Emerging Trend Prediction in Biomedical Literature

Fabian Moerchen, PhD, Dmitriy Fradkin, PhD, Mathaeus Dejori, PhD, Bernd Wachmann, PhD
Integrated Data Systems, Siemens Corporate Research, Princeton, NJ

Abstract

We present a study on how to predict new emerging trends in the biomedical domain based on textual data. We thereby propose a way of anticipating the transformation of arbitrary information into ground truth knowledge by predicting the inclusion of new terms into the MeSH ontology. We also discuss the preparation of a dataset for the evaluation of emerging trend prediction algorithms that is based on PubMed abstracts and related MeSH terms. The results suggest that early prediction of emerging trends is possible.

Introduction

The information landscape is rapidly growing and humans are struggling to keep up with it. Scientific research, for example, is highly dynamic with ground-breaking technologies changing established fields and creating new research territories. Especially in the biomedical domain breakthrough technologies are increasing the fragmentation and the invention of new fields make it impossible for humans to keep up with the latest information, trends, and findings in a reasonable amount of time. Gathering up-to-date information is crucial for the business success and indispensable at any level of organization. Thus, emerging trend prediction algorithms that scan large amounts of content to extract trends in an early stage are needed, especially for industry investigators looking for new promising directions established by researchers at the cutting edge of the field.

PubMed[1], the largest biomedical bibliographic text database with over 17 million articles and more than 10,000 newly submitted research abstracts every week, represents a good basis for monitoring breakthrough technologies and extracting trends. The U.S. National Library of Medicine (NLM) is hosting and maintaining the database and takes responsibility for the categorization and annotation of incoming documents with metadata based on the Medical Subject Headings (MeSH)[2] ontology. Besides its usefulness as a resource for associating semantic tags (annotations) to PubMed abstracts, the MeSH ontology also provides a formal and explicit specification of the present biomedical knowledge.

The MeSH ontology is manually curated and undergoes a major update every year based on new technologies or findings in the biomedical field. For example, genes or proteins that are crucial for the development and progress of a certain disease are added as concepts to the MeSH ontology once a certain degree of consensus in the scientific community has been established. Anticipating changes in the ontology can therefore be seen as predicting new established knowledge. In the following we study the development of the MeSH terms relating to cancer and show how our system can be used to predict emerging trends in a very early stage.

For the remainder of this paper, we will focus on the task of predicting emerging biomarkers from PubMed abstracts. In other words, we predict whether a gene becomes a disease relevant biomarker in the next 1-5 years given the observed frequency of the related molecular term. Unlike many other proposed methods on emerging trend *detection* we carefully prepare a dataset with a known ground truth derived from an external source and try to predict trends rather than to retrospectively detect them.

Related work

One of the earliest papers mentioning trends in document archives is [7]. The frequency of phrases is tracked over time and the user can query for trends of predefined shapes. This approach could detect recently emerging trends. [3] discusses emerging trends in the context of itemsets as patterns with large difference in support between two datasets (subsets of a single dataset or a time-based partition of the data). No evaluation with ground truth was performed in these papers since the analysis was mostly exploratory.

So far studies concentrate only on analyzing the past topics by retrospective analysis of document archive. [12] uses significance tests to detect time periods where words have a higher than usual frequency. [13] proposes a generative probabilistic model approach to better model evolving topics.

A state-based model is used in [5] to model a hierarchical structure of bursts in a document stream. One application is the automated categorization of emails into topics and subtopics. The burst detection is used in [4] to cluster trends of words. We have so far concentrated on the ranking of individual features but this could be extended to clusters of features.

Several approaches use a partition of the time axis into large intervals to detect changes in document archives. Finite mixture models are used in [9] to represent the documents of each interval. A comparison of the models among subsequent time intervals is used to report new trends. In [8] clustering is performed for each time step and clusters are connected with a graph model to obtain a temporal representation of the document collection. A similar method is used in [11] to detect emergent and persistent topics for the extension of ontologies. New clusters that cannot be connected to clusters from the previous time interval are candidates for new ontology concepts. [2] decomposes the time axis related to a document archive, trying to minimize the information loss in the set of significant features per time step. The method is applied to a small excerpt of PubMed in [14]. The methodology of Literature-related discovery proposed in [6] aims to combine literature from separate scientific fields to discover previously unknown relations and possibly synergies. Candidates are validated by searching patents for prior art.

