Mining Temporal Patterns of Technical Term Usages in Bibliographical Data

Hidenao Abe, Shusaku Tsumoto

Mining Temporal Patterns of Technical Term Usages in Bibliographical Data

Hidenao Abe[1], Shusaku Tsumoto[1]

Department of Medical Informatics, Shimane University, School of Medicine
89-1 Enya-cho, Izumo, Shimane 693-8501, Japan
abe@med.shimane-u.ac.jp, tsumoto@computer.org

Abstract. In text mining framework, data-driven indices are used as importance indices of words and phrases. Although the values of these indices are influenced by usages of terms, many conventional emergent term detection methods did not treat these indices explicitly. In order to detect research keys in academic researches, we propose a method based on temporal patterns of technical terms by using several data-driven indices and their temporal clusters. The method consists of an automatic term extraction method in given documents, three importance indices from text mining studies, and temporal patterns based on results of temporal clustering. Then, we assign abstracted sense of the temporal patterns of the terms based on their linear trends of centroids. Empirical studies show that the three importance indices are applied to the titles of four annual conferences about data mining field as sets of documents. After extracting the temporal patterns of automatically extracted terms, we compared the emergent patterns and one of the keyword of this article between the four conferences.

Keywords: Text Mining, Trend Detection, TF-IDF, Jaccard's Matching Coefficient, Temporal Clustering, Linear Regression

1 Introduction

In recent years, the accumulation of document data has been more general, according to the development of information systems in every field such as business, academics, and medicine. The amount of stored data has increased year by year. Document data includes valuable qualitative information to not only domain experts in the fields but also novice users on particular domains. However, detecting adequate important words or/and phrases, which are related to attractive topics in each field, is one of skilful techniques. Hence, the topic to support the detection has been attracted attentions in data mining and knowledge discovery fields. As for one solution to realize such detection, emergent term detection (ETD) methods have been developed [1, 2].

However, because the frequency of the words were used in earlier methods, detection was difficult as long as each word that became an object did not appear. These methods use particular importance index to measure the statuses of the words. Although the indices are calculated with the words appearance in each temporal set of documents, and the values changes according to their usages, most conventional methods do not consider the usages of the terms and importance indices separately. This causes difficulties in text mining applications, such as limitations on the extensionality of time direction, time consuming post-processing, and generality expansions. After considering these problems, we focus on temporal behaviors of importance indices of phrases and their temporal patterns.

In this paper, we propose an integrated for detecting temporal patterns of technical terms based on data-driven importance indices by combining automatic term extraction methods, importance indices of the terms, and trend analysis methods in Section 2. After implementing this framework as described in Section **??**, we performed an experiment to extract temporal patterns of technical terms. In this experiment, by considering the sets of terms extracted from the titles of four data mining relating conferences as examples, their temporal patterns based on three data-driven importance indices are presented in Section 3. With referring to the result, we discuss about the characteristic terms of the conferences. Finally, in Section 4, we summarize this paper.

2 An Integrated Framework for Detecting Temporal Patterns of Technical Terms based on Importance Indices

In this section, we describe a framework for detecting various temporal trends of technical terms as temporal patterns of each importance index consisting of the following three components:

1. Technical term extraction in a corpus
2. Importance indices calculation
3. Temporal pattern extraction

There are some conventional methods of extracting technical terms in a corpus on the basis of each particular importance index [2]. Although these methods calculate each index in order to extract technical terms, information about the importance of each term is lost by cutting off the information with a threshold value. We suggest separating term determination and temporal trend detection based on importance indices. By separating these phases, we can calculate different types of importance indices in order to obtain a dataset consisting of the values of these indices for each term. Subsequently, we can apply many types of temporal analysis methods to the dataset based on statistical analysis, clustering, and machine learning algorithms. An overview of the proposed method is illustrated in Figure 1.

