Anticipating Annotations and Emerging Trends in Biomedical Literature

Fabian Mörchen, Mathäus Dejori, Dmitriy Fradkin, Julien Etienne, Bernd Wachmann
Integrated Data Systems
Siemens Corporate Research
755 College Road East
Princeton, NJ, 08540, USA
firstname.lastname@siemens.com

Markus Bundschus
Department of Computer Science
Ludwig-Maximilians-University
Munich, 80538, Germany
bundschu@dbs.ifi.lmu.de

ABSTRACT

The BioJournalMonitor is a decision support system for the analysis of trends and topics in the biomedical literature. Its main goal is to identify potential diagnostic and therapeutic biomarkers for specific diseases. Several data sources are continuously integrated to provide the user with up-to-date information on current research in this field. State-of-the-art text mining technologies are deployed to provide added value on top of the original content, including named entity detection, relation extraction, classification, clustering, ranking, summarization, and visualization. We present two novel technologies that are related to the analysis of temporal dynamics of text archives and associated ontologies. Currently, the MeSH ontology is used to annotate the scientific articles entering the PubMed database with medical terms. Both the maintenance of the ontology as well as the annotation of new articles is performed largely manually. We describe how probabilistic topic models can be used to annotate recent articles with the most likely MeSH terms. This provides our users with a competitive advantage because, when searching for MeSH terms, articles are found long before they are manually annotated. We further present a study on how to predict the inclusion of new terms in the MeSH ontology. The results suggest that early prediction of emerging trends is possible. The trend ranking functions are deployed in our system to enable interactive searches for the hottest new trends relating to a disease.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Algorithms, Measurement

KDD'08, August 24–27, 2008, Las Vegas, Nevada, USA.

Keywords

Text mining, Trends, Prediction PubMed, MeSH, LDA

1. INTRODUCTION

The information landscape is rapidly growing and humans are struggling to keep up with it. Scientific research, for example, is highly dynamic with groundbreaking 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. However, gathering up-to-date information is crucial for the business success and indispensable at any level of organization: from the discovery of a new technology in the R&D department up to the definition of new strategies in the management.

Automated text analysis methods that scan large amounts of content for interesting knowledge are needed. We present a system that supports information retrieval tasks for the purpose of technology monitoring in biomedical research. It integrates several information sources in a single user interface and provides added value not found in existing applications using the same or similar data.

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 perfect 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. Given these properties, the BioJournalMonitor processes textual and MeSH ontology meta-data together in a twofold way in order to effectively monitor technologies and extract trends:

- Currently, PubMed abstracts are annotated mostly

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

manually by human experts who read the entire document before adding MeSH terms. This procedure causes a massive time delay between the inclusion of a new document in PubMed and its annotated version. In this paper we present results from a study on automatically predicting MeSH metadata for newly incoming biomedical abstracts based on a probabilistic topic model approach.

- The MeSH ontology is also curated manually undergoing a major update every year based on newly upcoming 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 included as concepts to the MeSH ontology once a certain degree of concensus in the scientific community has been established. Anticipating changes in the ontology can therefore be seen as predicting new established knowledge. We studied the development of the MeSH terms relating to cancer and show how our system can be used to predict emerging trends very early.

The remainder of this paper is structured as follows: In the next section, we give an overview of the trend detection system. In Section 3 we describe a probabilistic approach to automatically annotate incoming articles with MeSH terms. We present results on monitoring new emerging technologies by predicting the inclusion of terms into the MeSH ontology in Section 4. Related work is listed in Section 5 before we conclude with a discussion of our findings in Section 6.

2. BIOJOURNALMONITOR

The rapid advances in bioinformatics, medicine, biomolecular sciences and related scientific areas led to a dramatic increase of publications. The large amount of available knowledge leads to the common problem that finding the right information at the right time itself constitutes a challenge and valuable time of researchers needs to be spent for this task instead of focusing on the actual research itself.

We have developed the BioJournalMonitor platform that supports screening multiple textual resources for potential targets such as biomarkers and related technologies. The users can evaluate possible target candidates and narrow down the list of candidates for subsequent experimental pipelines.

Several publicly available data sources are utilized by the screening system, including

- PubMed with more than 17M biomedical research abstracts,
- MeSH Ontology with more than 23k medical terms,
- Gene Ontology with more than 22k processes and functions,
- UniProt with more than 200k proteins,
- FDA Clinical Trials [3] with more than 50k reports.

Proprietary data sources include patent information and news articles. Most of the sources are dynamic requiring continuous updates to the system to provide the user with up-to-date information on current research in this field. The information is processed with state-of-the-art text mining methods to provide an added value on top of the mere integration of the data sources.

