Trend Analysis on Medical Treatments and Topics in Clinical Trials

Yu-chen Tuan
12/14/2012

Motivation

In the 21th century, many people are suffered from various kinds of chronic diseases without effective medicines to cure. The trend analysis on medical treatment and topics can benefit not only the professionals in biomedical domain but also those individuals who want to educate themselves and to discover whether there is an alternative solution or growing trend for curing a specific disease. The project exposes the trend information and builds up a basic framework for future extended research.

Abstract

There are two parts of trend analysis in this project. The first part is to measure medical treatment trend given a disease or condition. The second part is to extract topics from clusters and predict the topic trend. The analysis is performed against of 130,000 clinical trial articles. Technically, natural language processing and machine learning techniques are applied to the work, such as linear regression, locally weighted linear regression and K-means clustering algorithm. Programming tools such as Java-based Lucene indexing engine is also applied in order to facilitate the processing. Future work for improvement such as named entity recognition is recommended.

Introduction

The analysis is based on clinical trial articles, which can be downloaded from ClinicalTrials.gov[1]. The website is a service of US National Institutes of Health that contains over 130,000 articles. In a clinical trial (also called an interventional study), participants receive specific interventions according to the research plan or protocol created by the investigators. These interventions may be medical products, such as drugs or devices; procedures; or changes to participants' behavior, for example, diet. Clinical trials may compare a new medical approach to a standard one that is already available or to a placebo that contains no active ingredients or to no intervention. When a new product or approach is being studied, it is not usually known whether it will be helpful, harmful, or no different than available alternatives. The investigators try to determine the safety and efficacy of the intervention by measuring certain outcomes in the participants.

There have been many researches in trend analysis. Tomonari et al. [2] proposes the method of adopting and extending Latent Dirchlet Allocation (LDA) for analyzing reference relationship among scientific articles. Nallapati et al. [4] proposes topic-flow model that basically combines ideas from network flow and topic modeling. Ren et al. [3] proposes dynamic hierarchical Dirchlet process to model the time-evolving statistical properties of sequential data sets. The statistical properties of data collected at consecutive time points are linked via a random parameter that controls their probabilistic similarity.

It seems the project is a forerunner project, because it is hard to find research papers in trend analysis on medical treatments. In the project, a foundation framework is built up as the first step. In the future, more sophisticated algorithms and methodologies can be applied.

Each clinical trial article is in XML format where multiple fields are defined including disease/condition and intervention.   Those two fields are extracted for disease and treatment information, respectively. Furthermore, many other fields are extracted such as title, description and summary. Co-occurred terms plays an important role in natural language processing. For example, given the sentence, "smoking can cause lung cancer". "Smoking" and "lung cancer" are co-occurred terms which indicate relationship through the verb "cause". By analyzing the disease and intervention fields along with co-occurred terms in summary, title and description fields, medical treatment trend can be exposed.

A cluster of terms represents the topic of the cluster. In this project, K-means clustering algorithm is applied first before extracting the topic and computing the trend. For better topic granularity, we classify terms inside the cluster into categories by year distribution. By looking at terms in each category, we can discover how topics or terms are changed year after year.

Detailed description is described in the following methods section and results section.

Methods

Multiple techniques have been applied to the projects.

XML Parsing and Lucene Indexing

Clinical trials downloaded from clinicaltrials.gov are in XML format, which has well defined fields, including disease condition, intervention, title, summary, description, study start date, study completion date and primary completion date information. For the date field, we only choose study start date as our temporal information. The fields and values are extracted into key and value pairs and indexed using Lucene indexing engine. Lucene default standard analyzer is applied without being customized in the project.

Lucene search query and co-occurred term relatedness

After indexing, search query and relatedness computation are performed by given any two terms. The project focus is to find the trend on disease/condition term and medical treatment term, so any disease term such as "Rheumatoid Arthritis" can be given and treatment terms "fish oil" can be given.

