Chapter 2

Analysis for Finding Innovative Concepts Based on Temporal Patterns of Terms in Documents

Hidenao Abe

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/52210

1. Introduction

In recent years, information systems in every field have developed rapidly, and the amount of electrically stored data has increased day after day. Electrical document data are also stored in such systems mainly for recording and for holding the facts. As for the medical field, documents are also accumulated not only in clinical situations, but also in worldwide repositories by various medical studies. Such data now provide valuable information to medical researchers, doctors, engineers, and related workers by retrieving the documents depending to their expertise. They want to know the up-to-date knowledge for providing better care to their patients. Hence, the detection of novel, important, and remarkable phrases and words has become very important to aware valuable evidences in the documents. However, the detection is greatly depending on their skills for finding the good evidences.

Besides, with respect to biomedical research documents, the MeSH [1] vocabulary provides overall concepts and terms for describing them in a simple and an accurate way. The structured vocabulary is maintained by NIH for reflecting some novel findings and interests on each specific field, considering amount of published documents and other factors based on the studies. Through such consideration, new concepts, which appear as new concepts every year, are usually added to the vocabulary if the concepts are useful. One criterion for adding new concepts is related to how attention paid to them by the researchers appears as an emergent pattern in published documents. As the fact, around few hundred of new concepts are added every year, and the maintenance of the concepts and their related structure has been done by manually. Thus, MeSH has another aspect as an important knowledge base for the biomedical research field. However, the relationships between particular data-driven trends and the newly added concepts did not be clarified.

By clarifying the relationship between such maintained vocabulary and the trends of term usages, readers in the field can detect the important terms for understanding the up-to-date trends in his/her field more clearly. Under the above-mentioned motivation, I developed a

method for analyzing the similarity of terms on the structured taxonomy and the trend of a data-driven index of the terms [2]. In this chapter, I describe a result of the analysis by using the method for identifying similar terms based on the temporal behavior of usages of each term. The temporal pattern extraction method on the basis of term usage index consists of automatic term extraction methods, term importance indices, and temporal clustering in the next section. Then, in Section 4, a case study is carried out for showing the differences between similar terms detected by the temporal patterns of medical terms related to migraine drug therapy in MEDLINE documents. Finally, I conclude the analysis result in Section 6.

2. The method for analyzing distances on taxonomy and temporal patterns of term usage

In this section, I describe a method for detecting various trends of words and phrases in temporally published corpora. In order to analyze the relationships between usages of words and phrases in temporally published documents and the difference on a taxonomy, this trend detection method use a temporal pattern extraction method based on data-driven indices [2]. By using the similar terms identified on the basis of temporal patterns of the indices, the method measures their similarities between each term on the taxonomy that can be assumed as the sets of tree structures of concepts on a particular domain. One of the reasons why the method uses temporal patterns is for detecting various trends. The important aim of the method is not only detecting particular trend that a user set, but also detecting various trends based on the nature of the given corpora. The other reason is for finding representing terms, called as 'keywords' in each specific field, on the basis of the trends. Considering these two aims, the method uses the temporal pattern extraction process based on temporal behaviors of terms by measuring an importance index.

Then, on the basis on the temporal behavioral similarity of each index, the distances between the similar terms, which are the members of each temporal pattern, are measured. By using the distance on the structured vocabulary, the averages of the distances between the terms included in temporal patterns is compared for analyzing the relationship between the trends of temporal patterns and the similarities of terms on the vocabulary.

In the following sections, the method for detecting temporally similar terms based on each importance index is described firstly. Subsequently, the distance measure on the structured vocabulary is explained.

2.1. Obtaining temporal patterns of data-driven indices related to term usages

In order to discover various trends related to usages of the terms in temporally published corpus, the framework [2] is developed as a method for obtaining temporal patterns of an importance index. This framework obtains some temporal patterns based on the importance index from the given temporally published sets of documents. It consists of the following processes.

- Automatic term extraction in overall documents
- Calculation of importance indices
- Obtaining temporal clusters for each importance index
- Assignment of some meanings for the obtained temporal patterns

http://dx.doi.org/10.5772/52210

2.1.1. Automatic term extraction in a given corpus

Firstly, a system determines terms in a given corpus. Considering the difficulties of constructing particular dictionaries on each domain, term extraction without any dictionary is required. As for the representative method for extracting terms automatically, a term extraction method [3] based on the adjacent frequency of compound nouns is selected. This method involves the detection of technical terms by using the following values for a candidate compound noun CN:

$$FLR(CN) = f(CN) \times (\prod_{i=1}^{L}(FL(N_i)+1)(FR(N_i)+1))^{\frac{1}{2L}}$$

where $f(CN)$ means frequency of a candidate compound noun CN separately, and $FL(N_i)$ and $FR(N_i)$ indicate the frequencies of different words on the right and the left of each noun N_i in *bi*-grams included in each CN. Each compound noun CN is constructed $L(L \geq 1)$, one or more, nouns.