Named-entity recognition Named-entities such as genes, diseases, or substances are detected with a combination of dictionary based recognition and Natural Language Processing [16] techniques. The results are used to emphasize such words in the feature representation to increase the quality of other algorithmic steps like clustering and classification. Disease-biomarker relations are automatically extracted.

Classification Recent documents are automatically annotated with terms of the MeSH ontology using a probabilistic topic model. See Section 3 for details.

Trend analysis The frequency of all terms including named-entities is tracked over time. Features are extracted from the trends to automatically detect emerging trends. See Section 4 for details.

Clustering Stream clustering is used to discover global trends in the document streams [20]. On-demand clustering is used to structure the results of a user-defined search.

Ranking Documents, document clusters, and trends are ranked using extracted features and weights assigned to these features. Since different users have different interests, the weights used in the ranking functions for features like age or relevance to the query can be set as a preference. We are currently collecting user data to be able to enhance the ranking automatically [13].

The documents and trends are presented to the user in an interactive web-based user interface shown in Figure 1. The user can query for keywords and is presented with a list of document clusters on the right hand side. Each cluster is described by a tag cloud of the most important words and entities detected in the documents of the cluster. For each cluster the list of documents and the distribution of the documents over time can be displayed. Large clusters can be refined by breaking them into smaller clusters. The left hand side of the user interface is dedicated to the trend analysis. For each query the system automatically determines the most relevant terms used in the document. The user can browse and compare the trends from different categories such are genes, protein, diseases, or substances. For each trend the corresponding document can be analyzed.

Figure 1: User interface of BioJournal Monitor. Search results are clustered and relevant emerging trends are automatically extracted.

[3] http://clinicaltrials.gov

3. AUTOMATED ONTOLOGY INDEXING

3.1 Motivation

The annotation of unstructured textual data with structured machine readible information is an important step for further applications such as information retrieval, document clustering etc. Articles selected for inclusion in PubMed, for example, are indexed with descriptors from the Medical Subject Headings thesaurus to facilitate later retrieval. We will refer to these descriptors as MeSH "terms" and MeSH "concepts" interchangeably. These MeSH terms are mostly assigned manually by medical experts who read the full text of an article, resulting in a very expensive and time-consuming procedure. The annotation process itself is therefore a major bottleneck for keeping the PubMed database up-to-date given the enormous amount of submissions per day. However, in order to detect trends as early as possbile, it is essential to be able to overcome this information bottleneck to provide up-to-date annotations for the latest articles.

Here, we present a document annotation module for the BioJournalMonitor that is based on a generative model for document collections. The so-called Topic-Concept (TC) model [5] simultaneously models the content of documents and the process of annotating documents. As in [4] each document is represented as a mixture of probabilistic topics. It extends previous work by modeling the process of indexing based on assigned topic distributions for words (see Figure 2).

3.2 Linking ontology terms via latent topic variables

Let $\mathbf{D} = \{d_1, d_2, ..., d_D\}$ be a set of documents, where D denotes the number of documents in the corpus. A document d is represented by a vector of N_d words, $\vec{w}_d$, where each word w_i is chosen from a vocabulary of size N. In our extension, a document d is additionally described by a vector of M_d MeSH concepts $\vec{c}_d$, where each concept c_i is chosen from a set of MeSH concepts of size M. The collection of D documents is defined by $\mathbf{D} = \{(\vec{w}_1, \vec{c}_1), ..., (\vec{w}_D, \vec{c}_D)\}$.

3.2.1 Extension of the classical LDA framework

The Topic-Concept model extends the LDA framework by simultaneously modeling the generative process of *document generation* and the process of *document indexing*. In addition to the steps performed in the classical LDA framework (see Figure 2(a)), two further steps are introduced to model the process of indexing. For each of the M_d concepts in the document a topic $\tilde{z}$ is uniformly drawn based on the topic assignments for each word in the document. Finally, each concept c is sampled from a multinomial distribution over concepts specific to the sampled topic. This generative process corresponds to the hierarchical Bayesian model shown in Figure 2(b). In this model, Γ denotes the vector of multinomial distribution over M concepts for each of T topics being drawn independently from a symmetric Dirichlet prior γ. After the generation of words, a topic $\tilde{z}$ is drawn from the document specific distribution, and a concept c is drawn from the $\tilde{z}$ specific distribution Γ. The probability distribution over M MeSH concepts for the generation of a concept c_i within a document is specified as:

$$p(c_i) = \sum_{t=1}^{T} p(c_i|\tilde{z}_i = t)p(\tilde{z}_i = t|\mathbf{z}) \quad (1)$$

where $\tilde{z}_i = t$ represents the assignment of topic t to the ith concept, $p(c_i|\tilde{z}_i = t)$ is given by the concept-topic distribution Γ. The topic for the concept is selected uniformly out off the assignments of topics in the document model, i.e., $p(\tilde{z}_i = t|\mathbf{z}) = \text{Unif}(z_1, z_2, \ldots, z_{N_d})$ leading to a coupling between both generative components. The generative process

