Proceedings of the 2009 IEEE International Conference on Systems, Man, and Cybernetics
San Antonio, TX, USA - October 2009

Detecting Temporal Patterns of Technical Phrases by using Importance Indices in a Research Documents

Hidenao Abe
Dept. of Medical Informatics
Shimane University
Izumo, Japan
abe@med.shimane-u.ac.jp

Shusaku Tsumoto
Dept. of Medical Informatics
Shimane University
Izumo, Japan
tsumoto@computer.org

***Abstract*—In text mining processes, temporal text mining have attracted considerable attention as an one of the important issues for finding remarkable terms with temporal patterns in temporal set of documents. Although importance indices of the technical terms play a key role in finding valuable patterns from various documents, temporal changes of them are not explicitly treated by conventional methods. Since those methods depend on particular index in each method, they are not robust in changes of terms. In order to detect remarkable temporal trends of technical terms in given textual datasets robustly, we propose a method based on temporal changes in several importance indices by assuming the importance indices of the terms to be a dataset. Our empirical study shows that two representative importance indices are applied to the documents from a research area. After detecting the temporal trends, we compared the emergent trend of the technical phrases to some emergent phrases given by a domain expert.**

***Index Terms*—Text Mining, Trend Detection, TF-IDF, Jaccard Coefficient, Linear Regression**

I. Introduction

In recent years, the development of information systems in every field such as business, academics, and medicine, and the amount of stored data have increased year by year. Accumulation is advanced to document data by not the exception but various fields. Document data provides valuable findings to not only domain experts in headquarter sections but also novice users on particular domains such as day trading, news readings and so on. Hence, the detection of new phrases and words has become very important. In order to realize such detection, emergent term detection (ETD) methods have been developed [1], [2].

However, because the frequency of the word was used in earlier methods, detection was difficult as long as the word that became an object did not appear. In addition, most conventional methods do not consider the nature of 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 changes of importance indices of phrases and their temporal patterns. Temporal change of the importance indices of extracted phrases is paid attention so that a specialist may recognize emergent terms and/or such fields. The detected terms and their patterns lead to capture human recognition behind the given documents.

In this paper, we propose a method for detecting trends of phrases by combining term extraction methods, importance indices of the terms, and trend analysis methods in Section III. Then, by considering the titles and abstracts of the IEEE Transaction of SMC [1] as an example, two kinds of temporal trends of extracted phrases based on two importance indices are presented in Section IV. With referring to the result, we discuss about the feasibility of the proposed method by comparing the phrases showing an emergent trend with the emergent technical terms provided by a domain expert in Section V. In this section, we also discuss the advantage of the computational complexity of the proposed framework. Finally, in Section VI, we summarize this paper.

II. Related Work

There exist some conventional studies on the detection of emergent terms/themes/topics in textual data. As the fist step, Lent et. al [1] proposed a method for finding temporal trends of words. Then, by applying various metrics such as frequency [3], n-gram [4], and tf-idf [5], researchers developed some emergent term detection (ETD) methods [2]. In [6] and [7], they suggested a method for finding emergent theme patterns on the basis of a finite state machine by using Hidden Markov Model (HMM) as one of the advanced ETD method. Topic modeling [8] is a related method from the viewpoint of temporal text analysis. In these methods, researchers consider the changes in each particular index of the terms rather than considering the nature of the terms in each language model.

Further, in the field of natural language processing, there are studies to find out meaningful terms in a document [9], [10]. One method to do so is based on χ^2 statistics of co-occurrence of nouns. Nakagawa [10] proposes a method of determining meaningful terms on the basis of adjacent frequency of compound nouns. By focusing on the methods for finding out meaningful terms consisting of two or more words on the basis of co-occurrence, Matsuo et. al [11] suggested a method for extracting technical terms consisting of co-existing nouns by calculating χ^2 statistics on a contingency matrix of occurrences of each pair of nouns in a given corpus.