For example, there is a set of compound nouns $S = \{data\ mining, text\ mining, mining\ method\}$ from a corpus, and they appear just one time in the corpus. Then, we want to now the FLR score of *datamining*, $FLR(data\ mining)$. The left frequency of 'data' is 0, because of $FL(data) = 0$. The right frequency of 'data' is 1, because 'mining' appears just one time on the right of 'mining'. So $FR(data)$ is 1. As the same way, the frequencies of 'mining' are $FL(mining) = 2$ and $FR(mining) = 1$. Then, the $FLR(data\ mining)$ is calculated as follows.

$$FLR(data\ mining) = 1 \times \sqrt{(0+1)(1+1) \times (2+1)(1+1)} = 3.464\cdots$$

2.1.2. Calculation of data-driven indices for each term in each set of documents

After determining terms in the given corpus, the system calculates importance indices of these terms in the documents in each time period for representing the usages of the terms as the values. For the temporally published corpora, users can set up a period optionally. Most of the cases, the period is set up yearly, monthly, and daily, because the published documents are given timestamps. In this framework, each set of documents, that are published in each period, is denoted as D_{period}.

Some importance indices for words and phrases in a corpus are well known. Term frequency divided by inverse document frequency (tf-idf) is one of the popular indices used for measuring the importance of terms [4]. The tf-idf value for each term $term_i$ can be defined for the documents in each period, D_{period}, as follows:

$$TFIDF(term_i, D_{period}) = tf(term_i, D_{period}) \times log\frac{|D_{period}|}{df(term_i, D_{period})}$$

where $tf(term_i, D_{period})$ is the frequency of each term $term_i$ in a corpus with $|D_{period}|$ documents. Here, $|D_{period}|$ is the number of documents included in each period, and $df(term_i, D_{period})$ is the frequency of documents containing term.

In the proposed framework, the method suggests treating these indices explicitly as a temporal dataset. This dataset consists of the values of the terms for each time point by using each index $Index(\cdot, D_{period})$ as the features. Figure 1 shows an example of such a dataset consisting of an importance index for each period. The value of the term $term_i$ is described as $Index(term_i, D_{period})$ in Figure 1.

Selecting m time periods as the features for the dataset

	⋯	$Index(\bullet, D_{2000})$	⋯	$Index(\bullet, D_{year})$	⋯
$term_1$	⋯	$Index(term_1, D_{2000})$	⋯	$Index(term_1, D_{year})$	⋯
⋮		⋮		⋮	
$term_i$	⋯	$Index(term_i, D_{2000})$	⋯	$Index(term_i, D_{year})$	⋯
⋮		⋮		⋮	
$term_n$	⋯	$Index(term_n, D_{2000})$	⋯	$Index(term_n, D_{year})$	⋯

Figure 1. Example of dataset consisting of an importance index.

2.1.3. Generating temporal patterns by using temporal clustering

After obtaining the dataset, the framework provides the choice of an adequate trend extraction method to the dataset. A survey of the literature shows that many conventional methods for extracting useful time-series patterns have been developed [5, 6]. Users can apply an adequate time-series analysis method and identify important patterns by processing the values in the rows of Figure 1. By considering these patterns with temporal information, users can understand the trends related to the terms such as transition of technological development with technical terms. The temporal patterns as the clusters also provide information about similarities between the terms at the same time. The system denotes the similar terms based on the temporal cluster assignments as $term_i \in c_k$.

2.1.4. Assigning meanings of the trends of the obtained temporal patterns

After obtaining the temporal patterns c_k, in order to identify the meanings of each pattern by using trends of the extracted terms for each importance index, the system applies linear regression analysis. The degree of the centroid of a temporal pattern c is calculated as follows:

$$Deg(c) = \frac{\sum_{j=1}^{M}(c_j - \bar{c})(x_j - \bar{x})}{\sum_{j=1}^{M}(x_j - \bar{x})^2}$$

http://dx.doi.org/10.5772/52210

where $\bar{x}$ is the average of $t_j - t_1$ for M time points and $\bar{y}$ is the average of the values c_j. Each value of the centroid, c_j, is a representative value of the importance index values of assigned terms in the pattern as $Index(term_i \in c_k, D_{period})$. Each time point t_j corresponds to each period, and the first period assigns to the first time point as t_1.