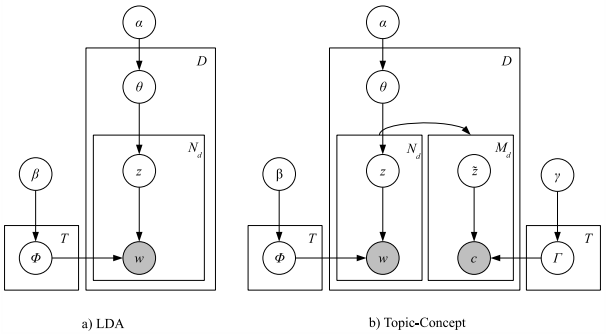


Figure 2: Graphical model for a) LDA and b) Concept-LDA in plate notation. Shaded nodes represent observed random variables, unshaded nodes represent latent random variables

of the TC model is essentially the same as the Correspondence LDA model proposed in [3] with the difference that the TC model imitates the generation of documents and their subsequent annotation, while [4] models the dependency between image regions and captions.

3.2.2 Learning the Topic-Concept model from text collections

Estimating Φ, θ and Γ provides information about the underlying topic distribution in a corpus and the respective word and MeSH concept distributions in each document. Given the observed documents, the learning task is to infer these parameters for each document. Instead of estimating the parameters directly [11, 3] we follow the idea of [8] and estimate Φ and θ from the posterior distribution over the assignments of words to topics $p(\mathbf{w}|\mathbf{z})$. As the posterior cannot be computed directly, we resort to Gibbs sampling generating samples from the posterior by repeatedly drawing a topic for each observed word from its probability conditioned on all other variables. In the LDA model the algorithm goes over all documents word by word. For each word w_i, a topic z_i is assigned by drawing from its conditional distribution

$$\begin{aligned} p(z_i = t|w_i = n, \mathbf{z_{-i}}, \mathbf{w_{-i}}) &\propto \\ p(w_i = n|z_i = t)p(z_i = t) &\propto \\ \frac{C^{WT}_{nt} + \beta}{\sum_{n'} C^{WT}_{n't} + N\beta} \frac{C^{DT}_{dt} + \alpha}{\sum_{t'} C^{DT}_{dt'} + T\alpha} & \end{aligned} \quad (2)$$

where $z_i = t$ represents the assignments of the ith word in a document to topic t, $w_i = n$ represents the observation that the ith word is the nth word in the lexicon, and $\mathbf{z_{-i}}$ represents all topic assignments not including the ith word. C^{WT}_{nt} is the number of times word n is assigned to topic t and C^{DT}_{dt} is the number of times topic t has occurred in document d, both excluding the current instance. In the TC model the posterior $p(c|\tilde{\mathbf{z}})$ is approximated by assigning for

each concept c_i, a topic $\tilde{z}_i$ from the following distribution:

$$\begin{aligned} p(\tilde{z}_i = t | c_i = m, \tilde{\mathbf{z}}_{\mathbf{i}}, \mathbf{z}_{-\mathbf{i}}, \mathbf{w}_{-\mathbf{i}}) &\propto \\ p(c_i = m | \tilde{z}_i = t) p(\tilde{z}_i = t | \mathbf{z}) &\propto \\ \frac{C^{CT}_{mt} + \gamma}{\sum_{m'} C^{CT}_{m't} + M\gamma} \frac{C^{TD}_{td}}{N_d} & \end{aligned} \quad (3)$$

$\tilde{z}_i = t$ represents the assignments of the ith concept in a document to topic t, $c_i = m$ represents the observation that the ith concept in the document is the mth concept in the lexicon, and $\mathbf{z}_{-\mathbf{i}}$ represents all topic assignments not including the ith concept. Furthermore, C^{CT}_{mt} is the number of times concept m is assigned to topic t, not including the current instance, and C^{TD}_{td} is the number of times topic t has occurred in document d, not including the current instance.

Parameters were estimated by averaging samples from 10 randomly-seeded runs, each running over 100 iterations, with an initial burn-in phase of 500 iterations (resulting in a total of 1.500 iterations). We found 500 iterations to be a convenient choice by observing a flattening of the log likelihood. The training time ranged from 10 hours to 15 hours depending on the data set (run on a standard Linux PC with Opteron Dual Core processor, 2.4 GHz). Instead of estimating the hyperparameters α, β and γ, we fix them to $50/T$, 0.001 and $1/M$ respectively in each of the experiments. We also fix the number of topics to $T = 400$. The values were chosen according to [27, 8].