[1] http://www.ieeesmc.org/

978-1-4244-2794-9/09/$25.00 ©2009 IEEE

In conventional studies on the detection of emergent words and/or phrases in documents such as Web pages and particular electronic message boards, researchers did not explicitly treat the trends of the calculated indices of words and/or phrases. However, on the basis of two different techniques, we consider a method for detecting temporal trends of phrases that consist of from two to nine words. We have focused on short phrases because a considerably long phrase may be a pattern including grammatical structure and anonymous words, as shown in [7].

III. Method of Detecting Trends of Technical Terms based on Importance Indices

In this section, we describe a method for detecting various temporal trends of technical terms by using multiple importance indices consisting of the following three subprocesses:

1) Technical term extraction in a corpus
2) Importance indices calculation
3) Trend detection

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 multiple 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 Fig.1.

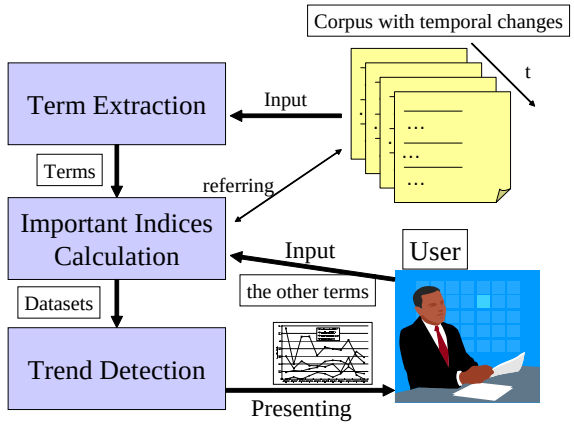

Fig. 1. An overview of the proposed remarkable temporal trend detection method.

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 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. Therefore, we apply a term extraction method that is based on the adjacent frequency of compound nouns. This method involves the detection of technical terms by using the following values for each candidate CN:

$$FLR(CN) = f(CN) \times (\prod_{i=1}^{L}(FL(N_i)+1)(FR(N_i)+1))^{\frac{1}{L}}$$

where $f(CN)$ means frequency of the candidates CN, and $FL(N_i)$ and $FR(N_i)$ indicate the frequencies of the right and the left of each noun N_i. In order to determine terms in this part of the process, we can also use other term extraction methods and terms/keywords from users.

After determining terms in the given corpus, the system calculates multiple importance indices of the terms for the documents of each period. As for importance indices of words and phrases in a corpus, there are some well-known indices. Term frequency divided by inversed document frequency (tf-idf) is one of the popular indices used for measuring the importance of the terms. tf-idf for each term t can be defined as follows:

$$TFIDF(t) = TF(t) \times log\frac{|D_{period}|}{DF(t)}$$

where $TF(t)$ is the frequency of each term t in the corpus with the documents included in each period as $|D_{period}|$. $DF(t)$ is the frequency of documents including t, which consists of the L words as shown in w_i. In the following experiments, we set up the threshold for FLR as $FLR > 1.0$ to extract terms from the whole set of documents.

As another importance index, we use Jaccard's matching coefficient [12] [2]. Jaccard coefficient can be defined as follows:

$$Jaccard(t) = \frac{DF(w_1 \cap w_2 \cap ... \cap w_L)}{DF(w_1 \cup w_2 \cup ... \cup w_L)}$$

where $DF(w_i)$ means the number of hit documents in the corpus D_{period} for the word w_i. Each value of Jaccard coefficient shows strength of co-occurrence of multiple words as an importance of the terms in the given corpus. 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. Fig.2 shows an example of the dataset consisting of an importance index for each year.

Then, the method provides the choice of some adequate trend extraction method to the datasets. In the following case study, we apply the linear regression analysis technique in order to detect the degree of existing trends based on the two

[2]Hereafter, we refer to this coefficient as "Jaccard coefficient".

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	0.006006006	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.

importance indices. The degree of each term t is calculated as the following:

$$Deg(t) = \frac{\sum_{i=1}^{M}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{M}(x_i - \bar{x})^2}$$

where $\bar{x}$ is the average of the M time points, and $\bar{y}$ is the average of each importance index for the period. Simultaneously, we calculate the intercept $Int(t)$ of each term t as follows:

$$Int(t) = \bar{y} - Deg(t)\bar{x}$$

IV. Experiment: Detecting Remarkable Trends of Technical Phrases in Two Temporal Sets of Documents

In this experiment, we show the results of detecting two trends by using the method described in Section III. As the input of temporal documents, annual sets of titles and abstracts of TSMC from 1996 to 2006 are taken.