We focus on analyzing textual data from the biomedical domain aiming at *continuously* detecting those medical entities, e.g. genes, proteins or medical devices, that are likely

[1] `http://www.ncbi.nlm.nih.gov/PubMed/`
[2] `http://www.nlm.nih.gov/mesh/`

to become a breakthrough technology *in the future*. This is in contrast to the methods above that use a coarse grained temporal resolution or focus on retrospective analysis. We believe the results of such approaches are of limited use because of the associated time lags in the detection. Unlike retrospective approaches that analyze whole collections to extract patterns in the past data, we use only data available before a time point to make predictions about the future (with respect to that time point) trends.

Data collection

Ground truth

We utilized the temporal information associated with MeSH terms in the 2008 version to obtain a ground truth for emerging trends. We assume that MeSH terms are added once the NLM considers a term an established and relevant medical concept and that there must have been significant research activity prior to the inclusion in the ontology. The terms are placed in the ontology and cross-referenced with existing terms such as diseases.

We searched the MeSH ontology for cancer related terms that were added between 1975 and 2007 and would have been interesting to a researcher monitoring the literature for scientific breakthroughs. Our filtering strategies are very strict to ensure that we only have truly interesting concepts as true positives. We started out with all MeSH terms that were observed in at least one of the cancer related documents. These 22,169 terms covered almost the complete ontology. Obviously not all these terms are related to cancer. We therefore used the tree structure to filter the MeSH terms for entries that are listed in a tree that has one of the cancer keywords (see above) in the path name. This results in 224 relevant trees with 759 relevant terms. From these 65 were removed because they were established before 01/01/1975. 524 of the remaining 694 terms were associated with the top level tree *Neoplasms [C04]*. Since we are interested in predicting biomarkers and technologies relevant for the diagnosis and treatment of diseases but not so much in predicting disease type themselves we excluded these terms. We only kept terms in one of the top level trees listed in Table 1 resulting in 140 remaining terms. For the evaluation the MeSH terms

Tree number	Tree name
A11	Cells
D08	Enzymes and Coenzymes
D12	Amino Acids, Peptides, and Proteins
D23	Biological Factors
D27	Chemical Actions and Uses
G05	Genetic Processes
G14	Genetic Structures

Table 1: Top level MeSH trees used to filter terms for biomarkers.

needed to be mapped to the observed word stems from the abstracts. In this process we removed umbrella concepts such as *Genes, neoplasm* or *Cell line, tumor* and terms that cannot be easily identified with a single word(stem) such as *b-cell maturation antigen*. We plan to use n-grams to cover such terms but this will increase the computational complexity. We also removed duplicates where the same word stem was matched to a gene and a protein entry in MeSH. The MeSH term with the earlier creation time was used in this case. The final result was a list of 81 MeSH terms as true positives for cancer-related biomarkers. The distribution of these terms over time is shown in Figure 2 and some examples are listed in Table 2.

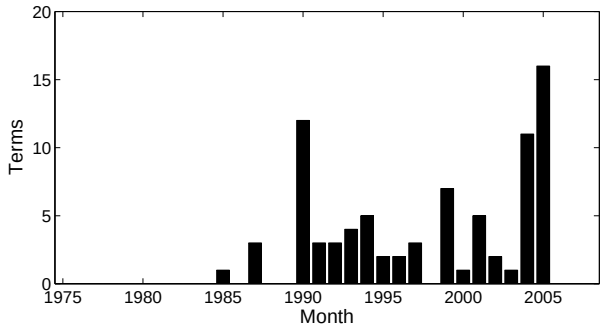

Figure 1: Addition of new MeSH terms describing biomarkers related to cancer.

Type	Count	Examples
genes	41	ras, p53, dcc, erbb-2, brca1
antigen	8	cd27, cd137, cd70, ca-125, ca-19-9
receptors	18	tnf1, tnfsf14, traf4, ox40, xedar
proteins	6	wt1, p14arf, p130, p107, fas
cells	8	pc12, hl-60, caco-2, k562, jurkat

Table 2: True positive cancer-related biomarkers.

