: «переключателя», «локальной сети»), сочетания символов, которые имеют тот же смысл, что и термины, но не являются таковыми («коммутатор сети», «технология Ethernet»). Среди часто используемых цепочек символов бессмысленных сочетаний иероглифов не встретилось.

Естественно, чистый частотный анализ позволяет выявлять только часто используемые термины предметной области и общеупотребительные слова, не делая между ними различий.

Извлечение терминов и использованием контрастной коллекции

Традиционным подходом к улучшению качества извлечения терминов предметной области является использование контрастной коллекции, относящейся к другой предметной области, или общей коллекции, не относящейся ни к какой предметной области [9]. Все существующие методы, так или иначе, поощряют присутствие кандидата в термины в целевой коллекции и штрафуют за присутствие в контрастной. Одной из первых работ в этом направлении следует назвать работу [10], в которой вычисляется «странность» (*Weirdness*) как отношение частот слова в целевой и контрастной коллекциях. Следует отметить, что в китайских текстах, не разбитых на слова, при таком подходе должна достаточно часто проявляться сингулярность, когда знаменатель обращается в ноль. Это связано с тем обстоятельством, что последовательности, включающие в себя фрагменты терминов и общеупотребительные слова, вполне могут быть уникальными. Другие модификации подхода TF-IDF [11–15] также являются эмпирическими, опираются на разные гипотезы о характере распределения терминов в целевой и контрастной коллекциях. Разнообразие, а также отсутствие явно выраженного преимущества какого-либо из существующих методов выявления терминологичности слов свидетельствует о том, что решение данной проблемы еще не найдено.

N	Последовательность	Перевод	Тип последовательности
54	线局域网	линия LAN	
52	无线局域	беспроводной доступ в Интернет	термин
22	交换机的	переключателя (род. падеж)	
20	局域网的	локальной сети (род. падеж)	
20	兆以太网	Гигабитная сеть	
19	无线信号	радиосигнал	термин
19	千兆以太	Гигабитный Ethernet	
18	数据传输	передача данных	термин
16	菲涅耳区	зона Френеля	термин
16	传输速率	скорость передачи данных	термин
16	以太网的	Интернета (род. падеж)	
15	网交换机	коммутатор сети	
13	线路由器	линия маршрутизатора	
13	无线路由	беспроводной маршрут	
13	无线网络	беспроводная сеть	термин
13	传输距离	дальность передачи	термин
12	质访问控	управление доступом к среде	
12	访问控制	контроль доступа	термин
12	覆盖范围	покрытие	термин
12	的计算机	компьютера (род. падеж)	
12	的局域网	локальной сети (род. падеж)	
12	的以太网	Ethernet (род. падеж)	
12	波束成形	формирование пучка	
12	无线接入	беспроводной доступ	термин
11	以太网技	Технология Ethernet	
10	载波信号	сигнал несущей частоты	термин
10	无线设备	беспроводное оборудование	термин
10	无线电波	радиоволны	термин
10	换式局域	локальная формула	
10	式局域网	LAN (повелит. наклонение)	
10	局域网络	LAN	термин
10	台计算机	один компьютер	

Таблица 1. Результаты частотного анализа последовательностей символов

При внимательном рассмотрении результатов чистого частотного анализа китайского текста была обнаружена закономерность, состоящая в том, что непереводимые последовательности из четырех иероглифов чаще всего состоят из общеупотребительных слов с добавлением предлогов или других фрагментов. Следовательно, если с помощью контрастной коллекции выявить общеупотребительные слова, то можно отфильтровать бессмысленные сочетания символов. Таким образом, для фильтрации списка, полученного на основе чистого частотного анализа, предлагается следующий подход. Пусть в целевой коллекции встречается последовательность *abcd*, которая включает в себя фрагменты *abc*, *bcd*, *ab* и *cd*. Дан-

ная последовательность включается в список терминов только в том случае, если вероятность присутствия не только самой последовательности, но и любого из ее фрагментов в целевой коллекции выше, чем в контрастной c:

$$p(abcd_g) > p(abcd_c),\ p(abc_g) > p(abc_c),\ p(bcd_g) > p(bcd_c),\ p(ab_g) > p(ab_c),\ p(ab_g) > p(cd_c).$$

Результаты экспериментального исследования

В табл. 2 приведены результаты фильтрации слов для четырехсимвольных последовательностей из упомянутого выше текста по локальным компьютерным сетям. В данном тексте присутствует $|D_{rel}|$=25 терминов, встречающихся четыре и более раза. В качестве контрастного корпуса использовался набор статей из китайской газеты «Женьминь жибао» по темам «политика», «культура», «спорт», «происшествия» общим объемом около 480 тыс. иероглифов. Из выбранного текста с помощью предложенного алгоритма фильтрации было отобрано $|D_{retr}|$=46 слов, из которых $|D_{rel}| \cap |D_{retr}|$=20 оказались терминами предметной области.