We determine technical terms by using the term extraction method [10][3] for each entire set of documents.

Subsequently, the values of tf-idf and Jaccard coefficient are calculated for each term in the annual documents on each corpus. To the datasets consisting of temporal values of the importance indices, we apply linear regression to detect the following two temporal trends of the phrases: Emergent and Subsiding.

A. Extracting technical terms

We use the titles and the abstracts of all parts (A, B, and C) of the IEEE Transactions on SMC from 1996 to 2006 as the corpus. The description of the corpus is shown in TABLEI.

TABLE I
DESCRIPTION OF THE NUMBERS OF THE TSMC CORPUS.

Year	Abstract		Title	
	# of documents	# of words	# of documents	# of words
1996	165	22,271	165	1,510
1997	177	24,884	177	1,697
1998	233	32,633	233	2,290
1999	223	31,714	223	2,150
2000	226	32,083	226	2,231
2001	214	30,434	214	2,119
2002	201	30,366	201	2,231
2003	209	33,213	209	2,065
2004	370	58,876	370	3,773
2005	276	43,504	276	2,664
2006	139	22,975	139	1,410
TOTAL	2,433	362,953	2433	23,852

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

As for the sets of documents, we assume each title and abstract 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 52,235 terms from the TSMC abstracts. Similarly, 5,511 terms are extracted from the TSMC titles.

B. Results for automatically extracted terms

By using the degree and the intercept of each term, we attempt to determine the following two trends:

- Emergent
  - sorting the degree in ascending order
  - sorting the intercept in descending order
- Subsiding
  - sorting the degree in descending order
  - sorting the intercept in ascending order

Fig.3 and Fig.4 shows the top phrases of the two trends based on the two importance indices respectively.

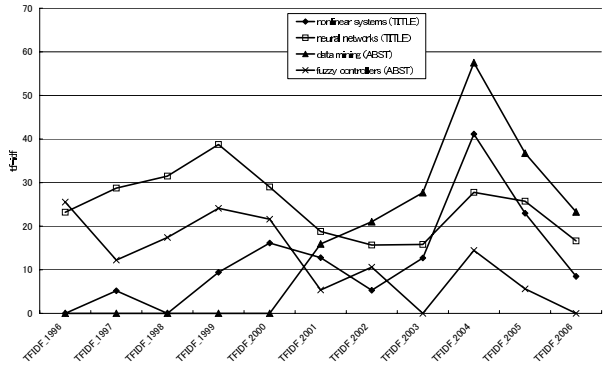

Fig. 3. Example of the trends based on tf-idf values in the TSMC corpus.

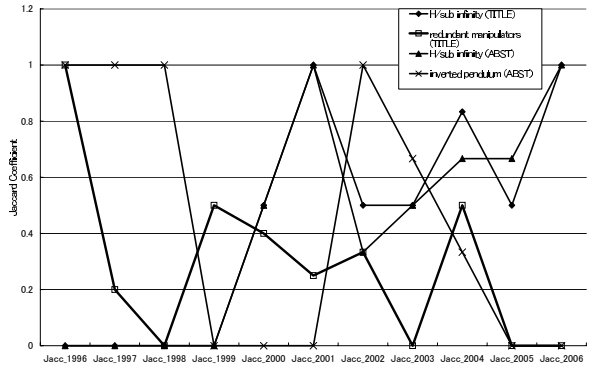

Fig. 4. Example of the trends based on Jaccard coefficient values in the TSMC corpus.

As shown in these charts, the trends of the extracted phrases are represented by an importance index. According to the degrees and the interruptions of each importance index calculated by using linear regression, we identify both emergent phrases and subsiding phrases in each temporal set of documents.

TABLEII shows top ten terms extracted from the TSMC abstracts, these tables show the two trends based on the two importance indices. The top ten terms extracted from the TSMC titles are shown in TABLEIII.