First, the system determines terms in a given corpus. There are two reasons why we introduce term extraction methods before calculating importance indices. One is that the cost of building a dictionary for each particular domain

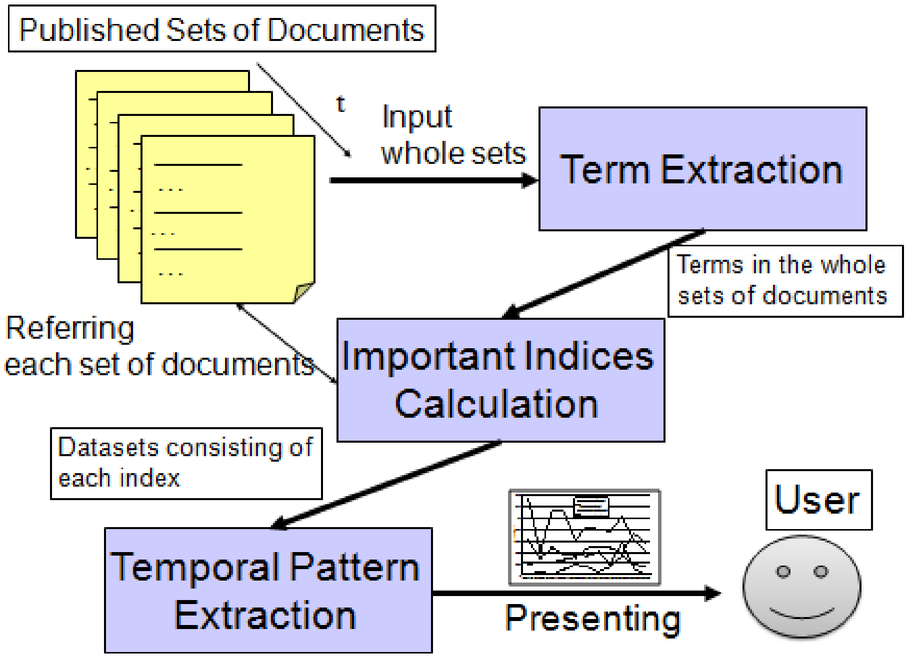

Fig. 1. An overview of the proposed remarkable temporal trend detection method.

is very expensive task. The other is that new concepts need to be detected in a given temporal corpus. Especially, a new concept is often described in the document for which the character is needed at the right time in using the combination of existing words.

After determining terms in the given corpus, the system calculates multiple importance indices of the terms for the documents of each period. Further, in the proposed method, we can assume the degrees of co-occurrence such as the χ^2 statistics for terms consisting of multiple words to be the importance indices in our method.

In the proposed method, we suggest treating these indices explicitly as a temporal dataset. The features of this dataset consist of the values of prepared indices for each period.

Figure 2 shows an example of the dataset consisting of an importance index for each year.

Then, the framework provides the choice of some adequate trend extraction method to the dataset. In order to extract useful temporal patterns, there are so many conventional methods as surveyed in the literatures [3, 4]. By applying an adequate time-series analysis method, users can find out valuable patterns by processing the values in rows in Figure 2.

Term	Jacc_1996	Jacc_1997	Jacc_1998	Jacc_1999	Jacc_2000	Jacc_2001	Jacc_2002	Jacc_2003	Jacc_2004	Jacc_2005
output feedback	0	0	0	0	0	0	0	0	0	0
H/sub infinity	0	0	0.012876	0	0.00885	0	0	0	0.005405	0.003623
resource allocation	0.006060606	0	0	0	0	0	0	0	0	0
image sequences	0	0	0	0	0	0	0	0.004785	0	0
multiagent systems	0	0	0	0	0	0	0.004975	0	0	0
feature extraction	0	0.005649718	0	0.004484	0	0	0	0	0	0
images using	0	0	0	0	0	0.004673	0	0	0	0
human-robot interaction	0	0	0	0	0.004425	0	0	0	0	0
evolutionary algorithm	0	0.005649718	0	0.004484	0	0	0	0	0.002703	0.003623
deadlock avoidance	0	0	0	0	0.004425	0	0	0	0	0
ambient intelligence	0	0	0	0	0	0	0	0	0	0.003623
feature selection	0	0	0	0	0	0	0	0	0.002703	0
data mining	0	0	0	0	0.004425	0	0	0	0.002703	0

Fig. 2. Example of a dataset consisting of an importance index.

3 Experiment: Extracting Temporal Patterns of Technical Terms by Using Temporal Clustering

In this experiment, we show the results temporal patterns by using the implementation of the method described in Section 2 and Section **??**. As the input of temporal documents, we used the annual sets of the titles of the following four academic conferences[1]; KDD, PKDD, PAKDD, and ICDM.

We determine technical terms by using the term extraction method [6][2] for each entire set of documents.