3.3 Experiments

3.3.1 Data

Two large MEDLINE corpora, previously generated by [23, 22], were used to train the TC model. The first data set is a collection of PubMed abstracts randomly selected from the MEDLINE 2006 baseline database provided by the NLM. The collection consists of $D = 50000$ abstracts, $M = 17716$ unique MeSH main headings and a $N = 22531$ unique word stems. Word tokens from title and abstract were stemmed with a standard Porter stemmer [25] after stop words were removed using the PubMed stopword list. Additionally, word stems occurring less than five times in the corpus were removed. For each abstract MeSH main headings were linked to the corresponding concept in the MeSH thesaurus of 2006. Note that no filter criteria were defined for the MeSH vocabulary. The second data set contains $D = 84080$ PubMed abstracts, with $M = 18350$ unique MeSH main headings and a total of $N = 31684$ word stems and the same text-processing steps were applied. This corpus is composed of genetics abstracts from the MEDLINE 2005 baseline corpus. See [23, 22] for more information about both corpora. In the following, the data sets are referred to as *random 50K* data set and *genetics* data set respectively.

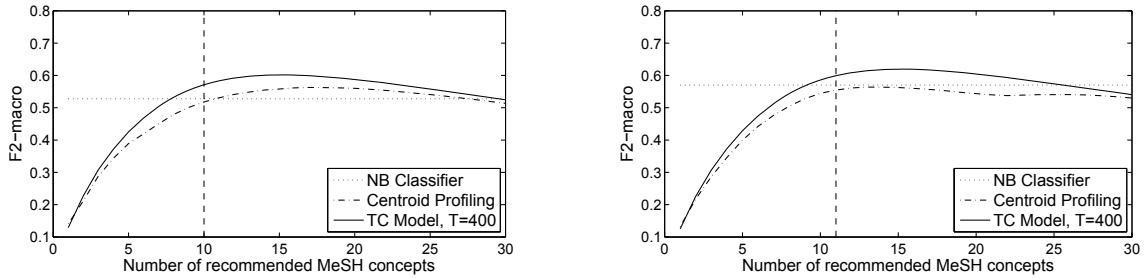
For each document in the training and test set, we prune each assigned MeSH descriptor to the first level of each taxonomy-subbranch resulting in 108 unique MeSH concepts. For example, if a document is indexed with the MeSH descriptor *Muscular Disorders, Atrophic [C10.668.550]* the concept is pruned to *Nervous System Diseases [C10]*. We believe that this setting represents a reasonable prerequisite for discipline-based indexing. Note that from a machine learning point of view, this is a very challenging 108 multi-label classification problem. In the pruned setting of our task, we have on average 9.6/10.5 (random 50K/genetics) pruned MeSH labels per document.

3.3.2 Results

The prediction of MeSH terms for unseen documents can be formulated as follows: based on the word-topic and concept-topic count matrices learned from the training data, the likelihood of a concept c given the test document d is $p(c|d) = \sum_t p(c|t)p(t|d)$. The first probability in the sum, $p(c|t)$, is given by the learned topic-concept distribution. The mixture of topics for the document $p(t|d)$ is estimated by drawing for each word token in the test document a topic based on the learned word-topic distribution $p(w|t)$. Thus, given an unseen document d, the TC model returns a ranked list of MeSH terms based on $p(c|d)$.

We benchmarked the performance of the TC model against a method called *centroid profiling* [12] as well as a multi-label naive Bayes classifier. Centroid profiling computes for each word token w_i and each MeSH concept c_j, in a training corpus, a term frequency measure $TF_{i,j} = w_{i,c_j} / \sum_{j=1}^{M} w_{i,c_j}$ with M equals to the total number of MeSH concepts. Thus, $TF_{i,j}$ measures the number of times a specific word w_i co-occurs with the MeSH concept c_j, normalized by the total number of times the word w_i occurs. As a consequence, each word token in the training can be represented by a vector consisting of the term frequency distribution over all M MeSH concepts. When indexing a new unseen document, the centroid over all word token vectors in the test document is computed returning a ranked list of MeSH terms. We assumed a bag of words representation for the multi-label naive Bayes classifier and trained it for each of the MeSH concepts (108 labels). We used a multi-variate Bernoulli Model for naive Bayes [18]. Recall, that in contrast to the TC model and centroid profiling approach, the multilabel naive Bayes classifier does not return a ranked list MeSH terms. For both data sets, models were trained on 90% of the documents and tested on the remaining 10%. The training and test splits are available online[4]. Figure 3 plots F2-macro measure against the number of recommended MeSH concepts. According to [7], we decided to weight recall over precision and average F2-macro measure over documents rather than over indexing categories. We evaluate the results until a cut-off value of 30 recommended MeSH concepts, except for the naive Bayes classifier. The TC model clearly outperforms the naive Bayes classifier and the centroid profiling with the best result at a cut-off value of 15 recommendations for both data sets (0.60 (random 50K)/0.62 (genetics)). Using a cut-off value which equals to the number of average MeSH assignments (rounded-up) in the two training corpora the F2-macro is 0.570 (random 50K) and 0.599 (genetics) for the TC model, while the centroid profiling reaches only 0.517 (random 50K) and 0.55 (genetics). The multi-label naive Bayes classifier reaches a F2 measure of 0.525 (random 50K) and 0.570 (genetics). Note that using the average number of MeSH assignments is the most simple way to determine an appropriate cut-off value. A more analytical way of determining the cut-off value would be to set up an independent development set for the given corpus and to maximize the F2-macro measure with respect to the number of recommendations.