The phrases selected on the basis of the degree and the averaged values of each importance index are very similar to each other in the TSMC abstracts and the TSMC titles.

The difference between the two sets of documents is related to the object of each article because of the difference between the phrases detected by the two important indices. In the TSMC abstracts and the TSMC titles, the two different importance indices detected almost the same phrases for both the emergent and the subsiding trends. Therefore, we should select an appropriate importance index to the given temporal sets of documents.

In order to extract the other remarkable trends in the temporal changes of the importance indices, we should consider their statistical features such as average, minimum, and maximum values.

V. Discussion

A. Comparing the above results to the emergent terms given by a domain expert

In order to compare the trends based on the importance indices with a human criterion, we consulted emergent technical terms/research areas provided by a domain expert. He provided 20 keywords related to his emergent research areas.

TABLEIV shows the result of the degrees and the intercepts of the eleven years' importance indices values for the keywords.

As shown in TABLEIV, our method detected more than a half of the keywords appeared in each set of documents as 'Emergent'. The trends detected by the two importance indices represent the same trends in the same sets of documents respectively. Comparing the degrees of each keyword on the titles and the abstracts, the trend of 'autonomous robotics' is different in each set of documents. The computation time for these keywords is linearly decreased, because it only depends on the number of input terms in this method.

However, the extracted technical terms do not cover every term that is provided by the domain expert. We improve this term extraction phrase by introducing other term extraction criteria, such as χ^2 statistics, Web based keyword expansion [13], [14].

B. Comparison of time complexity for post-processing

Focusing on the post-process to interpret the output from ETD methods, the proposed method reduces the time complexity dramatically. The worst order of the post-processing of ETD methods based on trends of words is $O(2^n)$, where n is the number of words. Since a domain expert should interpret each meaning of the combinations between emergent words and the other words, the cost of the post-processing increases exponentially. In contrast, by using the proposed method, a domain expert can interpret the trends of m phrases, which were determined by a term extraction method. The order of time complexity is $O(m)$. Usually, the number of extracted terms m is very smaller than the number of combinations of the n words. If there is no suitable phrase in the extracted phrases, the user can get the trend of multiple importance indices of a phrase $O(1)$ by assuming the time complexity depends only on the human interpretation procedure. Alternatively, the user can find similar phrases in the candidates, and then the user may revise the given corpus to another one.

VI. Conclusion

In this paper, we proposed the method to detect trends of technical terms by focusing on the temporal changes of the importance indices. We implemented the method by combining the technical term extraction method, the two important indices, and linear regression analysis.

The case study shows that the temporal changes of the importance indices can detect the trend of each term, according to the degree of the values for each annual document. The emergent terms, which detected by a domain expert, are ranked as the terms with increasing degrees of the importance indices. Regarding to the result, our method can support to find out trends of terms in documents based on the temporal changes of the importance indices.

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 trends, we will introduce temporal pattern recognition methods, such as temporal clustering. Then, we will apply this framework to other documents from various domains.

Acknowledgement

We would like to thank Prof. Hideyuki Takagi, who provided the documents of IEEE Transactions Systems, Man, and Cybernetics and this opportunity to analyze research trends in textual data. We also thank Dr. Stuart Rubin, who provided the emergent keywords in his research areas as a SMC TC Chair.

References

[1] B. Lent, R. Agrawal, and R. Srikant, "Discovering trends in text databases." AAAI Press, 1997, pp. 227–230.

[2] A. Kontostathis, L. Galitsky, W. M. Pottenger, S. Roy, and D. J. Phelps, "A survey of emerging trend detection in textual data mining," *A Comprehensive Survey of Text Mining*, 2003.

[3] R. Swan and J. Allan, "Automatic generation of overview timelines," in *SIGIR '00: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*. New York, NY, USA: ACM, 2000, pp. 49–56.

[4] C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, pp. 379–423, 623–656, 1948.

[5] K. Sparck Jones, "A statistical interpretation of term specificity and its application in retrieval," *Document retrieval systems*, pp. 132–142, 1988.