Subsequently, the values of tf-idf, Jaccard coefficient, and Odds are calculated for each term in the annual documents. To the datasets consisting of temporal values of the importance indices, we extract temporal patterns by using k-means clustering. Then, we apply the meanings of the clusters based on their linear trends calculated by the linear regression technique for the timeline.

3.1 Extracting technical terms

We use the titles of the four data mining related conferences as temporal sets of documents. The description of the sets of the documents is shown in Table 1.

As for the sets of documents, we assume each title of the articles to be one document. Note that we do not use any stemming technique because we want to consider the detailed differences in the terms.

By using the term extraction method with simple stop word detection for English, we extract technical terms as shown in Table 2. After merging all of titles of each conference into one set of the documents, these terms were extracted for each set of the titles.

3.2 Extracting temporal patterns by using k-means clustering

In order to extract temporal patterns of each importance index, we used k-means clustering. We set up the numbers of one percent of the terms as the maximum

[1] These titles are the part of the collection by DBLP [5].

[2] The implementation of this term extraction method is distributed in `http://gensen.dl.itc.u-tokyo.ac.jp/termextract.html` (in Japanese).

Table 1. Description of the numbers of the titles.

	KDD		PKDD		PAKDD		ICDM	
	# of titles	# of words	# of titles	# of words	# of titles	# of words	# of titles	# of words
1994	40	349						
1995	56	466						
1996	74	615						
1997	65	535	43	350				
1998	68	572	56	484	51	412		
1999	93	727	82	686	72	628		
2000	94	826	86	730	52	423		
2001	110	942	45	388	63	528	109	908
2002	140	1,190	43	349	62	515	121	1,036
2003	108	842	44	340	60	520	127	1,073
2004	133	1,084	64	504	83	698	105	840
2005	113	868	76	626	101	882	150	1,161
2006	139	1,068	67	497	128	1,159	317	2,793
2007	131	1,065	67	537	196	1,863	213	1,779
2008	134	1,126	110	832	136	1,224	264	2,225
TOTAL	1,498	12,275	783	6,323	1,004	8,852	1,406	11,815

Table 2. Description of the numbers of the extracted terms.

	KDD	PKDD	PAKDD	ICDM
# of extracted terms	3,232	1,653	2,203	3,033

number of clusters k for each dataset. Then, the system obtained the clusters with minimizing the sum of squared error within clusters. By iterating less than 500 times, the system obtains the clusters by using Euclidian distance between instances consisting of the values[3] of the same index.

Table 3 shows the result of the SSE of k-means clustering. As shown in this table, the SSE values of Jaccard coefficient are higher than the other two indices: tf-idf and odds. Since we were not selected the terms with two or more words, the values of Jaccard coefficient of the terms with just one word, which are 0 or 1, are not suitable to make clusters.

3.3 Details of a temporal pattern of the technical terms

As shown in Table@4, there are several kind of clusters based on the averaged linear trends. The centroid terms mean the terms that are the nearest location to the centroids. Then, by using the averaged degree and the averaged intercept of each term, we attempt to determine the following three trends:

[3] The system also normalized the values for each year.

Table 3. The sum of squared errors of the clustering for the technical terms in the titles of the four conferences.

Conf. Name	SSE (tf-idf)	SSE (Jaccard)	SSE (Odds)
KDD	46.71	689.44	8.87
PKDD	58.76	432.21	18.17
PAKDD	35.13	325.53	10.01
ICDM	21.05	286.91	4.93

- Popular
  - the averaged degree is positive, and the intercept is also positive.
- Emergent
  - the averaged degree is positive, and the intercept is negative.
- Subsiding
  - the averaged degree is negative, and the intercept is positive.

Since the terms assigned as the centroid have the highest FLR score in each pattern, the term is frequently used in the cluster by comparing to the other terms. As for the centroids of the degree and the intercept, they are the same as the average of each cluster, because the calculation of the centroid is assumed as the least-square method.

The emergent temporal patterns of the tf-idf index are visualized in Figure 3. According to the meanings based on the linear trend, the patterns #5,#6, and #8 of KDD have the emergent patterns. The emergent patterns that are #4 for PKDD, #1, #2, and #4 for PAKDD, and #4 for ICDM are also visualized.

Although these conferences share some emergent and subsiding terms based on the temporal patterns, characteristic terms can be also determined. The centroids of terms assigned as the emergent patterns[4] express the research topics that have attract the attentions of researchers.