The relatedness of those two terms are computed using Pointwise Mutual Information (PMI). PMI indicates the strength of relatedness between terms. The equation is described in the following.

PMI = log (p(term1, term2) / (p(term1) p(term2)))

Assuming N is total number of docs and DF is a function of document frequency.

p(term1) = DF(term1) / N
p(term2) = DF(term2) / N
p(term1, term2) = DF(term1 AND term2) / N

Above PMI equation is revised in the following in order to incorporate temporal information:

PMI = log (p(term1, term2, year) / (p(term1) p(term2) p(year)))

The value of DF can be easily acquired by counting the size of the search query result set through Lucene search engine. The total number of documents also can be easily acquired using Lucene API.

Using Lucene to store the data for computing the PMI is not only for quick implementation but also for memory consideration. The project runs under 32-bit desktop machine, instead of 64-bit. However, there are 130,000 clinical trial articles which can contain lots of terms in the text (There are over one million unique terms in biomedical domain.). If storing DF values for all terms into a matrix, out of memory problem will most likely occur. However, by default lucene only store only one term into memory for every 128 terms in its dictionary. Therefore, Using Lucene can achieve the memory efficiency.

Linear Regression, Locally Weighted Linear Regression using Matlab

After PMI are generated, linear regression and locally weighted linear regression are applied to generate analysis reports.

Topic Extraction and Topic Trend Analysis

Topic is extracted and trend of the topic is measured through the following steps.

1. At first, a term vector is constructed for each document of the whole clinical trials, where the TF.IDF based term weight is retrieved from Lucene index. Each term vector contains various length of term-weight pairs, including the investigation starting time of the trial.

2. Then K-means clustering algorithm is applied to generate clusters. The number of clusters is set 200. Then for each cluster, classify vectors into categories by year distribution. Then for each category, compute the center vector (mean).

3. The term vector of the mean in each category is sorted by the term weight in descending order. After removing noisy terms, we extract only top 20 highly weighted terms to represent the topic of the category. The sum of the weight of those top 20 terms represents relevant importance of the topic in the cluster to all topics in the

whole clusters. Common terms, such as "study", "clinical" etc are considered noisy terms, because they don't expose significant information. Therefore, they are taken out from the list.

The K-means clustering program runs over two days under a 32-bit machine.

Results

Medical Treatment Trend

One trend is measured at a time by choosing only one disease condition and only one medical treatment. In our cases, we measure two trends by choosing "arthritis" as the disease condition and "stem cell" and "fish oil" as one of the corresponding medical treatments. Figure 1-2 shows the trends for each case. The red points shows the trend prediction, developed using locally weighted linear regression algorithm.