Simultaneously, the system calculates the intercept $Int(c)$ of each pattern c_k as follows:

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

Then, by using the two linear trend criteria, users assigne some meanings of the temporal patterns related to the usages of the terms.

2.2. Defining similarity of terms on a structured taxonomy

In this chapter, a tree structure of concepts that are defined with a relation such as is-a as 'structured taxonomy' is used. In the biomedical domain, MeSH (Medical Subjects Headings) [1] is one of the important structured taxonomy for representing key concepts of biomedical research articles. MeSH consists of 16 categories including not only proper categories for biomedicine but also general categories such as information science. It contains 25,588 concepts as 'Descriptor', and 464,282 terms as 'Entry Terms' in the version of 2001. Each concept has one or more entry terms and the tree numbers as the identifier in the hierarchy structure.

For this structure, the similarity of each pair of terms represented by using distance in the tree structure of MeSH is defined, as shown in Figure 2.

For example, when the distance between the two terms, $term_{i1}$ and $term_{i2}$, denotes as $Dist(term_{i1}, term_{i2})$, the distance between 'migraine' and 'sharp headache', $Dist(migraine, sharpheadache)$, is calculated as 8 or 9. By using this distance, the similarity between each pair of terms is defined as the following:

$$Sim(term_{i1}, term_{i2}) = \frac{1}{1 + Dist(term_{i1}, term_{i2})}$$

where the similarity can be calculated when the both terms have tree numbers in MeSH.

For overall terms belonging to some group g, representative values are also defined their averaged similarity in the group as the following:

$$Avg.Sim(g) = \frac{1}{numPair} \sum_{term_i \in g} Sim(term_{i1}, term_{i2})$$

where $numPair$ is the number of matched pairs of the terms included in the group g. The definition is $numPair =_m C_2$, where the number of appeared terms m is $m = |term_i \in g \cap hasTreeNumber(term_i)|$.

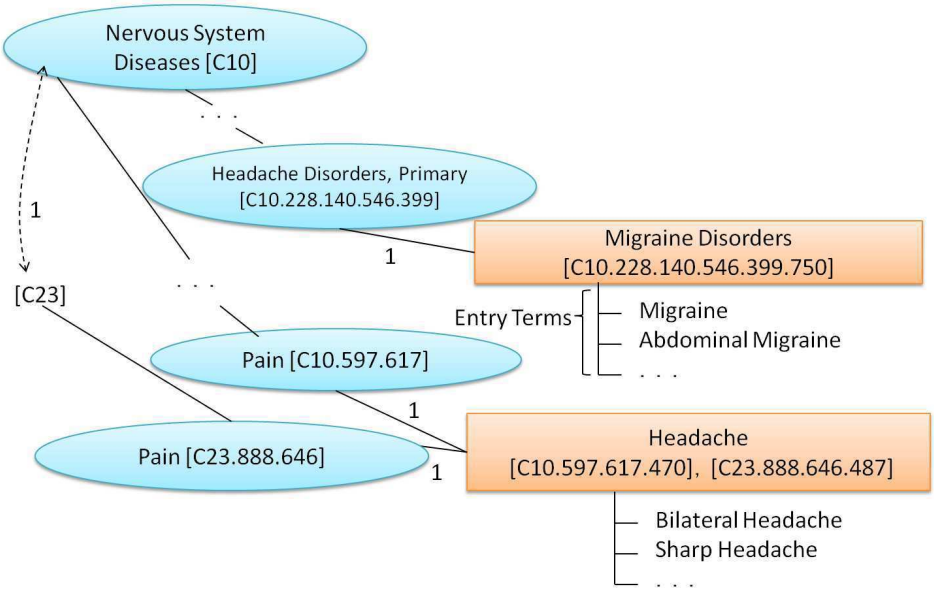

Figure 2. Example of MeSH hierarchy structure for migraine disorders and headache.

3. A case study for detecting trends of terms by obtaining temporal patterns

In this section, I describe a case study for analyzing similarity of terms detected some temporal patterns in medical research documents. For obtaining the temporal patterns, I used an importance index of the terms in each set of documents that ware published year by year. The medical research documents are retrieved from MEDLINE by using a search scenario over time. The scenario is related to migraine drug therapy similar to the first one in a previous paper on MeSHmap [7].

In this case study, I consider the search scenario and three meanings of trends as temporal clusters by using the degrees and intercepts of the trend lines for each term as follows. As for the meanings, the following two trends are assigned; "emergent" to ascending trend lines with negative intercepts, "subsiding" to decending trend lines with positive intercepts, and "popular" to ascending trend lines with positive intercepts.