Observed terms

We filtered the PubMed database from 01/1975 through 10/2007 for abstracts with the following cancer related keywords (substrings): *cancer*, *carcinoma*, *tumor*, *neopla*, *malignant*. About 1.5M documents were found and processed with the standard text mining pipeline: word level parsing, stop word removal, word stemming [10]. The stop word list included some very common medical terms such as *result*, *patient*, *study*, *method*. In total about 600 stop words were used. The MeSH annotations of the abstracts were not used and no named-entity recognition was performed to ensure that no information is used that would not have been available at the time the abstracts were published. Individual parts of composite terms with hyphens or slashes, e.g., *P53-induced*, were considered as separate words if they were longer than a single character and not a number. Some biomarker specific normalization was performed, for example by replacing the suffix *-ii* with *-2*.

Each document in PubMed has up to four date fields: creation date, completion date, revision date, and publication date. There is no total order among these dates. We used the earliest available date for each document as a time stamp. This corresponds to the time that a researcher that would have had access to all resources would have seen the article.

The frequency of all word stems found in the documents was tracked over time generating time series with a granularity of months. Using a minimum frequency threshold of 5 left about 170k trends for further analysis. Obviously these trends are not observed over the complete time period. The solid line in Figure 2 shows the number of terms

that are present in PubMed each month. However, since we aim to predict new MeSH terms, we are only concerned with word stems that have not yet appeared in the MeSH ontology. These are shown with a dashed line. As can be seen, the set of candidate terms is reduced only slightly. We further removed trends that were found in a list of 80k word stems derived from an English dictionary[3], the remaining terms are shown as the 3rd line from the top. This results in a significant reduction, though more than 100k candidate trends are left to be analyzed in the years following 1999 and about 140k in 2005. This is the set of trends we used for further analysis. We also considered an inclusive approach,

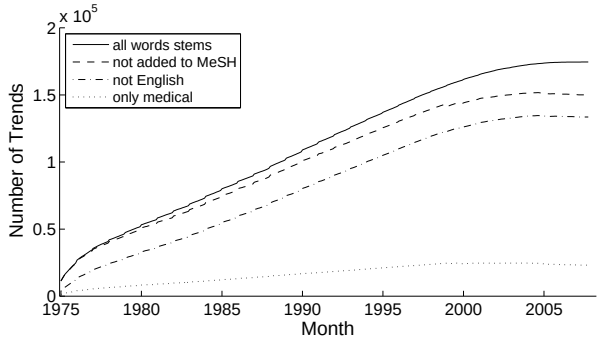

Figure 2: Number of terms after filtering

by generating a list of 1.1M medically relevant word stems derived from the MeSH descriptors, MeSH pharmacological actions and substances, UniProt[4], EntrezGene[5]. The dotted line Figure 2 shows the number of trends found in this list. This approach appeared to be too restrictive because many excluded word stems looked like potential biomarkers.

Trend analysis

Normalization

The number of cancer-related articles per month is shown in Figure 3. The clear upward trend is another proof of the information overload analysts are facing nowadays. The peaks have a yearly period and might be caused by the inclusion of journals or other publications on a yearly basis. In order to remove the yearly peaks and add some smoothing to the trends we used a moving average: For each month we consider the frequency of a word stem in all cancer-related documents from the last year up to the month. To account for the global trend in biomedical (cancer) research the trends were divided by the total number of documents in the same time range.

Figure 4 shows an example of such a normalized frequency trend for the Breast Cancer gene BRCA1. The dashed line indicates the time when the corresponding MeSH term was added to the ontology. There was significant research activity with a sharply increasing trend starting in January of 1993. The corresponding MeSH term was added in February 1996, more than 3 years later. Figure 5 shows the trend for the gene P53. The first mentioning as discovered by our system was in May 1979. In the figure the trend starts to

[3] `http://www.winedt.org/Dict/`
[4] `http://www.pir.uniprot.org/`
[5] `http://www.ncbi.nlm.nih.gov/sites/entrez?Db=gene`

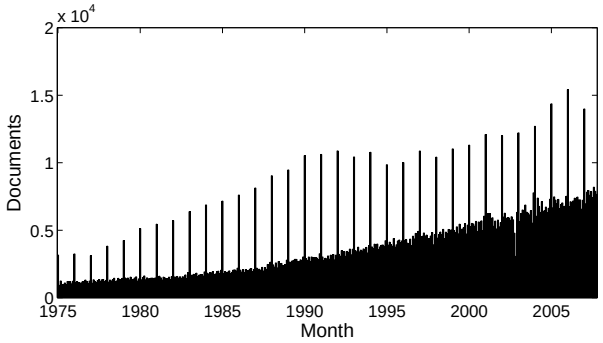

Figure 3: Number of cancer related documents per month in PubMed database. The frequency of publications is strongly increasing.