N	Последовательность	Перевод	Тип последовательности
52	无线局域	беспроводной доступ в Интернет	термин
19	无线信号	радиосигнал	термин
19	千兆以太	гигабитный Интернет	
16	菲涅耳区	зона Френеля	термин
16	传输速率	скорость передачи данных	термин
13	线路由器	линия маршрутизатора	
13	无线路由	беспроводной маршрут	
13	传输距离	дальность передачи	термин
12	波束成形	формирование пучка	
12	无线接入	беспроводной доступ	термин
12	太网交换	коммутации Интернет	
12	多模光纤	многомодовое волокно	термин
11	内置电源	встроенный блок	термин
10	无线电波	радиоволны	термин
10	换式局域	локальная формула	
10	交换式局	коммутационный центр	термин
9	线接入点	линия точки доступа	
9	全向天线	всенаправленная антенна	термин
9	个发射机	один передатчик	
8	要供电的	к источнику питания	
8	换机端口	переключатель 8-портовый	
8	双工端口	дуплексный порт	термин
7	无线介质	беспроводная среда	термин
7	据链路层	захватывать канал данных	
7	功率类别	класс мощности	термин
6	线的增益	линия усиления	
6	端口带宽	пропускная способность порта	термин
6	的覆盖范	крышка вентилятора	

Таблица 2. Результаты извлечения терминов с использованием контрастной коллекции

N	Последовательность	Перевод	Тип последовательности
6	工端口带	*нет перевода*	
6	屏蔽双绞	экранированная витая пара	термин
6	太网链路	ссылка Ethernet	
6	口带宽为	*нет перевода*	
5	电缆衰减	затухание кабеля	термин
5	同轴电缆	коаксиальный кабель	термин
5	发射功率	мощность передачи	термин
5	享式以太	*нет перевода*	
5	不重叠的	не перекрывать	
4	送方和接	отправитель, а затем	
4	送或接收	отправить или получить	
4	这意味着	это означает, что	
4	过无线介	через беспроводную среду	
4	笔记本电	ноутбук	
4	相移键控	фазовая манипуляция	
4	的灵敏度	чувствительность	
4	天线增益	усиление антенны	термин
4	发送或接	отправить или получить	

Таблица 2 (продолжение). Результаты извлечения терминов с использованием контрастной коллекции

Из табл. 2 видно, что предложенный алгоритм фильтрации успешно устранил слова с притяжательным признаком 的 (аналогом родительного падежа) без явного указания его в качестве стоп-слова. При этом он не справился с сочетаниями иероглифов, которые могут быть интерпретированы как «гигабитный Интернет», «линия маршрутизатора», «беспроводной маршрут», «формирование пучка», «линия точки доступа» и другими, которые встречаются в целевом тексте, отсутствуют в контрастной коллекции, но не являются терминами данной предметной области.

Таким образом, точность извлечения $P = (|D_{rel}| \cap |D_{retr}|) / |D_{retr}| = 0{,}48$, а полнота $R = (|D_{rel}| \cap |D_{retr}|) / |D_{rel}| = 0{,}8$. Комбинированная F-метрика, определяемая по формуле $F = 2PR/(P+R) = 0{,}6$. В сравнении с существующими методами извлечения терминов для английских текстов [16], где комбинированная метрика F варьируется на разных текстах от 0,35 до 0,85, предлагаемый подход показывает вполне сопоставимые результаты.

Приведенные выше результаты показывают, что исключение из кандидатов в термины предметной области сочетаний иероглифов, части которых присутствуют в контрастной коллекции, позволяет извлекать не только часто используемые, но и достаточно редкие термины предметной области.

Заключение

В результате проведенного исследования экспериментальным путем была подтверждена реализуемость метода извлечения терминов предметной области из иероглифических текстов без предварительной сегментации фраз, который продемонстрировал результаты, сопоставимые по точности и полноте с двухфазной процедурой, состоящей из разбиения предложений на слова и последующего извлечения терминов. Дальнейшие исследования будут направлены на уточнение методов выявления бессмысленных сочетаний иероглифов, а также на более точное выявление терминологичности слов с применением методов, изложенных в [17], где для улучшения качества отбора терминов учитывается их совместная встречаемость в предложениях.