3.1. Analysis of a disease over time

In this scenario, a user may be interested in exploring the progression of ideas in a particular domain, say, corresponding to a particular disease. By performing the search such that the disease is represented according to the year, one may obtain a temporal assessment of the changes in the field.

Let us assume that the user wants to explore the evolution of ideas about drugs used to treat chronic hepatitis. The user performs a search for abstracts of articles "chronic hepatitis/drug

http://dx.doi.org/10.5772/52210

therapy [MH:NOEXP] AND YYYY [DP] AND clinical trial [PT] AND english [LA]" through PubMed. The string "YYYY" is replaced with the four digits necessary for retrieving articles published each year. The retrieval for PubMed can be performed trough their WebAPI written in Perl [8]. By iterating the query by updating the years, the script can retrieve the published research documents in the field depending on the query string. In this example, the temporal sets of documents year by year are gathered on the field related to the drug therapy for chronic hepatitis.

With this search query, we obtain articles published between 1982 and 2009 with the abstract mode of PubMed[1]. Figure 3 shows the numbers of article titles and abstracts retrieved by the query. In this study, each abstract is assumed as text to be one document.

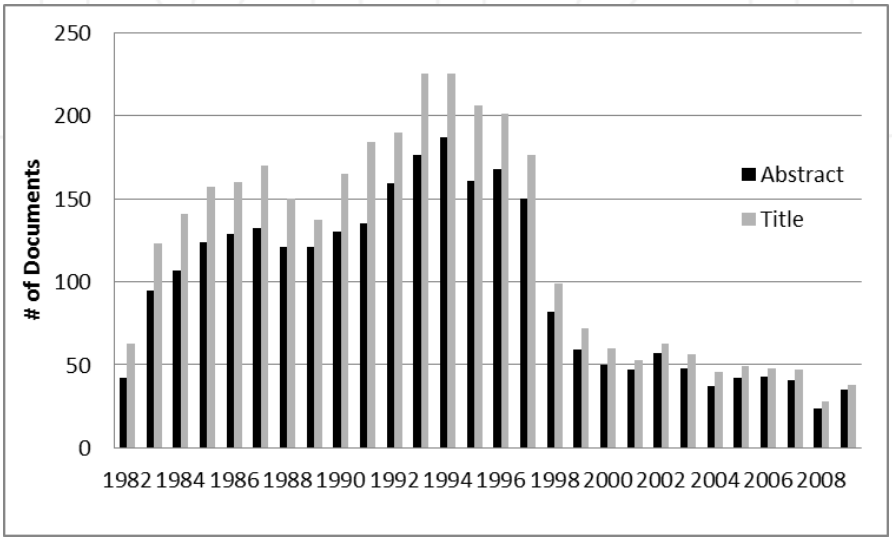

Figure 3. Numbers of documents with titles and abstracts related to hepatitis drug therapy published from 1982 to 2009.

From all of the retrieved abstracts, the automatic term extraction method identifies 12,194 terms. As for the titles, the method extracted 1,428 terms.

3.2. Obtaining temporal patterns of medical terms about chronic hepatitis drug therapy studies

By calculating the document frequency and the tf-idf values as the importance indices for each year on titles and abstracts respectively. By using the document sets and the index, the system obtained the dataset to obtain temporal clusters that consist of temporal behavior of each index year by year for each term.

As for the clustering algorithm, the k-means clustering algorithm implemented in Weka [9](Weka-3-6-2) are applied. Since the implementation search better cluster

[1] The current heading of chronic hepatitis has introduced 1982.

assignments by minimizing the sum of squared errors (SSE), the upper limits of the number of clusters are set up 1% of the number s of terms. And the maximum iteration to search better assignment is set up 500 times.

Table 1 shows the result of the k-means clustering on the sets of documents.

Dataset		# of Clusters	SSE	Total trend		# of patterns		
				Avg.Deg	Avg.Int	Emergent	Popular	Subsiding
Abstrtacts	tf-idf	119	309.70	-0.04	2.65	1	47	71
	df	118	32.02	-0.01	0.64	0	32	86
Titles	tf-idf	14	136.58	-0.03	2.01	0	4	10
	df	14	15.03	-0.01	0.69	0	2	12

Table 1. Overall result of temporal clustering on titles and abstracts about chronic hepatitis drug therapy by using the three importance indices.