The emergent terms in KDD, they are related to web data and graphs. As for PKDD, the phrases 'feature selection' determine as emergent phrases only for this conference. The mining techniques that are related to items and text are also determined in PAKDD and ICDM. These terms indicate some characteristics of these conferences, relating to people who have been contributed for each conference.

By comparing these patterns of the indices, we can understand not only the remarkable terms but also similarity and dissimilarity of the conferences.

[4] The emergent terms are emphasized in Table 4.

Table 4. Whole of the temporal patterns as the k-means clustering centroids on the three data-driven indices.

KDD	Cluster No.	tf-idf			Jaccard Coeffieicent			Odds		
		Term	Avg. Deg	Avg. Int	Term	Avg. Deg	Avg. Int	Term	Avg. Deg	Avg. Int
	1	sequence using data mining	0.007	0.039	graph mining	0.000	0.005	sequence using data mining	0.0000	0.0001
	2	data mining	0.759	15.346	*machine learning*	0.006	−0.021	mining	−0.0060	0.3012
	3	database mining	−0.088	1.271	databases	0.018	0.351	database mining	−0.0008	0.0085
	4	web usage mining	0.022	0.255	pattern discovery	0.002	0.014	web usage mining	0.0000	0.0004
	5	*web data*	0.094	−0.273	*graphs*	0.025	−0.026	*web data*	0.0001	−0.0003
	6	*relational data*	0.132	−0.444	latent	0.020	−0.054	*relational data*	0.0002	−0.0004
	7	web mining	−0.001	0.448	constraints	−0.003	0.122	web mining	0.0000	0.0014
	8	*graph mining*	0.140	−0.558	prediction models	0.007	−0.017	*graph mining*	0.0002	−0.0006
	9	bayesian network	−0.069	0.997	*interactive exploration*	0.025	−0.097	bayesian network	−0.0004	0.0049
	10	data streams	0.045	0.094	rule induction	−0.009	0.092	data streams	0.0001	0.0004
	11	knowledge discovery	0.519	4.485	predictive modeling	0.009	0.034	data mining	−0.0093	0.1430
	12	mining knowledge	−0.055	0.898	mining	0.022	0.627	mining knowledge	−0.0003	0.0030
	13	high-dimensional data	−0.029	0.798	data mining	−0.014	0.176	high-dimensional data	−0.0002	0.0039
	14	distributed data mining	−0.017	0.543	learning bayesian networks	−0.003	0.027	distributed data mining	−0.0001	0.0015
	15	data sets	0.354	1.565	scale space exploration	0.004	0.075	databases	0.0003	0.0185
	16				knowledge discovery	−0.008	0.137			
	17				efficient algorithms	−0.020	0.232			
	18				bayesian networks	−0.025	0.275			
	19				abstract	−0.005	0.128			
	20				categorical datasets	0.001	0.062			
PKDD	Cluster No.	tf-idf			Jaccard Coeffieicent			Odds		
		Term	Avg. Deg	Avg. Int	Term	Avg. Deg	Avg. Int	Term	Avg. Deg	Avg. Int
	1	classification learning	0.004	0.096	spatial data	0.002	0.004	classification learning	0.0000	0.0007
	2	knowledge discovery	−0.168	1.932	document collections	−0.033	0.262	data mining	−0.0136	0.1382
	3	data mining	−0.104	10.324	feature selection	−0.004	0.110	learning	−0.0017	0.1188
	4	*feature selection*	0.195	−0.559	learning	0.007	0.668	*pattern discovery*	0.0004	−0.0012