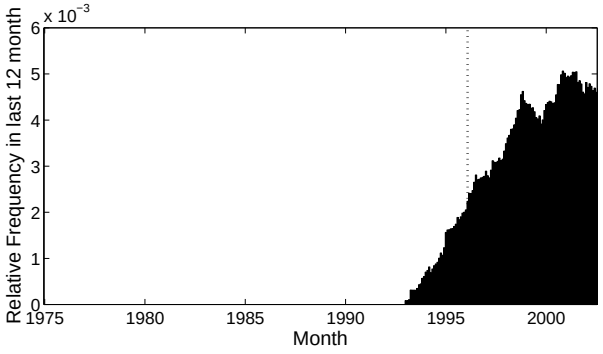

Figure 4: Trend of word stem *brca1* that corresponds to the MeSH term *Genes, BRCA1* added on 2/16/1996 (dashed line).

become visible in the early 80's with several bumps in the mid and late 80's. Ideally we would already detect these relatively small peaks. In the early 90's the relative frequency starts to increase rapidly. The Mesh term *Genes, P53* is added in 1992.

Scoring

A trend scoring system works by monitoring candidate terms, and, at each time point, assigning each term a score indicating the interestingness of the term. Since we are interested in newly emerging trends we would like the score to reflect the rate at which the relative frequency $f(w,t)$ of a term w changes with time. We compute the score $s(w,t_c)$ of trend w at time t_c in the following way:

$$s(w,t_c) = \sum_{i=23}^{0} f(w,t_c - i) > \max_{j=24,\ldots,i+1} f(w,t_c - j) \quad (1)$$

- in other words we count the number of times that frequency of a trend has increased compared to the previous maximum, within the 24 month period. High score indicates a steady increase in frequency.

We have considered several other scoring functions, but they did not perform as well. Due to space constraints we omit their descriptions and discussion of their results.

Note that this approach is computationally simple. To

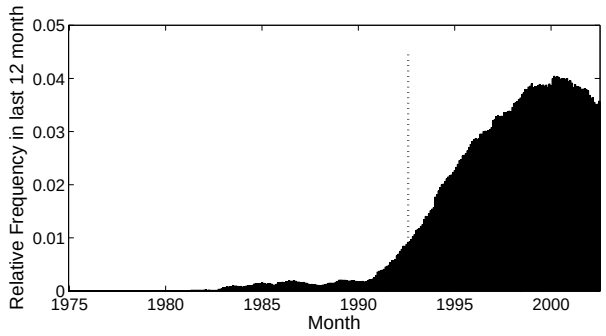

Figure 5: Trend of word stem *p53* that corresponds to the MeSH term *Genes, P53* added on 8/12/1992 (dashed line).

compute normalized frequencies for a term in a particular month, we need frequencies of that term for the preceding 12 month. To make a prediction, we need normalized frequencies for 24 months. Therefore, we need to keep 36 numbers for each term. The term frequencies are updated each month by scanning through the new documents. The normalized frequencies are then recomputed, and scores for each term are calculated. The cost is linear in the size of the vocabulary and the total length of all new documents.

Evaluation

We consider a set containing all 81 positive examples and all *observed terms*, selected as described above (up to 140K terms). In order to evaluate our approach we consider several different types of measures, illuminating different aspects of the problem.

One question of interest is whether in fact our approach is capable of detecting new important trends early, i.e., whether we can predicting MeSH terms in advance. One way to answer that is to consider the time between positive term's first appearance in top-k terms and its inclusion into MeSH. Figure 6 shows the distribution of differences between these two events for $k = 300$. We detect 48 out of 81

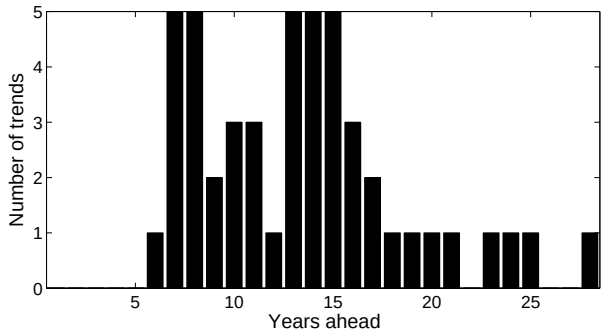

Figure 6: Time difference in years between inclusion in MeSH and earliest detection for positive trends using the top 300 trends every month. Larger value indicates earlier detection. 48 out of 81 positive terms are detected.