Figure 4 shows the centroid values of the temporal clusters and the representative terms of each temporal pattern on the title corpus based on the temporl tf-idf. The centroid values mean the averages of the yearly values of the terms in each cluster. The representative terms are selected with their FLR scores, that are the highest in each temporal custer. The cluster is selected with the following conditions: including phrases, highest linear degree with minimum intercepts to y-axis by sorting the average degrees and the average intercepts of the 14 clusters.

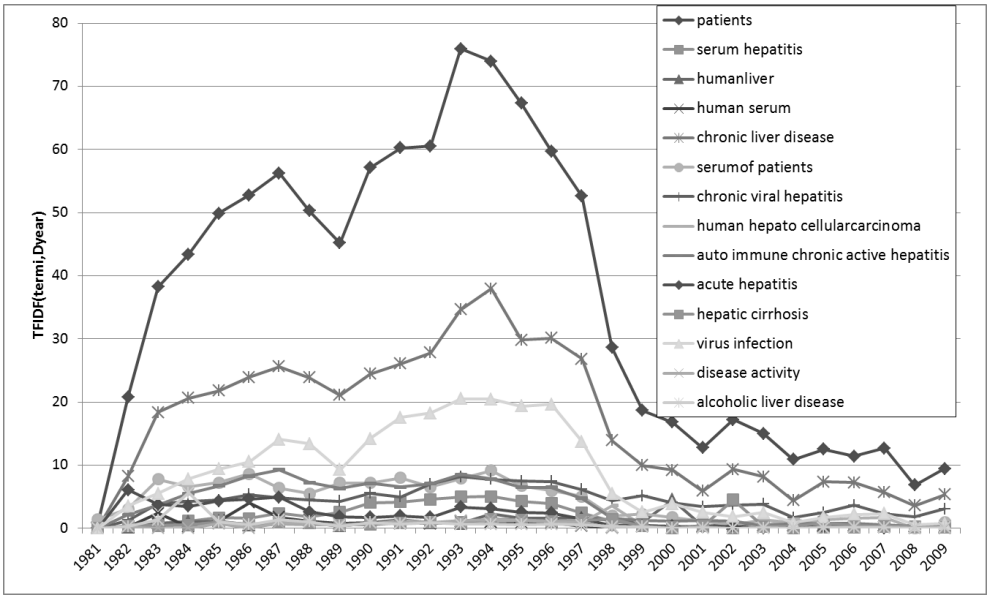

Figure 4. The representative terms and values of tf-idf temporal patterns on the titles of the chronic hepatitis articles.

As shown in Figure 4, the method can detect the trends based on the temporal behaviors of terms. Although the temporal patterns and the similar terms as the member of the clusters

http://dx.doi.org/10.5772/52210

show the trends and the similar group at the same time, the meaning of each group indicates should be evaluated by medical experts.

4. Analyzing temporal trends of terms and the similarities on MeSH structure

In this example, the sets of documents published year by year on the field related to the drug therapy for migraine are gathered. Let us assume that the user wants to explore the evolution of ideas about drugs used to treat migraine[2]. The user performs a search for abstracts of articles "migraine/drug therapy [MH:NOEXP] AND YYYY [DP] AND clinical trial [PT] AND english [LA]" through PubMed. As same as the retrieval for chronic hepatitis in Section 3.

With this search query, articles published between 1980 and 2009 with the abstract mode of PubMed are retrieved. Figure 5 shows the numbers of article titles and abstracts retrieved by the query.