[4] `http://www.dbs.ifi.lmu.de/~bundschu/research/data/`



(a) *random 50K corpus* (b) *genetics corpus*

Figure 3: F2-macro measure plotted as a function of the number of top n recommended MeSH terms. The vertical dotted line marks the average number of MeSH assignments in our experimental setting.

4. EMERGING TREND PREDICTION

A key component of our text monitoring system is the detection of emerging trends. This provides users with a key advantage in scouting tasks. Our text monitoring system tracks the frequency of all words over time to support the trend discovery. The user can search the document database and is not only returned a set of document clusters but also relevant emerging trends. We performed a study to investigate the feasibility of discovering trends by trying to predict the inclusion of important medical terms in the MeSH ontology. The goal is to automatically report new biomarkers that potentially represent a scientific breakthrough.

4.1 Data collection

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 [25]. 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*. The word stems and trends as well as the ground truth described below are available online [5].

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.

We utilized the temporal information associated with MeSH terms 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.

Similar to the documents, MeSH terms are associated with several dates. The creation date usually indicated the inclusion in the ontology and the established date indicate the date of the earliest annotated article. This can be earlier than the creation date, because the PubMed database is then retro-annotated by searching for relevant existing documents and adding new MeSH terms to those records. Unfortunately some terms did not have a created date when the MeSH ontology was restructured in 1999. They were assigned the creation date 01/01/1999. We ignored these terms in this analysis. The 2008 version of MeSH was used.

We searched the MeSH ontology for cancer related terms that were added within the time period under study 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 223 relevant trees with 759 relevant terms. From these 187 were removed because they had a creation date of 01/01/1999 and 65 were removed because they had a creation date before 01/01/1975. 524 of the remaining 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 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*. While we could have processed the document with n-grams this would have increased the complexity of the study tremendously. 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

[5] `http://www.mybytes.de/research/pubmed`

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

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

terms as true positives for cancer-related biomarkers. The distribution of these terms over time is shown in Figure 4 and some examples are listed in Table 2.

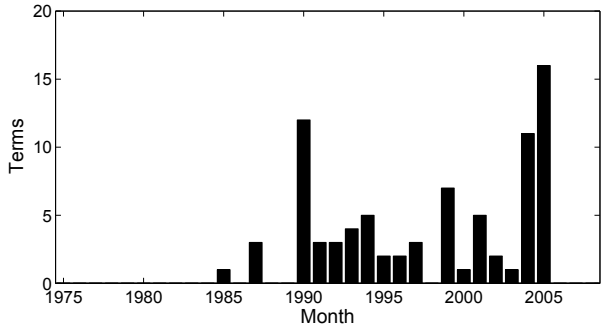


Figure 4: Addition of new Mesh terms describing biomarkers related to cancer.

Table 2: True positive cancer-related biomarkers.

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

4.2 Trend analysis

The number of cancer-related articles per month is shown in Figure 5. 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 6 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. A trend analysis system works

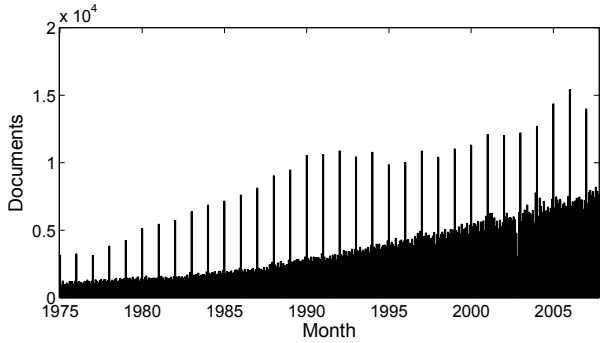


Figure 5: Number of cancer related documents per month in PubMed database. The frequency of publications is strongly increasing.