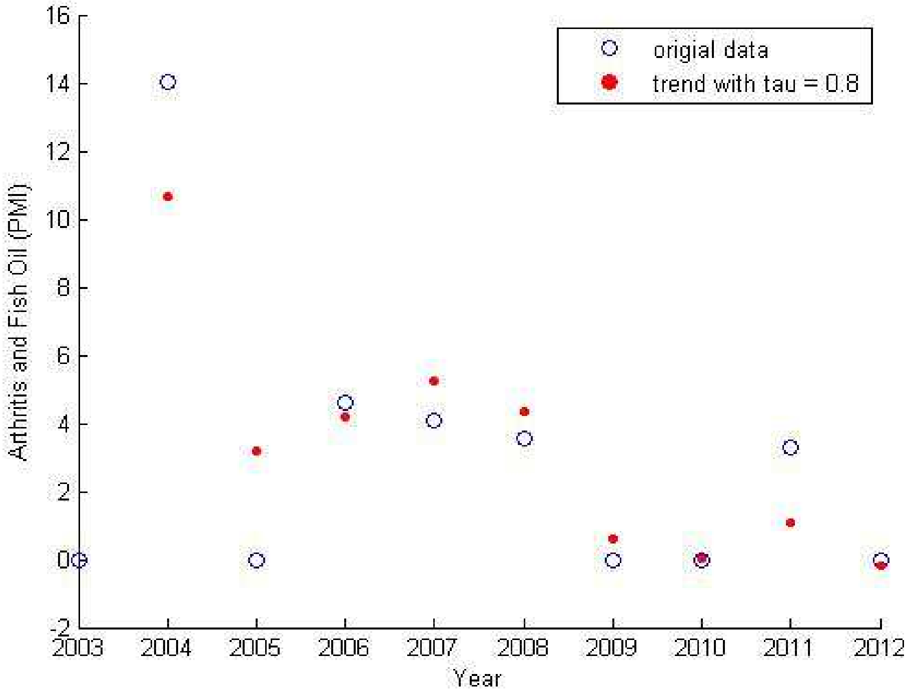


Fig. 1 Trend of arthritis treatment using fish oil.

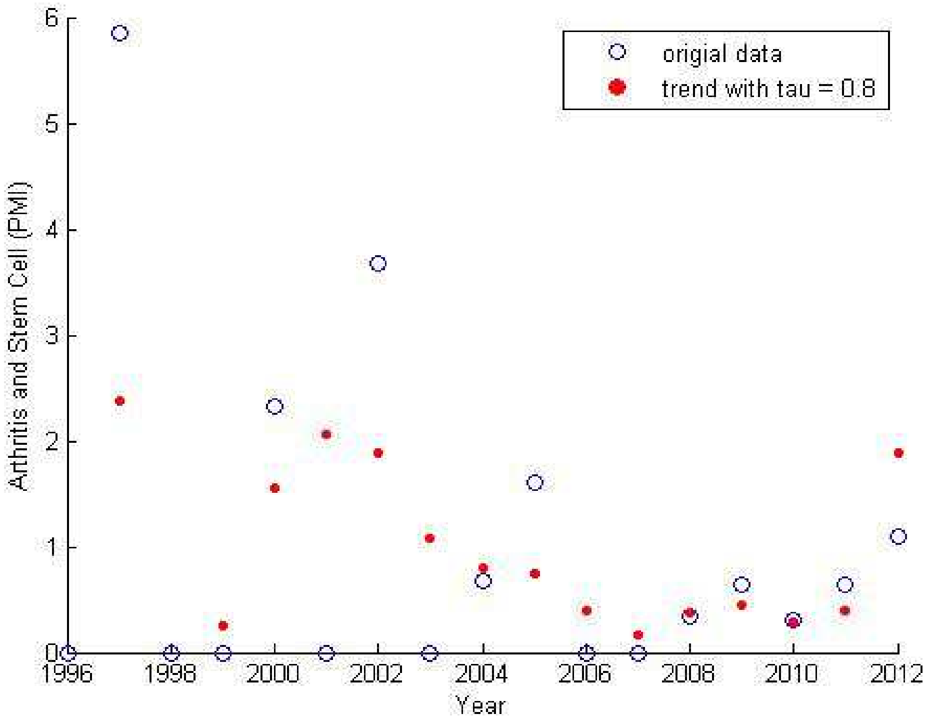
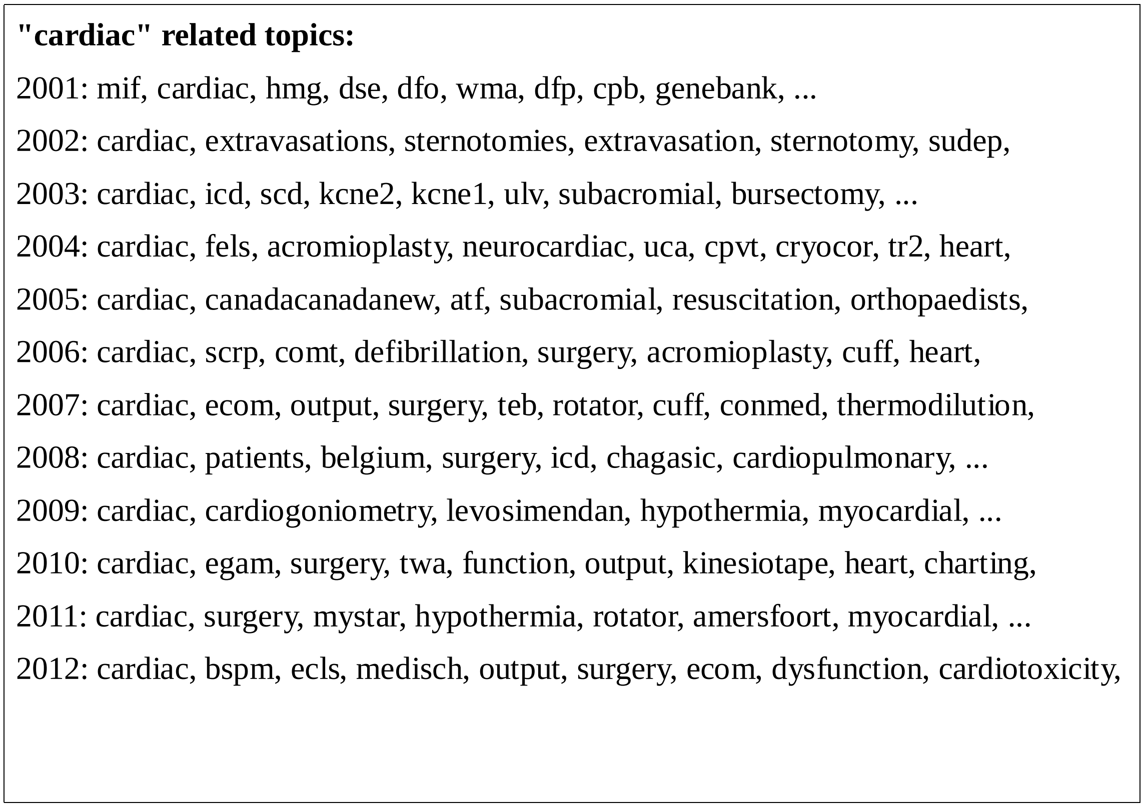