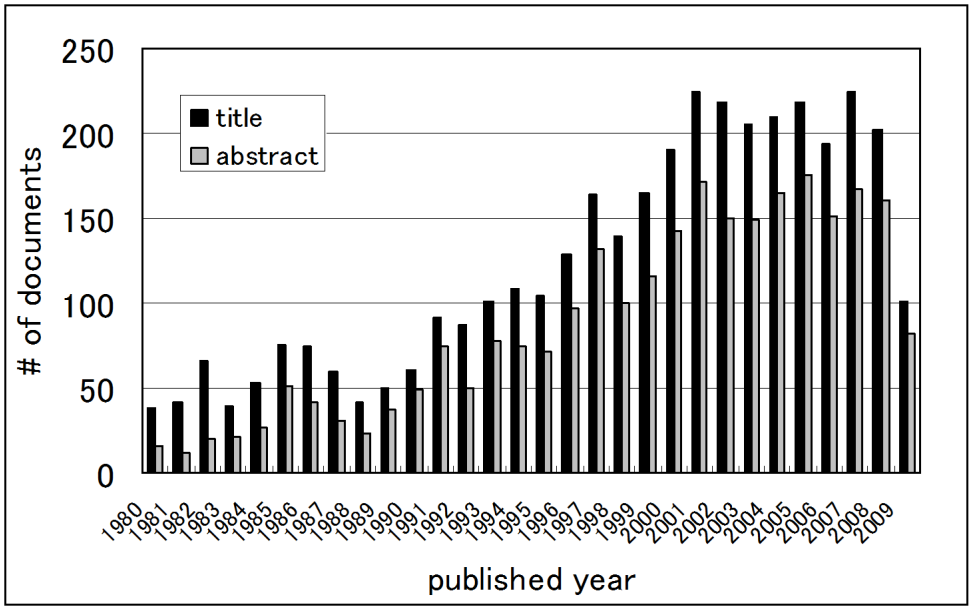

Figure 5. Numbers of documents with titles and abstracts related to migraine drug therapy published from 1980 to 2009.

By assuming the abstracts and the titles as each corpus, the automatic term extraction is applied. From all of the retrieved abstracts, the automatic term extraction method identifies 61,936 terms. Similarly, from all of the titles, the system extracts 6,470 terms.

[2] Migraine causes sharp and severe headaches to people. People who have migraine exist commonly in world-wide. The severe headaches give economical disadvantages not only to the patients, but also socially sometimes.

4.1. Obtaining temporal patterns of medical terms about migraine drug therapy studies

By calculating the document frequency and the tf-idf values as the importance indices for each year on titles and abstracts respectively. Then, the temporal clusters that consist of temporal behavior of each index year by year for each term are obtained. The clustering algorithm that is used in the following experiment is also the same as the setting in Section 3.

Table 2 shows the result of the k-means clustering on the sets of documents.

Dataset		# of Clusters	SSE	Total trend		# of patterns		
				Avg.Deg	Avg.Int	Emergent	Popular	Subsiding
Abstrtacts	tf-idf	129	216.81	0.17	−0.44	81	37	0
	df	129	36.67	0.04	−0.06	103	25	1
Titles	tf-idf	14	125.69	0.10	−0.03	5	9	0
	df	14	10.14	0.03	0.02	4	9	1

Table 2. Overall result of temporal clustering on titles and abstracts about migraine drug therapy by using the three importance indices.

Figure 6 shows the emergent cluster centroid and the top ten emergent terms on the abstracts on the basis of tf-idf. The cluster is selected with the following conditions: including phrases, highest linear degree with minimum intercepts to y-axis by sorting the average degrees and the average intercepts of the 14 clusters.

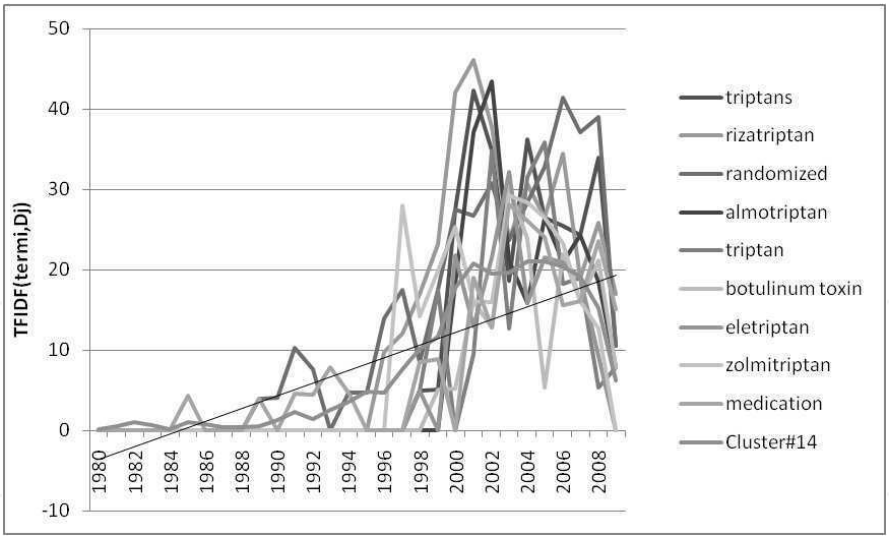

Figure 6. The detailed tf-idf temporal values included in the emergent temporal pattern (Cluster #14).

As shown in Figure 6, the method detected the emergent terms included in the emergent pattern that related to triptans drug therapy. The cluster also includes some terms related to the time for the therapy. The drugs including triptans, which are appeared in this pattern, are

http://dx.doi.org/10.5772/52210

approved later 1990s in US and European countries, and early 2000s in Japan. Based on the result, the method obtained the temporal patterns related to the topics that attract interests of researchers in this field. In addition, the degree of the increasing and the shapes of the temporal patterns of each index show some aspects the movements of the research issue.