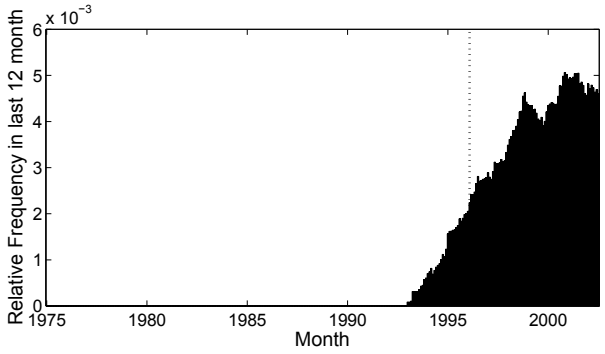


Figure 6: Trend of word stem *brca1* that corresponds to the MeSH term *Genes, BRCA1* added on 2/16/1996 (dashed line).

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:

1. Define time t_0 as the time of the first occurrence of term w, i.e. $f(w,t_0) > 0$ and $\forall t < t_0 f(w,t) = 0$
2. Consider an interval $I = [max(t_c - 24, t_0), t_c]$ - we consider at most the previous 24 months.
3. Select pairs $\{t, f(w,t) | t \in I, f(w,t) > 0\}$. If there are not such pairs, the score is undefined: $s(w,t_c) = NaN$.
4. Fit a line to pairs $\{t, log(f(w,t))\}$. The slope a of this line $log(f(w,t)) \approx at + b$ corresponds to the power of an exponential curve fitting these points and is used as the score: $s(w,t_c) = a$.

The scores are then used to rank trends - a higher score indicates a more interesting trend.

4.3 Evaluation

Due to computational costs of conducting the evaluation on all 180K terms, we consider a set containing all 81 positive examples and a randomly selected 10K other terms. In order

to evaluate our approach we consider several different types of measures, illuminating different aspects of the problem. Since in our application the trends are the behaviors of the terms, we will use "trends" and "terms" interchangeably.

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 7 shows the distribution of differences between these two events for $k = 100$ and $k = 200$. We detect 75 out of 81 terms with $k = 200$, and 66 with $k = 100$ and most of the trends are detected between 5 and 20 years in advance. This shows that our simple approach can detect true positives far in advance of their acceptance into MeSH. Obviously there are some costs attached. Considering 200 terms each month for 20 years could potentially lead to a total of 48000 terms - certainly a high cost for detecting only 81 true trends. It turns out however, that the real cost is significantly lower because terms appear in top-k sets multiple times. For example, only 5760 unique terms appear in top 200 from 01/1980 to 01/2004. 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 the user would have the option to remove in the actual system.

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.

- 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 8 we show precision and recall of our approach for $k = 100$ and $k = 200$. The fact that the precision stays between 1% and 10% (except in the end) is not surprising, since we only look at top k terms at each time point and because real trends are rare. For many time points this is significantly better than random: the overall expected value is less than 1% (81/10081) and the number of positive trends is commonly much smaller (at most 40). 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 somewhere between 5 and 35 positive trends are potentially detectable, and the recall is mostly greater than 20% (10% for $k = 100$) except for a few small regions of sharp drops and the final period which we discuss in more detail below.

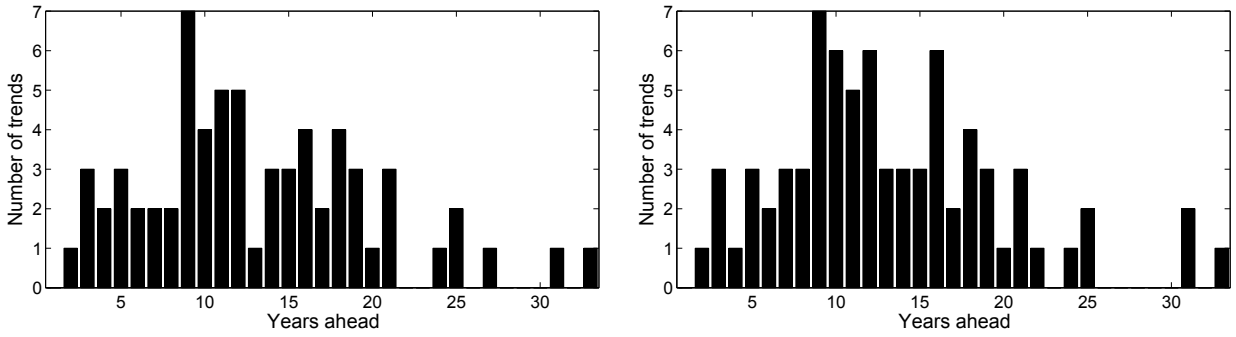
The performance drops in the last 40 months of the observed period. This occurs for the following reason. Many of true trends that are added to MeSH around 2005 have already slowed down their growth so they are no longer among the top k (though they were detected earlier). Meanwhile, since we reached 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 to be added to MeSH, but in the evaluation we are forced to treat them as false positives. In other words, the change in last period reflects the nature of our evaluation rather than the weakness of the method.