[6] J. M. Kleinberg, "Bursty and hierarchical structure in streams," *Data Min. Knowl. Discov.*, vol. 7, no. 4, pp. 373–397, 2003.

TABLE II
TOP 10 TERMS FOR TWO TRENDS BASED ON TF-IDF AND JACCARD COEFFICIENT VALUES IN THE TSMC ABSTRACTS.

(a) Emergent

	tf-idf			Jaccard coeffient		
Rank	t	Deg(t)	Int(t)	t	Deg(t)	Int(t)
1	data mining	4.657	−6.726	H/sub infinity	0.095	−0.053
2	fuzzy temporal	2.651	−6.595	facial expressions	0.072	−0.055
3	multiresolution learning	2.592	−6.328	supply chain	0.048	0.045
4	experimental results	2.553	16.380	facial expression	0.044	−0.106
5	evolutionary algorithm	2.514	−3.451	detailed description	0.041	−0.081
6	quality of	2.491	8.453	shop floor	0.037	−0.011
7	H/sub infinity	2.476	−2.491	breast cancer	0.036	0.045
8	decision tree	2.455	1.135	fault tolerance	0.030	−0.038
9	data sets	2.146	4.564	Web services	0.027	−0.052
10	results demonstrate	2.059	4.222	cellular automata	0.026	−0.035

(b) Subsiding

	tf-idf			Jaccard coeffient		
Rank	t	Deg(t)	Int(t)	t	Deg(t)	Int(t)
1	fuzzy controllers	−2.020	22.552	inverted pendulum	−0.079	0.848
2	expert systems	−1.772	14.081	traveling salesman	−0.073	1.000
3	blackboard system	−1.625	11.860	sliding mode	−0.042	0.566
4	fuzzy systems	−1.600	30.280	obstacle avoidance	−0.037	0.445
5	decision boundaries	−1.567	13.149	Kalman filter	−0.031	0.536
6	priori knowledge	−1.560	15.041	Data mining	−0.028	0.279
7	supervised learning	−1.523	13.729	mean square	−0.027	0.246
8	Petri nets	−1.520	31.265	risk assessment	−0.025	0.200
9	information processing	−1.498	12.787	Markov chain	−0.024	0.268
10	systems approach	−1.483	11.861	World Wide Web	−0.023	0.345

[7] Q. Mei and C. Zhai, "Discovering evolutionary theme patterns from text: an exploration of temporal text mining," in *KDD '05: Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*. New York, NY, USA: ACM, 2005, pp. 198–207.

[8] H. M. Wallach, "Topic modeling: beyond bag-of-words," in *ICML '06: Proceedings of the 23rd international conference on Machine learning*. New York, NY, USA: ACM, 2006, pp. 977–984.

[9] K. T. Frantzi and S. Ananiadou, "Extracting nested collocations," in *Proceedings of the 16th conference on Computational linguistics*. Morristown, NJ, USA: Association for Computational Linguistics, 1996, pp. 41–46.

[10] H. Nakagawa, ""automatic term recognition based on statistics of compound nouns"," *Terminology*, vol. 6, no. 2, pp. 195–210, 2000.

[11] Y. Matsuo and M. Ishizuka, "Keyword extraction from a single document using word co-occurrence statistical information," *International Journal on Artificial Intelligence Tools*, vol. 13, no. 1, pp. 157–169, 2004.

[12] M. R. Anderberg, *Cluster Analysis for Applications*, ser. Monographs and Textbooks on Probability and Mathematical Statistics. New York: Academic Press, Inc., 1973.

[13] "Google sets:," http://labs.google.com/sets.

[14] R. C. Wang and W. W. Cohen, "Language-independent set expansion of named entities using the web," in *ICDM '07: Proceedings of the 2007 Seventh IEEE International Conference on Data Mining*. Washington, DC, USA: IEEE Computer Society, 2007, pp. 342–350.

TABLE III
TOP 10 TERMS FOR THE TWO TRENDS BASED ON TF-IDF AND JACCARD COEFFICIENT VALUES IN THE TSMC TITLES.