Литература

1. Joachims T. Learning to Classify Text Using Support Vector Machines: Methods, Theory and Algorithms. Kluwer

References

1. Joachims T. *Learning to Classify Text Using Support Vector Machines: Methods, Theory and Algorithms*. Kluwer

Academic Publishers, 2002. 205 p.
2. Wallach H.M. Topic modeling: beyond bag-of-words // Proc. 23rd Int. Conf. on Machine Learning. Pittsburgh, USA, 2006. P. 977–984.
3. Nugumanova A., Bessmertny I. Applying the latent semantic analysis to the issue of automatic extraction of collocations from the domain texts // Communications in Computer and Information Science. 2013. V. 394. P. 92–101. doi: 10.1007/978-3-642-41360-5_8
4. Тайваньские принципы сегментации текстов [Электронный ресурс] // http://ip194097.ntcu.edu.tw/TG/CompLing/hunsu/hunsu.htm (дата обращения 28.10.2016).
5. Xue N. Chinese word segmentation as character tagging // Computational Linguistics and Chinese Language Processing. 2003. V. 8. N 1. P. 29–48.
6. Zeng D., Wei D., Chau M., Wang F. Domain-specific Chinese word segmentation using suffix tree and mutual information // Information Systems Frontiers. 2011. V. 13. N 1. P. 115–125. doi: 0.1007/s10796-010-9278-5
7. Huang Lei, Wu Yan-Peng, Zhu Qun-Feng. Research and improvement of TFIDF feature weighting method // Computer Science. 2014. V. 41. N 6. P. 204–208.
8. Li Xiaochao, Zhao Shang, Lao Yan, Chen Min, Liu Mengmeng. Statistics law of same frequency words in Chinese texts and its application to keywords extraction // Application Research of Computers. V. 33. N 4. P. 1007–1012.
9. 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.
10. 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.
11. Penas A., Verdejo F., Gonzalo J. Corpus-based terminology extraction applied to information access // Proceedings of Corpus Linguistics. 2001. V. 2001. P. 458–465.
12. 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.
13. Basili R. A contrastive approach to term extraction // Proc. 4th Terminological and Artificial Intelligence Conference (TIA2001). Nancy, France, 2001.
14. 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.
15. 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.
16. Астраханцев Н.А. Автоматическое извлечение терминов коллекции предметной области с помощью Википедии // Труды ИСП РАН. 2014. Т. 26. № 4. С. 7–20. doi: 10.15514/ISPRAS-2014-26(4)-1
17. Нугуманова А.Б., Бессмертный И.А., Пецина П., Байбурин Е.М. Обогащение модели Bag-of-Words семантическими связями для повышения качества классификации текстов предметной области // Программные продукты и системы. 2016. № 2. С. 89–99. doi: 10.15827/0236-235X.114.089-099

Academic Publishers, 2002, 205 p.
2. Wallach H.M. Topic modeling: beyond bag-of-words. *Proc. 23rd Int. Conf. on Machine Learning*. Pittsburgh, USA, 2006, pp. 977–984.
3. Nugumanova A., Bessmertnyi I. Applying the latent semantic analysis to the issue of automatic extraction of collocations from the domain texts. *Communications in Computer and Information Science*, 2013, vol. 394, pp. 92–101. doi: 10.1007/978-3-642-41360-5_8
4. *Taiwanese Principles of Text Segmentation*. Available at: http://ip194097.ntcu.edu.tw/TG/CompLing/hunsu/hunsu.htm (accessed 28.10.2016).
5. Xue N. Chinese word segmentation as character tagging. *Computational Linguistics and Chinese Language Processing*, 2003, vol. 8, no. 1, pp. 29–48.
6. Zeng D., Wei D., Chau M., Wang F. Domain-specific Chinese word segmentation using suffix tree and mutual information. *Information Systems Frontiers*, 2011, vol. 13, no. 1, pp. 115–125. doi: 0.1007/s10796-010-9278-5
7. Huang Lei, Wu Yan-Peng, Zhu Qun-Feng. Research and improvement of TFIDF feature weighting method. *Computer Science*, 2014, vol. 41, no. 6, pp. 204–208.
8. Li Xiaochao, Zhao Shang, Lao Yan, Chen Min, Liu Mengmeng. Statistics law of same frequency words in Chinese texts and its application to keywords extraction. *Application Research of Computers*, vol. 33, no. 4, pp. 1007–1012.
9. 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.
10. 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, pp. 717.
11. Penas A., Verdejo F., Gonzalo J. Corpus-based terminology extraction applied to information access. *Proceedings of Corpus Linguistics*, 2001, vol. 2001, pp. 458–465.
12. 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.
13. Basili R. A contrastive approach to term extraction. *Proc. 4th Terminological and Artificial Intelligence Conference, TIA2001*. Nancy, France, 2001.
14. 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.
15. 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.
16. Astrakhantsev N.A. Automatic term acquisition from domain-specific text collection by using Wikipedia. *Trudy ISP RAN*, 2014, vol. 26, no. 4, pp. 7–20. (In Russian)
17. Nugumanova A.B., Bessmertnyi I.A., Petsina P., Baiburin E.M. Semantic relations in text classification based on Bag-of-words model. *Programmnye Produkty i Sistemy*, 2016, no. 2, pp. 89–99.

Авторы

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

Юй Чуцяо – аспирант, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, yuchuqiao0607@qq.com

Ма Пенюй – аспирант, Университет ИТМО, Санкт-Петербург, 197101, 175964911@qq.com

Authors

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

Yu Chuqiao – postgraduate, ITMO University, Saint Petersburg, 197101, Russian Federation, yuchuqiao0607@qq.com

Ma Pengyu – postgraduate, ITMO University, Saint Petersburg, 197101, Russian Federation, 175964911@qq.com