Fig. 2 Trend of arthritis treatment using stem cell.

Topic Extraction and Topic Trend

A cluster of terms represent the topic of the cluster, and the topic trend in each cluster indicates how important of that topic relevant to the whole clinical trial corpus. In the following, we use three clusters related to, cardiac, vaccine and implant, respectively to demonstrate our results in topic extraction and topic trend analysis.

Fig. 3 shows "cardiac" related topics categorized by years. Fig. 4 shows "cardiac" related topic trend across years. Both linear regression and locally weighted linear regression are applied for the analysis. From the graph we know that the trend goes down. Fig. 5 and 6 show "vaccine" related topics and trend. Fir. 7 and 8 show "implant" related topics and trend.



"cardiac" related topics:

2001: mif, cardiac, hmg, dse, dfo, wma, dfp, cpb, genebank, ...

2002: cardiac, extravasations, sternotomies, extravasation, sternotomy, sudep,

2003: cardiac, icd, scd, kcne2, kcne1, ulv, subacromial, bursectomy, ...

2004: cardiac, fels, acromioplasty, neurocardiac, uca, cpvt, cryocor, tr2, heart,

2005: cardiac, canadacanadanew, atf, subacromial, resuscitation, orthopaedists,

2006: cardiac, scrp, comt, defibrillation, surgery, acromioplasty, cuff, heart,

2007: cardiac, ecom, output, surgery, teb, rotator, cuff, conmed, thermodilution,

2008: cardiac, patients, belgium, surgery, icd, chagasic, cardiopulmonary, ...

2009: cardiac, cardiogoniometry, levosimendan, hypothermia, myocardial, ...

2010: cardiac, egam, surgery, twa, function, output, kinesiotape, heart, charting,

2011: cardiac, surgery, mystar, hypothermia, rotator, amersfoort, myocardial, ...