(a) Emergent

	tf-idf			Jaccard coeffient		
Rank	t	Deg(t)	Int(t)	t	Deg(t)	Int(t)
1	nonlinear systems	2.116	1.616	H/sub infinity	0.095	-0.038
2	H/sub infinity	1.273	-1.178	output feedback	0.060	-0.029
3	output feedback	1.236	-0.961	time series	0.055	-0.053
4	recognition using	0.965	-2.043	simulated annealing	0.052	-0.011
5	deadlock avoidance	0.830	-0.896	image sequences	0.034	-0.036
6	human-robot interaction	0.826	-0.979	human-robot interaction	0.031	-0.005
7	uncertain nonlinear systems	0.790	-0.509	resource allocation	0.029	0.037
8	data mining	0.749	-0.073	deadlock avoidance	0.029	-0.002
9	pattern classification	0.742	-0.123	traveling salesman	0.027	0.227
10	resource allocation	0.707	-0.076	sliding mode	0.021	0.242

(b) Subsiding

	tf-idf			Jaccard coeffient		
Rank	t	Deg(t)	Int(t)	t	Deg(t)	Int(t)
1	neural networks	-1.048	29.928	redundant manipulators	-0.049	0.533
2	fuzzy logic	-1.015	17.240	mobile robot	-0.036	0.393
3	mobile robot	-0.892	13.034	fault diagnosis	-0.030	0.438
4	Petri nets	-0.816	17.628	function approximation	-0.022	0.267
5	neural network	-0.725	28.025	associative memory	-0.018	0.530
6	function approximation	-0.646	6.903	neural network	-0.016	0.367
7	learning automata	-0.629	8.148	genetic algorithms	-0.014	0.245
8	systems engineering	-0.614	7.328	membership functions	-0.013	0.210
9	dynamic environments	-0.607	4.809	neural networks	-0.011	0.379
10	fuzzy controllers	-0.538	7.187	learning automata	-0.010	0.125

TABLE IV
SET OF EMERGENT KEYWORDS DETECTED BY A DOMAIN EXPERT AND THEIR TRENDS BASED ON THE TWO IMPORTANCE INDICES.

	Titles				Abstracts			
	tf-idf		Jaccard coeffient		tf-idf		Jaccard coeffient	
Keyword	Deg(t)	Int(t)	Deg(t)	Int(t)	Deg(t)	Int(t)	Deg(t)	Int(t)
associative memory	-0.241	5.253	-0.018	0.530	-0.009	7.782	-0.010	0.285
fault diagnosis	-0.137	5.605	-0.030	0.438	0.132	7.995	0.007	0.149
Decision support	-0.052	1.239	-0.012	0.114	0.393	-0.673	0.005	-0.009
parallel solution	0.224	-0.673	0.015	-0.045				
autonomous robotics	0.224	-0.673	0.023	-0.068	-0.232	1.625	-0.004	0.029
knowledge-based robotic plan execution	-0.188	1.412	-0.009	0.068				
case-based condition assessment	0.000	0.488	0.000	0.045				
Probabilistic techniques	0.000	0.488	0.000	0.091				
man-machine interfaces	-0.098	0.983	-0.018	0.182	-0.098	0.983	-0.006	0.061
Human-robot interactions	0.097	0.000	0.018	0.000				
Sensor explication	-0.188	1.412	-0.036	0.273				
Multicriteria meta-heuristics	0.204	-0.511	0.018	-0.045				
feature selection	0.683	0.300	0.006	0.099	1.351	3.185	0.003	0.051
simultaneous feature selection	0.224	-0.673	0.011	-0.034				
heuristic method	0.224	-0.673	0.045	-0.136				
Knowledge-based fast evaluation	0.204	-0.511	0.012	-0.030				
possibilistic uncertainty	0.161	-0.269	0.014	-0.023				
Financial prediction	-0.149	1.239	-0.009	0.076				
Zadeh's paradigm of computing	0.161	-0.269	0.000	0.000	0.161	-0.269	0.000	0.000
automatic diagnosis	0.097	0.000	0.005	0.000	0.097	0.000	0.003	0.000
Emergent (Deg(t)>0)	13		13		5		5	
Subsidint(Deg(t)<0)	7		7		3		3	