	5	spatial data	−0.116	1.195	supervised learning	−0.028	0.252	spatial data	−0.0007	0.0059
	6	data clustering	−0.062	0.840	applications	−0.013	0.268	data clustering	−0.0002	0.0027
	7	data streams	0.089	0.046	knowledge discovery	0.022	0.018	data analysis	0.0002	0.0017
	8	relational learning	0.041	0.735	rule discovery	0.006	0.067	databases	0.0000	0.0071
	9	web	0.073	0.270	data mining	−0.008	0.082	web	0.0002	0.0016
	10				time series	0.009	0.072			
PAKDD	Cluster No.	tf-idf			Jaccard Coeffieicent			Odds		
		Term	Avg. Deg	Avg. Int	Term	Avg. Deg	Avg. Int	Term	Avg. Deg	Avg. Int
	1	*hierarchical clustering based*	0.143	−0.230	text mining	0.001	0.003	hierarchical clustering based	0.0002	−0.0004
	2	data mining based	0.004	0.101	decision trees	0.010	0.034	data mining based	0.0000	0.0005
	3	mining association rules	−0.122	1.201	density-based clustering	−0.012	0.090	databases	−0.0006	0.0053
	4	*text classification*	0.220	−0.525	*machine learning*	0.011	−0.015	*text classification*	0.0002	−0.0005
	5	*frequent pattern mining*	0.263	−0.709	*association rules*	0.034	−0.079	*mining frequent*	0.0004	−0.0009
	6	mining structured association patterns	−0.044	0.949	data mining	0.004	0.003	knowledge discovery	−0.0006	0.0069
	7	data mining	1.365	3.439	*continuous features*	0.050	−0.137	data mining	0.0003	0.0272
	8	knowledge discovery	0.570	1.739	*databases*	0.033	−0.005	algorithm	−0.0052	0.0933
	9	clustering	2.382	8.213	mixed similarity measure	−0.032	0.283	clustering	−0.0012	0.1575
	10	text mining	−0.020	0.790	*rule extraction*	0.014	−0.036	text mining	−0.0003	0.0033
	11	data clustering	0.030	0.597	applications	−0.017	0.202	data clustering	−0.0001	0.0031
	12				model	0.073	0.092			
	13				sequential patterns	−0.037	0.267			
	14				clustering	0.011	0.782			
	15				feature selection	−0.015	0.260			
	16				bayesian classifiers	−0.008	0.167			
ICDM	Cluster No.	tf-idf			Jaccard Coeffieicent			Odds		
		Term	Avg. Deg	Avg. Int	Term	Avg. Deg	Avg. Int	Term	Avg. Deg	Avg. Int
	1	using data mining	0.070	−0.061	data clustering	0.003	0.003	using data mining	0.0001	0.0000
	2	data clustering	−0.271	1.847	feature selection	−0.045	0.398	data clustering	−0.0006	0.0035
	3	data mining approach	−0.154	1.407	sequence modeling	−0.027	0.211	medical data mining	−0.0004	0.0031
	4	*text mining*	0.332	−0.005	data mining	−0.017	0.112	text classification	0.0007	0.0013
	5	text classification based	0.206	0.010	*data streams*	0.051	−0.065	text classification based	0.0003	0.0000
	6	mining	0.476	23.930	text classification	0.013	0.004	data mining	−0.0001	0.0408
	7	web mining	−0.407	2.537	mining	0.010	0.814	web mining	−0.0009	0.0053
	8	data mining	0.284	8.233	event sequences	−0.079	0.450	mining	−0.0110	0.2294
	9	spatial data mining	0.085	0.468	link prediction	0.050	0.002	data mining approach	0.0001	0.0014
	10				association rules	0.037	0.426			
	11				change	0.011	0.102			