2012: cardiac, bspm, ecls, medisch, output, surgery, ecom, dysfunction, cardiotoxicity,

Fig. 3 "Cardiac" related topics across years

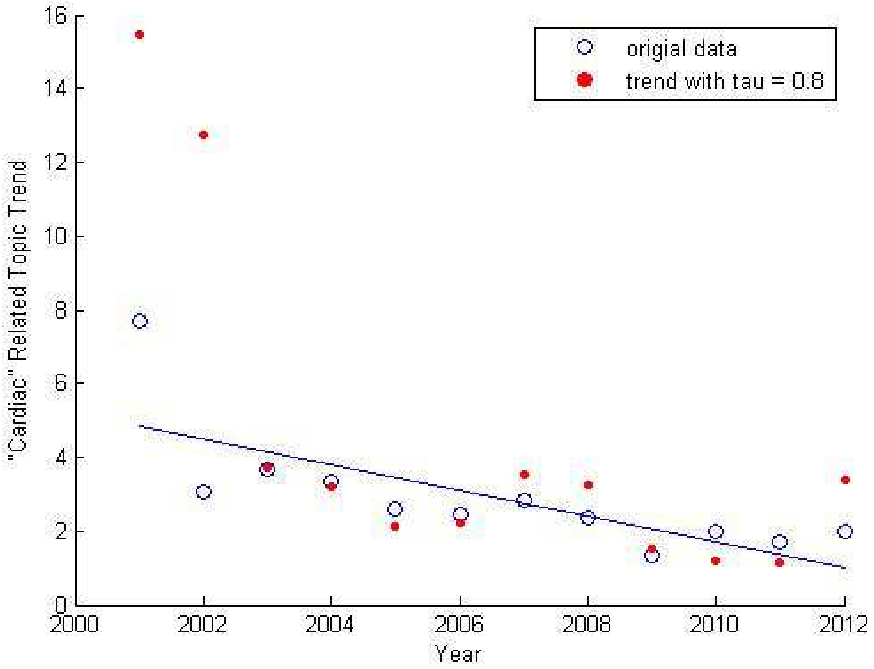


Fig. 4 "cardiac" related topic trend

"vaccine" related topics:

1997: mncc, meningococcal, vaccine, conjugate, production, antibodies, mmr, ...

1998: dtpa, hib, ipv, hbv, infanrix, biologicals, penta, hiberix, pentetic, hexa, beecham,

1999: dtpa, hib, ipv, hbv, infanrix, biologicals, penta, hiberix, pentetic, hexa, ...

2000: vaccine, pneumococcal, valent, conjugate, crm, antibody, ps, om, uk, infants, ...

2001: hib, vaccine, carriage, vaccines, medicin, rabies, alaska, conjugate,

2002: pneumococcal, ppv, vaccine, alaska, conjugate, pcv, natives, transplant, ...

2003: vaccine, pneumococcal, hib, typeb, pneumovax, conjugate, ppv23, ...

2004: w135, pneumococcal, vaccine, polysaccharide, 11pn, infanrix, hexa, conjugate,

2005: pneumococcal, vaccine, hib, conjugate, vaccination, dtpw, valent, menc, pnc,

2006: vaccine, pneumococcal, conjugate, 23vppv, repa, pneumum, acyw135, carriage,

2007: vaccine, menitorix, pneumococcal, conjugate, polysaccharide, booster, hib, ...

2008: pneumo23, vaccine, pneumococcal, conjugate, tetramen, vaccination, booster,

2009: vaccine, pneumococcal, conjugate, vaccination, gsk1024850a, menbvac, pps,

2010: pneumococcal, vaccine, pcv13, pcv7, conjugate, pcv10, dtap, carriage, ...