5. RELATED WORK

Modeling documents in a probabilistic framework via latent topics has first been introduced by [10], resulting in the so-called aspect model. However, the parameterization of the model was susceptible to overfitting. [4] addressed these limitations by introducing the LDA framework. Depending on the addressed generative process, the LDA framework has been extended e. g. to model the dependencies between authors, topics and documents [27] or the dependencies between author and recipients [17]. Further approaches include the modeling of images and their captions [3], the modeling of dependencies between topics and named entities [24] and the uncovering of latent topic structure hidden in biomedical text corpora [2].

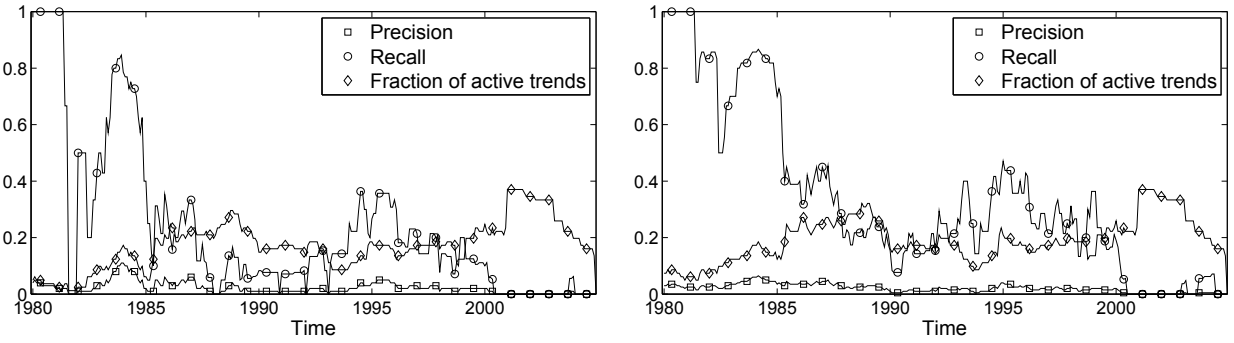
One of the earliest work mentioning trends in document archives is [15]. The frequency of phrases is tracked over time and the user can query for trends of predefined shapes. This could include recently emerging trends. No evaluation with ground truth was performed.

Many studies concentrate only on the main topics in a document archive and perform a retrospective analysis while we are interested in the early detection of new topics. [28] uses significance tests to detect time periods where words



$k = 100$: 66/81 trends detected k=200: 75/81 trends detected

Figure 7: Time difference in years between inclusion in MeSH and earliest detection for positive trends using the top k trends every month. Larger value indicates earlier detection.



$k = 100$ k=200

Figure 8: Precision, Recall, and Fraction of 81 positive trends within the time window (sum of true positives and false negatives) in period from January 1980 to January 2004.

have a higher than usual frequency. A temporal extension of LDA is proposed in [29] to better model evolving topics.

Even though the detection of interesting temporal phenomena is most useful in online settings where the data is continuously monitored, many studies have only described retrospective algorithms and leave online methods for future work. A state-based model is used in [14] 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 [6] 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. The burst detection is also used in [9] to emphasize features for subsequent text mining steps such as document clustering.

Several approaches use a partition of the time axis into large intervals to detect changes in document archives. Finite mixture models are used in [21] to represent the documents of each interval. A comparison of the models among subsequent time intervals is used to report new trends. In [19] 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 applied in [26] with the goal 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. While we have tried to predict the extension of the MeSH ontology this was mainly done to obtain a ground truth for the ranking evaluation. In our system we do not really care whether a trend is at some point included in MeSH or not, we just want to provide the user with pointers to emerging technologies as early as possible.

6. CONCLUSION

We have presented an integrated system for screening and monitoring of information feeds describing biomedical research. After an overview of the complete system we presented two particular modules that provide an added value using data mining methods.

The experimental evaluation on annotating abstracts shows the advantage of our generative approach. In the current setting, we restrict our experiments to subparts of the MeSH thesaurus. However, we plan to conduct experiments on the whole MeSH taxonomy. The extension of the generative process model for capturing the hierarchy of taxonomies is a matter of ongoing research.

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 de-

tection, 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. In the meantime the ranking is already deployed in our system and has received positive feedback. Some of the drawbacks revealed in our evaluation can be removed in a live system. For example, terms already in MESH or 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.

7. REFERENCES

[1] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. ACM Press, Addison-Wesley, 1999.

[2] D. M. Blei, K. Franks, M. I. Jordan, and I. S. Mian. Statistical modeling of biomedical corpora: mining the caenorhabditis genetic center bibliography for genes related to life span. *BMC Bioinformatics*, 7(1), 2006.