4.2. Similarity of the terms in obtained temporal patterns on MeSH

By using the similarity measure as described in Section 2, the averaged similarity of the medical terms included in each temporal pattern are calculated. In order to analyze the relationship between the trends and the similarities, a comparison is performed with the representative values of the averaged similarities of the term.

As shown in Table 3, the similarities for each temporal pattern are calculated. Smaller similarity value means that the terms included in the temporal pattern are defined separately on the MeSH structure. Besides, greater similarity value means that the similar terms on the temporal pattern are also defined similarly on the MeSH structure.

k	term_ck	Deg(ck)	Int(ck)	Meaning	Avg.Sim(ck)
2	clinical efficacy	0.063	−0.142	Emergent	0.133
3	placebo-controlled study	0.069	−0.299	Emergent	0.135
5	5-ht1b/1 dagonists	0.044	−0.139	Emergent	0.131
8	migraine therapy	0.339	−1.870	Emergent	0.141
14	acute treatment of migraine	0.789	−3.566	Emergent	0.135
1	migraine patients	0.059	3.874	Popular	0.120
4	cluster headache	0.050	0.203	Popular	0.144
6	migraine	1.816	3.589	Popular	0.174
7	double-blind study	0.058	0.206	Popular	0.121
9	patients	0.748	1.363	Popular	0.138
10	management of migraine	0.025	0.816	Popular	0.133
11	tension-type headache	0.058	0.196	Popular	0.133
12	oral sumatriptan	0.255	0.955	Popular	0.131
13	migraine prophylaxis	0.025	0.523	Popular	0.160

Table 3. Temporal patterns obtained for the tf-idf dataset on the titles and the similarities of the terms in each temporal pattern.

Then, for clarifying the relationships between the temporal patterns and the similarity on the taxonomy, the difference of the similarity values by separating the meanings of the linear trends is compared on the two trends; emergent or not emergent. As for the first representative values, the two groups of the average values are tested by using t-test. Then, the representative values of the two groups of the similarity values are compared by using Wilcoxon rank sum test. Table 4 shows the averages and the medians of the similarity values.

The similarity values around 0.13 means that the pair of terms defined in each place with from 6 to 7 paths. By testing the difference between the two groups based on the linear trends, for the abstracts, the similarities of the terms included in the emergent temporal patterns are significantly smaller than the terms included in the popular patterns based on the tf-idf values. This result indicates that the tf-idf index detects new combinations

	Abstracts		Titles	
	Emergent	**Not Emergent**	**Emergent**	**Not Emergent**
tf-idf	0.126∗	0.130∗	0.134	0.139
df	0.129∗	0.125∗	0.134	0.141

(a) Averages

	Abstracts		Titles	
	Emergent	**Not Emergent**	**Emergent**	**Not Emergent**
tf-idf	0.126∗	0.130∗	0.132	0.138
df	0.131∗	0.125∗	0.133	0.139

(b) Medians

Table 4. Comparison of the representative values. (a)Averages, (b)Medians. $*$ means significant difference on $\alpha = 0.05$.

of the concepts as its emergent trend. Besides, based on the temporal patterns by using the document frequency, the terms included in the emergent patterns are defined more similarly. More frequently used terms in the recently published documents are defined nearer than the other popular terms. This can be understandable by considering the process for maintaining the structure of the concepts manually.

5. Related work

Related to the method that I described, there are two separated research topics. One is to detect emergent trend in a given temporal corpora. The other is for learning structured taxonomy or ontology from a given corpus. They have not been combined as the method to analyze the relationship between the emergent terms and the place on the structure that the terms should be appeared. This work provides a novel idea not only as a text mining approach, but also for the two separated studies

5.1. Emergent Trend Detection (ETD) methods

In order to detect the emergent trend in a temporally published corpora, the method for detecting emergent trend have been developed [10, 11]. Most of these methods concentrated to find out just one trend at each setting. Moreover, they rather finding terms that represent emergent trends than the emergent trend itself. Thus, the user of these methods should interpret the meaning of the terms that are detected by the ETD method.