terms with $k = 300$. This shows that our simple approach can detect more than half of true positives far in advance of their acceptance into MeSH. Obviously there are some costs attached. Considering 300 terms each month for 25 years could potentially lead to a total of 90000 terms - certainly a very high cost. It turns out however, that the real cost is significantly lower because terms appear in top-k sets multiple times. For example, only 6290 unique terms appear in top 300 in the time covered in evaluation. The effort in following such number of trends over 25 years is quite manageable. Also, the numbers in the figures are "pessimistic", because in computing them we do not remove previously correctly identified terms or the irrelevant trends that a user of the ranking system would have the option to remove from further consideration.

The other measures we use are standard Information Retrieval (IR) metrics: precision and recall [1]. Before we define those, we first need to clarify how we categorize terms in a way that is consistent with our applications.

- "Discarded Terms" are the terms that will not be considered in computing quality measures. First of all, these are the terms that did not yet occur as of time t_c. Second, they are terms that have already been added to MeSH. Third, we discard terms that are added to MeSH at time $t \in [t_c, t_c+h_0)$. The last one requires additional explanation: in order for the trend detection to have value, it must come some time before inclusion in MeSH. Horizon h_0 is used to define how far in advance detection should be made. The terms that are about to be added to MeSH are not positives in a sense that it is too late to detect them, yet they are definitely not negatives. Therefore, we do not consider them when evaluating results.

- True Positives *tp* at time t_c are trends that will be added to MeSH at time $t \geq t_C + h_0$.

- False Positives *fp* at time t_c are trends that never be added to MeSH but appear in top k at time t_c. Note, that for recent years the inclusion into MeSH might occur in the future, so some *fp* may be currently unknown *tp*.

- False Negatives *fn* at time t_c are trends that will be added to MeSH a time $t \in [t_C + h_0, t_C + h_1]$, where $h_1 > h_0$ is another horizon parameter. In other words, it would hardly be reasonable to penalize a method for not detecting a trend that will only be recognized far in the future, possibly when a lot more research has been done in the area.

- All other terms are considered true negatives *tn*.

In the experiments described here we set $h_0 = 12$ months, and $h_1 = 60$ months. Given the numbers above precision P and recall R can be calculated. Note that since the earliest any of the 81 terms are added to MeSH is on 04/25/1985 and the latest is added on 12/21/2005, it only makes sense to consider time period from 01/1980 (which is when we could possibly detect the earliest trend) to the end of 12/2004 (which is when we would still be able to detect the last trend without being late). Towards the end of this period we would effectively be making predictions for trends which may only be added to MeSH in the coming years, and this could result in a reduced precision. In Figure 7 we show precision and recall of our approach for $k = 300$. The fact that the precision stays between 0.3% and 3.7% (except in the end) is not surprising, since we look at top k terms at

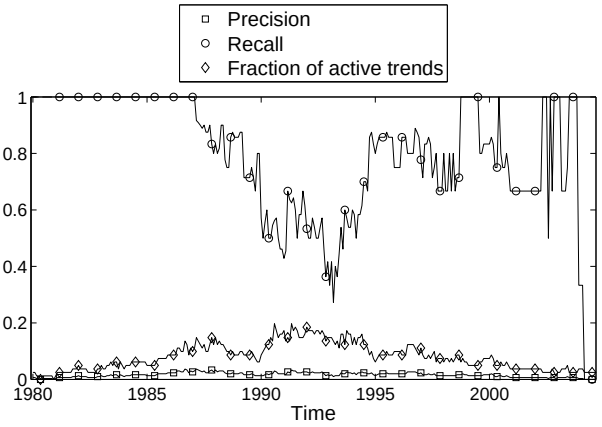

Figure 7: Precision, Recall, and Fraction of 81 positive trends within the time window (tp+fn) in period from January 1980 to January 2004.

each time point and real trends are rare. For many time points such precision is much better than what would be obtained randomly: the overall expected value is less than 0.06% (81/140,000) and the number of active positive trends at each time point is commonly much smaller (at most 16). The mean and median of precision are both above 1.5%. The recall is very high initially, when there are relatively few positive trends in the window and it decreases as more positive trends appear. At any time (with exception of few short intervals without positives) somewhere between 2 and 16 positive trends are potentially detectable. The recall is greater than 20% (except for a small period at the very end of evaluation) and both mean and median of recall are above 75%.