2011: pneumococcal, vaccine, pcv13, png, meningitis, conjugate, study, pneumoniae,

2012: menafrivac, polysaccharide, vaccine, meningococcal, conjugate, feta, vaccines, Fig. 5 "vaccine" related topics across years

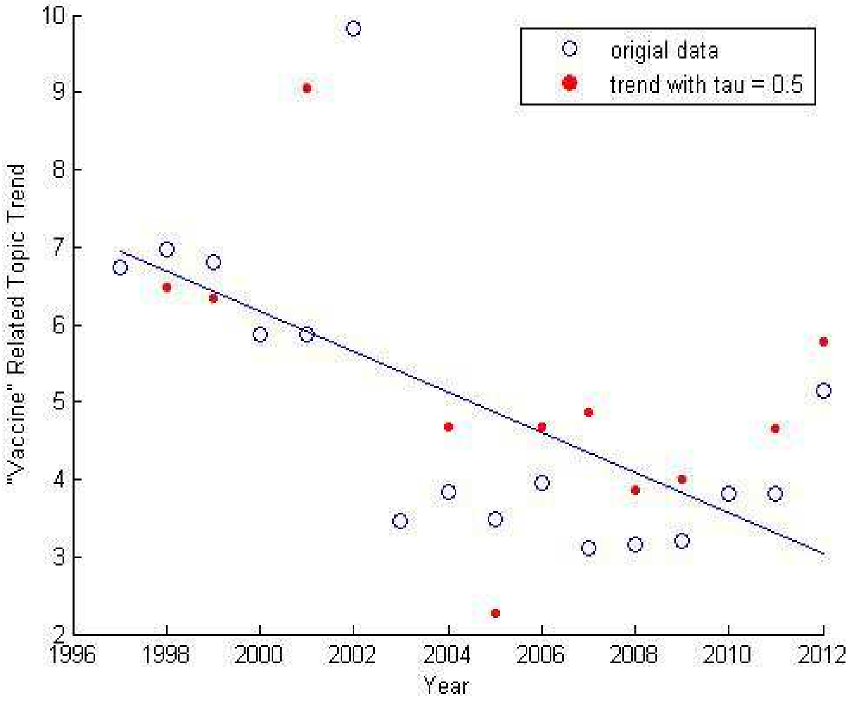


Fig. 6 "vaccine" related topic trend

"implant" related topics:

2002: sientra, implant, intravitreal, fluocinolone, acetonide, uveitis, ...

2003: temporized, endodontically, implant, telamon, crestal, extraction,...

2004: implant, straumann, lacrimal, polycaprolactone, tear, overdenture, ...

2005: tshr, implant, osseotite, hyperthyrotropinemia, edentate, lateralized, ...

2006: pess, anophthalmic, implant, orbital, vantas, sockets, enucleation, ...

2007: implant, microanastomosed, bone, brantigan, buccopharyngeal, ...

2008: implant, radiolucency, nanotite, osseospeedtm, isq, osteotome, bone, dental,

2009: implant, picf, endodontically, sensing, implantitis, patients, vf, posts, ...

2010: implant, implants, polycaprolactone, bone, dentures, denture, sinus, ...

2011: implant, osseospeedtm, fp8, nexplanon, abutment, countertorque, bone,...

2012: implant, baha, socket, tissue, cntf, esthetic, bone, cylinders, overdenture, gbr,