Conventional ETD methods are mostly based on the probabilistic transition of the term appearances as shown in the works such as [12, 13]. The method achieved for detecting emergent trend, which is actually a set of terms. However, they did not detect various trends as described in Section 3 and Section 4 at the same time. In addition to the difference, the proposed trend detection method has an availability to visualize both of the representing values of each temporal pattern and the detailed values for each term, using simple time-series charts.

http://dx.doi.org/10.5772/52210

5.2. Ontology Learning (OL) methods from domain corpora

In order to construct a taxonomy for each domain from a given corpus, the methods for leaning ontologies are proposed [14, 15]. However, they did not consider the difference over times of the structured taxonomy. The maintenances of the structure are majorly depending on the manual works of domain experts. Some support methods for the maintenance of the structured taxonomy are really required to the structured taxonomy useful and up-to-date. For this issue, the advanced method of the proposal of this chapter will provide the support with more objective evidences based on the temporal corpora of each particular domain.

6. Conclusion

In this chapter, I describe the method for detecting trends of terms in the published articles in MEDLINE as the case study on chronic hepatitis studies. With this case study, the result shows that the method can find various trends of terms and similar terms at the same time. In this case study, the similar terms detected by using their temporal behavior of the two importance indices; document frequency and tf-idf index. Then, the temporal patterns of the biomedical terms by using the two importance indices are obtained. The patterns indicate the similar usages of the terms on the biomedical research documents as the temporal corpus.

Subsequently, by using migraine drug therapy studies, a comparison of the similarity of the terms between the terms grouped up by our trend detection method and the terms in the structured vocabulary is shown. By using MeSH as the structured taxonomic definition of the medical terms, we compared the averaged similarity based on the distances on the tree structure between the terms included in each temporal pattern. By separating the trends of the temporal patterns based on the linear regression technique, the averaged similarities of the terms in each pattern show significant differences on the larger structured vocabulary. Based on the temporal patterns with the emergent trend of the tf-idf, the terms included in such patterns are not similar compared to the terms included in the popular patterns. This indicates that the novel concepts are obtained from new combination by using the existing concepts widely. Besides, the similarity of the different index detects the opposite relationship between its trend and the similarity on the taxonomic definition.

In the future, more indices for representing various aspects of term usages in a corpus will be introduced and compared. Then, based on the similarities on temporal behavior of each index as the temporal patterns, some predictive models such as numerical prediction models will be introduce for predicting adequate places of new concepts on a structured taxonomy.

Author details

Hidenao Abe

* Address all correspondence to: hidenao@shonan.bunkyo.ac.jp

Department of Information Systems, Faculty of Information and Communications, Bunkyo University, Japan

References

[1] Medical subject headings:. http://www.nlm.nih.gov/mesh/.

[2] Hidenao Abe and Shusaku Tsumoto. Trend detection from large text data. In *Proceedings of the 2010 IEEE International Conference on Systems, Man and Cybernetics*, pages 310–315. IEEE, 2010.

[3] Hiroshi Nakagawa. Automatic term recognition based on statistics of compound nouns. *Terminology*, 6(2):195–210, 2000.

[4] Karen Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. *Document retrieval systems*, pages 132–142, 1988.

[5] Eamonn Keogh, Selina Chu, David Hart, and Michael Pazzani. Segmenting time series: A survey and novel approach. In *an Edited Volume, Data mining in Time Series Databases.*, pages 1–22. World Scientific, 2003.

[6] T. Warren Liao. Clustering of time series data: a survey. *Pattern Recognition*, 38:1857–1874, 2005.

[7] P. Srinivasan. Meshmap: a text mining tool for medline. In *Proc. of AMAI Symposium 2001*, pages 642–646, 2001.

[8] E-utilities:. http://www.ncbi.nlm.nih.gov/books/NBK25500/.

[9] I. H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, 2000.

[10] Brian Lent, Rakesh Agrawal, and Ramakrishnan Srikant. Discovering trends in text databases. In *KDD '97: Proceedings of the third ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 227–230. AAAI Press, 1997.

[11] April Kontostathis, Leon Galitsky, William M. Pottenger, Soma Roy, and Daniel J. Phelps. A survey of emerging trend detection in textual data mining. *A Comprehensive Survey of Text Mining*, 2003.

[12] Jon M. Kleinberg. Bursty and hierarchical structure in streams. *Data Min. Knowl. Discov.*, 7(4):373–397, 2003.

[13] Qiaozhu Mei and ChengXiang 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*, pages 198–207, New York, NY, USA, 2005. ACM.

[14] Philipp Cimiano, Johanna Völker and Rudi Studer. Ontologies on Demand? - A Description of the State-of-the-Art, Applications, Challenges and Trends for Ontology Learning from Text *Information, Wissenschaft und Praxis*, 57, 2006

[15] Hazman Maryam, Samhaa R. El-Beltagy and Ahmed Rafea A Survey of Ontology Learning Approaches *International Journal of Computer Applications*, 22, 8, 2011