The performance drops in the last year of the observed period. This occurs for the following reasons. Many of true trends that are added to MeSH around 2005 have already slowed down their growth so they are no longer among the top k (though they were detected earlier). Meanwhile, since we reached the end of the experimental period, no new trends appeared that can be detected by the method. It is also possible that the method is detecting trends that will yet be added to MeSH in the future, but for now they are treated as false positives. In other words, the change in last period reflects the nature of our evaluation rather than the weakness of the method.

Conclusion

We demonstrated that early prediction of technological trends is feasible. Several key concepts for cancer were detected with high confidence years before they were introduced in the ontology. We also propose an approach for preparing ground truth in the realm of biomedical research and evaluating rankings for this task. To our knowledge this is the first attempt at *predictive* evaluation of trend detection, as opposed to retrospective detection. Our experience highlights some of the difficulties of such an evaluation, namely: complicated selection of candidates and definition of false negatives and true positives at each time point, and incomplete information on relevance of trends (due to the evolving and expanding vocabulary of MeSH). The task is complicated by existence of trends that do not pick up at the first few occurrences of a term, but develop later, since we have no way of knowing this in evaluation.

The ranking with a simple unsupervised scoring function and selection of the top k worked surprisingly well. We plan to expand the ground truth and move toward supervised algorithms. Some of the drawbacks revealed in our evaluation can be removed in a live system. For example, terms predicted to be included can be removed from future consideration. More importantly, terms not describing biomarkers, for example, or those deemed irrelevant by the user can be eliminated, improving performance.

References

1. Baeza-Yates R, Ribeiro-Neto B, Modern information retrieval, Addison-Wesley, 1999.
2. Chundi P, Rosenkrantz DJ, On lossy time decompositions of time stamped documents, Proc. ACM Intl. Conf. on Information and Knowledge Management, 2004:437-445.
3. Dong G, Li, J, Efficient mining of emerging patterns: discovering trends and differences Proc. ACM SIGKDD Intl. Conf. on Knowledge discovery and data mining, 1999:44-52.
4. Fung GPC, Yu JX, Yu PS, Lu H, Parameter free bursty events detection in text streams. Proc. Intl. Conf. on Very Large Data Bases, 2005:181-192.
5. Kleinberg J Bursty and hierarchical structure in streams. Proc. ACM SIGKDD Intl. Conf. on Knowledge discovery and data mining, 2002:91-101.
6. Kostoff RN, Briggs MB, Solka JL, Rushenberg RL, Literature-related discovery (LRD): methodology, Technological Forecasting & Social Change, 2007.
7. Lent B, Agrawal R, Srikant R, Discovering trends in text databases, Proc. ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 1997:227-230.
8. Mei Q, Zhai C, Discovering evolutionary theme patterns from text: an exploration of temporal text mining. Proc. ACM SIGKDD Intl. Conf. on Knowledge discovery and data mining, 2005:198-207.
9. Morinaga S, Yamanishi K, Tracking dynamics of topic trends using a finite mixture model. Proc. ACM SIGKDD Intl. Conf. on Knowledge discovery and data mining, 2004:811-816.
10. Porter, MF, An algorithm for suffix stripping, Program, 1980:14:3:130-137.
11. Schult R, Spiliopoulou M. Discovering emerging topics in unlabelled text collections. Proc. East European ADBIS Conf., 2006:353-366.
12. Swan R, Allan J, Automatic generation of overview timelines, Proc. ACM SIGIR Intl. Conf. on information retrieval, 2000:49-56.
13. Wang X, McCallum, A, Topics over time: a non-markov continuous-time model of topical trends, Proc. ACM SIGKDD Intl. Conf. on Knowledge discovery and data mining, 2006:424-433.
14. Zhang R, Chundi, P, Using time decompositions to analyze pubmed abstracts. Proc. IEEE Symposium on Computer-Based Medical Systems, 2006:569-576.