Fig. 7 "implant" related topics across years

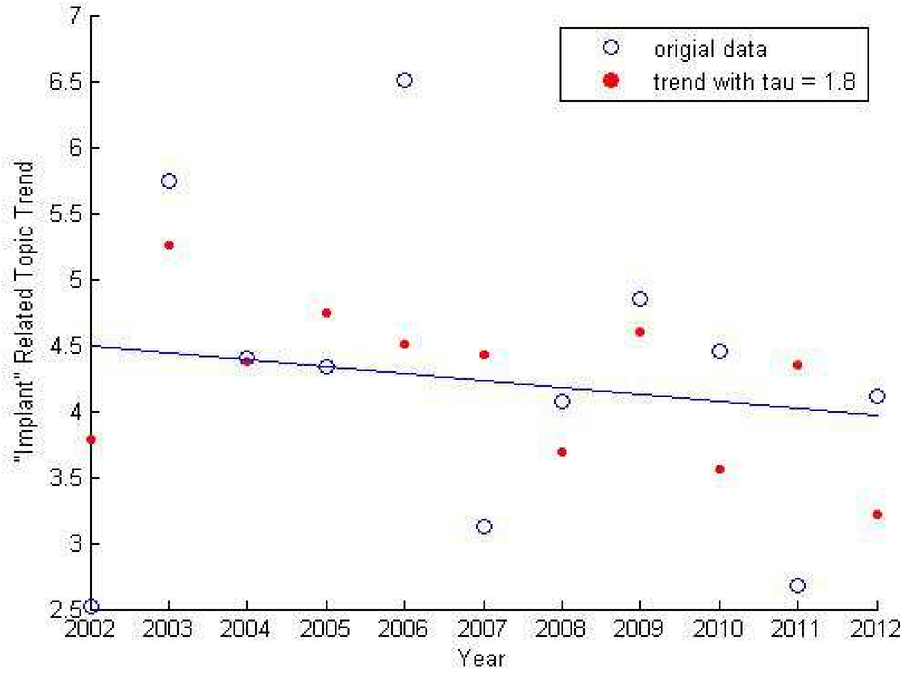


Fig. 8 "implant" related topic trend

Conclusion and Future Work

Multiple improvements can be done naturally in the next step.

Named Entity Recognition and Semantic Distance Measurement

The trend analysis currently doesn't consider of matching synonyms or near concepts. For example, we don't measure how close of the meaning of "arthritis" to "polyarthritis". So during clustering process, those two terms are considered orthogonal. Thus, it is possible that a document containing "arthritis" and another document containing "polyarthritis" are not fall into the same cluster.

Named Entity Recognition (NER) is a process of annotating concepts on those terms meaningful to the system. With NER, synonyms will be annotated with the same concept. Thus, term like "bursitis" is the same as "frozen shoulder". Therefore, PMI will be more accurately measured. The semantic distance measurement (SDM) is to measure how close of two concepts are each other. With SDM, the clustering quality can be improved if the term weight TF.IDF is revised by plugging SDM. The NER can be achieved using UMLS thesaurus[6]. The other advantage of using NER and SDM is to reduce the term vector dimension so as to contribute to the computational performance.

Latent Dirchlet Allocation and Hierarchical Clustering

Latent Dirchlet Allocation (LDA) [5] is an unsupervised generative model based clustering algorithm. It has been popularly applied to clustering and trend analysis. The main difference between LDA and K-means is LDA is soft link and K-means is hard link based clustering. It will be interesting to compare the quality between LDA and K-means.

No matter which clustering algorithm is used, converting the clustering algorithm into a hierarchical clustering is an important step in order to have better computational speed and quality. For example, currently for K-means, it is set 200 clusters. To determine the cluster the document will fall into, we need to compute the distance to each cluster center, meaning the computation complexity is O(200). If converting the clustering algorithm into a hierarchy, instead of setting 200, we can set maximal 10 clusters in each node of the hierarchy. The distance is measured in top-down fashion by traversing the hierarchy. Because each node has at most 10 clusters, the computation complexity could be less than O(200) when the best cluster of some node is found in the hierarchy.

References

[1] ClinicalTrials. http://www.clinicaltrials.gov/
[2] Tomonari Masada, Atsuhiro Takasu. Extraction of Topic Evolutions from References in Scientific Articles and Its GPU Acceleration. In CIKM’12, October 29–November 2, 2012.
[3] Lu Ren, David B. Dunson, Lawrence Carin. The Dynamic Hierarchical Dirichlet Process. In Proceedings of the 25 th International Conference on Machine Learning, Helsinki, Finland, 2008.
[4] Ramesh Nallapati and Christopher Manning. TopicFlow Model: Unsupervised Learning of Topic-specific Influences of Hyperlinked Documents. In Internet (no publication)
[5] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. JMLR, 2003.
[6] UMLS Thesaurus, https://uts.nlm.nih.gov/home.html