3.4 Visualizing the trend of a key word of this article in the different conferences

Figure 4 shows the trend of ‘text mining’, which is included in the titles of the different four conferences, by using the tf-idf values. Their tf-idf values are increased in around 2000 and 2007 respectively. The later peak is not observed in the titles of PKDD. The trend shows that the technical topics related to this term are different in each peak. Since a novel technique itself is attractive in the earlier period, the technique tends to apply other topics by using the technique in the later periods. The trend of ‘text mining’ also shows that the technique was paid attentions in the earlier period, and the technique was applied to the other objects such as stream data.

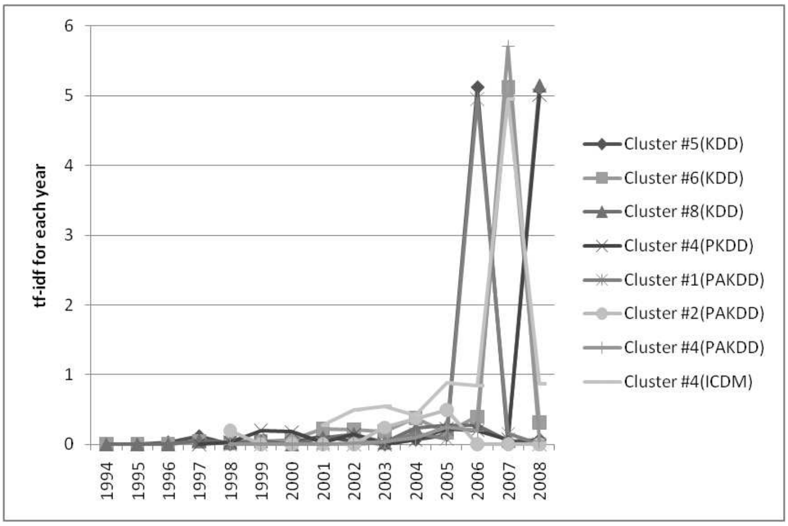

Fig. 3. The emergent temporal patterns of tf-idf through the four conferences.

4 Conclusion

In this paper, we proposed a framework to detect temporal patterns of the usages of technical terms appeared as the temporal behaviors of the importance indices. We implemented the framework with the automatic term extraction, the three importance indices, and temporal pattern detection by using k-means clustering.

The empirical results show that the temporal patterns of the importance indices can detect the trends of each term, according to their values for each annual set of the titles of the four academic conferences. Regarding the results, we detected not only the emergent temporal patterns in the conferences, but also the difference of the research topics between the conferences by comparing the temporal patterns and their representative terms. By focusing on the trend of one keyword of this article, 'text mining', we show the trend of this technical topic and the difference of the trends in the different conferences.

In the future, we will apply other term extraction methods, importance indices, and trend detection method. As for importance indices, we are planning to apply evaluation metrics of information retrieval studies, probability of occurrence of the terms, and statistics values of the terms. To extract the temporal patterns, we will introduce temporal pattern recognition methods [7], which can consider time differences between sequences with the same meaning. Then, we will apply this framework to other documents from various domains.

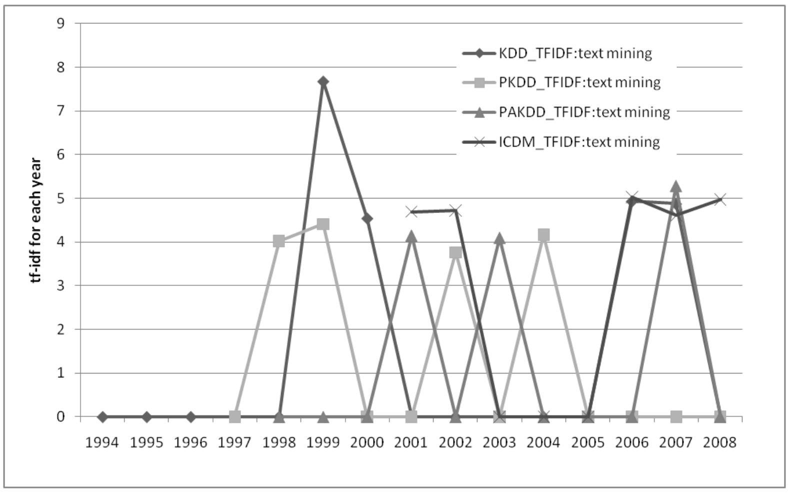

Fig. 4. The tf-idf values of ‘text mining’ in the titles of the four conferences.

References

1. Lent, B., Agrawal, R., Srikant, R.: Discovering trends in text databases. In: KDD ’97: Proceedings of the third ACM SIGKDD international conference on Knowledge discovery in data mining, AAAI Press (1997) 227–230
2. Kontostathis, A., Galitsky, L., Pottenger, W.M., Roy, S., Phelps, D.J.: A survey of emerging trend detection in textual data mining. A Comprehensive Survey of Text Mining (2003)
3. Keogh, E., Chu, S., Hart, D., Pazzani, M.: Segmenting time series: A survey and novel approach. In: an Edited Volume, Data mining in Time Series Databases., World Scientific (2003) 1–22
4. Liao, T.W.: Clustering of time series data: a survey. Pattern Recognition **38** (2005) 1857–1874
5. : The dblp computer science bibliography. `http://www.informatik.uni-trier.de/~ley/db/`
6. Nakagawa, H.: ”automatic term recognition based on statistics of compound nouns”. Terminology **6**(2) (2000) 195–210
7. Ohsaki, M., Abe, H., Yamaguchi, T.: Numerical time-series pattern extraction based on irregular piecewise aggregate approximation and gradient specification. New Generation Comput. **25**(3) (2007) 213–222