[3] D. M. Blei and M. I. Jordan. Modeling annotated data. pages 127–134, 2003.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[5] M. Bundschus, M. Dejori, S. Yu, V. Tresp, and H.-P. Kriegel. Statistical modeling of medical indexing processes for biomedical knowledge information discovery from text. *Submitted*, 2008.

[6] G. P. C. Fung, J. X. Yu, P. S. Yu, and H. Lu. Parameter free bursty events detection in text streams. In *Proc. 31st Intl. Conf. on Very large data bases*, pages 181–192, 2005.

[7] C. W. Gay, M. Kayaalp, and A. R. Aronson. Semi-automatic indexing of full text biomedical articles. In *AMIA Annu Symp Proc*, pages 271–275, 2005.

[8] T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proc Natl Acad Sci U S A*, 101 Suppl 1:5228–5235, 2004.

[9] Q. He, K. Chang, E.-P. Lim, and J. Zhang. Bursty feature representation for clustering text streams. In *Proc. SIAM Int. Conf. on Data Mining*, 2007.

[10] T. Hofmann. Probabilistic latent semantic analysis. In *Proc. of Uncertainty in Artificial Intelligence*, Stockholm, 1999.

[11] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, V42(1):177–196, 2001.

[12] S. M. Humphrey, T. C. Rindflesch, and A. R. Aronson. Automatic indexing by discipline and high-level categories: methodology and potential applications, 2000.

[13] T. Joachims. Optimizing search engines using clickthrough data. In *Proc. ACM Conf. on Knowledge Discovery and Data Mining*, 2002.

[14] J. Kleinberg. Bursty and hierarchical structure in streams. In *Proc. 8th ACM SIGKDD Intl. Conf. on Knowledge discovery and data mining*, pages 91–101, 2002.

[15] B. Lent, R. Agrawal, and R. Srikant. Discovering trends in text databases. In *Proc. 3rd Int. Conf. Knowledge Discovery and Data Mining*, pages 227–230, 1997.

[16] C. D. Manning and H. Schutze. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.

[17] A. McCallum, A. Corrada-Emmanuel, and X. Wang. Topic and role discovery in social networks. 2005.

[18] A. McCallum and K. Nigam. A comparison of event models for naive bayes text classification, 1998.

[19] Q. Mei and C. Zhai. Discovering evolutionary theme patterns from text: an exploration of temporal text mining. In *Proc. 11th ACM SIGKDD Intl. Conf. on Knowledge discovery in data mining*, pages 198–207, 2005.

[20] F. Mörchen, K. Brinker, and C. Neubauer. Any-time clustering of high frequency news streams. In *Proc. Data Mining Case Studies Workshop, KDD*, 2007.

[21] S. Morinaga and K. Yamanishi. Tracking dynamics of topic trends using a finite mixture model. In *Proc. 10th ACM SIGKDD Intl. Conf. on Knowledge discovery and data mining*, pages 811–816, 2004.

[22] A. Névéol, S. E. Shooshan, S. M. Humphrey, T. C. Rindflesch, and A. R. Aronson. Multiple approaches to fine-grained indexing of the biomedical literature. In *Pacific Symp. on Biocomputing*, pages 292–303. World Scientific, 2007.

[23] A. Névéol, S. E. Shooshan, J. G. Mork, and A. R. Aronson. Fine-grained indexing of the biomedical literature: Mesh subheading attachment for a medline indexing tool. In *Proc. AMIA Symp*, 2007.

[24] D. Newman, C. Chemudugunta, and P. Smyth. Statistical entity-topic models. In *Proc. 12th ACM SIGKDD Intl. Conf. on Knowledge discovery and data mining*, pages 680–686, New York, NY, USA, 2006.

[25] M. F. Porter. An algorithm for suffix stripping. pages 313–316, 1997.

[26] R. Schult and M. Spiliopoulou. Discovering emerging topics in unlabelled text collections. In *Proc. East European ADBIS Conf.*, pages 353–366, 2006.

[27] M. Steyvers, P. Smyth, M. Rosen-Zvi, and T. Griffiths. Probabilistic author-topic models for information discovery. In *Proc. of the 10th ACM SIGKDD Intl. Conf. on Knowledge discovery and data mining*, pages 306–315, 2004.

[28] R. Swan and J. Allan. Automatic generation of overview timelines. In *Proc. 23rd Intl. ACM SIGIR Conf. on information retrieval*, pages 49–56, 2000.

[29] X. Wang and A. McCallum. Topics over time: a non-markov continuous-time model of topical trends. In *Proc. 12th ACM SIGKDD Intl. Conf. on Knowledge discovery and data mining*, pages 424–433, 2006.