Final Project Report

F1303026 5130309756 Yiyi Shen

June 25, 2016

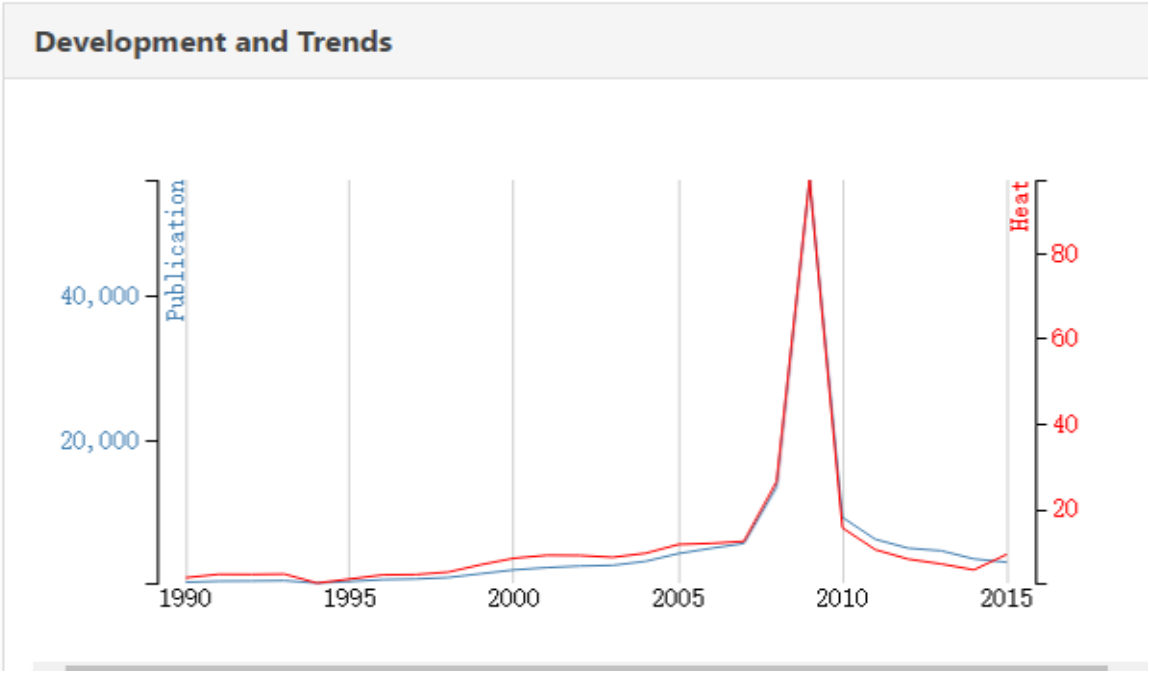
1 Background

How to leverage information technologies to improve the productivity of scientific research is a highly important challenge with clearly huge impact on the society. One bottleneck in research productivity is that as a research community grows, it would be increasingly difficult for researchers to see the complete picture of how a field has been evolving. That may cause several problems: junior researchers can often get lost in the overwhelming amount of related papers; while researchers who seek to shift to a new topic may spend lots of time preparing a reading list on his or her own.

All these clearly hinder the progress of scientific research, and it would be highly beneficial to develop mining techniques to help researchers more easily and more efficiently understand research themes in scientific literature. Thus, it leads to quite a lot of works to do and we would like to show: when did the topic become popular and is it still attracting attention today?

2 Motivation

There is already some works been done on on Acemap web page, and here gives out one current graph of topic 'Data Mining' in *Fig*.1:



We can find there are two lines on the graph: publication and heat. Publication means the number of papers published in this year and heat means how popular this field or topic is in this year.

There are two obvious problems on this graph. First, we can find these two lines are almost overlapped, which arouses a question: is heat simply equal to publication? As we all know, writing papers is a tough work and it has a time delay, which means, before a paper is published, you need to work seriously for a very long time until a new idea comes to your mind and then you should continue working to implement it, perfect it and do lots of experiments to confirm it. So there should be a time lag for publication according to the heat. What's more, sometimes though lots of papers are published, they have small influence. Few people care about it, few people cite it. Maybe, the influence of hundreds of such papers have on one topic is even less than that of a very important paper. Thus, there are lots of factors need to be considered besides publication.

The second problem is obviously cold boot. We can find that in this graph after year 2010, the publication and heat rapidly decreased, which gives rise to the incompleteness of our database rather than imply the topic is not popular any more. Usually it is hard to cover most papers published in recent years without official authorization and it is a common problem for many system.

For above reasons, we determine to do something make it better.

3 Model

We need to dig information other than publication numbers to build our model and there are many existing models such as LDA, PLSA and so on to extract information from words in a document. LDA (latent Dirichlet allocation), is a generative statistical model widely used in natural language processing. In conventional LDA, each word in document may be viewed as belonging to one topic with particular probability while each topic, with particular probability as well, choose a word.

However, empirical study has already noticed that it is a difficult task to annotate for each word its belonging topic even manually and it is usually computational complex. Thus we try to change our mind. We notice that for citations in a published paper written by experienced authors, it would be much easier to annotate since most authors make citations prudently and thus citations are much less noisy than text. Another advantage of using citation is that since our database is very large, use citation rather than words can greatly reduced the computational complexity because the number of citations is much less. What's more in our database, papers and abstract are not completed. Sometimes Acemap only gives out a link to a paper rather than a full paper. Thus we finally decide to use citations information instead of words.

In the paper we referred**?**, the model aims at finding two distributions between documents and topics. One is a probability distribution over topics conditioned on document d and the other is a reverse conditional distribution of documents given a topic. It shows how a topic is characterized by a set of documents that are cited, which are useful to our purpose. These two distributions can be obtained by LDA topic model. But it can be too complex and high time cost in large database. Fortunately, Acemap has great cooperations with Microsoft and they provide us a database which contains much information.

The Microsoft database has already classified, layered and ranked these academic data. Papers are classified to different topics with corresponding confidence and ranks and the topics are further layered according to scientific branches to 19 L0 topics, 290 L1 topics and so on. We then based on the 290 L1 topics and construct our model referring to citation-LDA. We define the second distribution ϕ as: when given a topic, we fetch all

papers belonging to it and its corresponding citations with other information such as paper year, paper ID, paper rank, confidence for further use.

$$\phi_k = \{d_k^{(i)}, c_k^{(i)} | z_k\}$$

where z_k is the topic k. $d_k^{(i)}$ is a paper i belonging to topic k with other information and $c_k^{(i)}$ stands for the citations for paper $d_k^{(i)}$.

4 Topic Strength

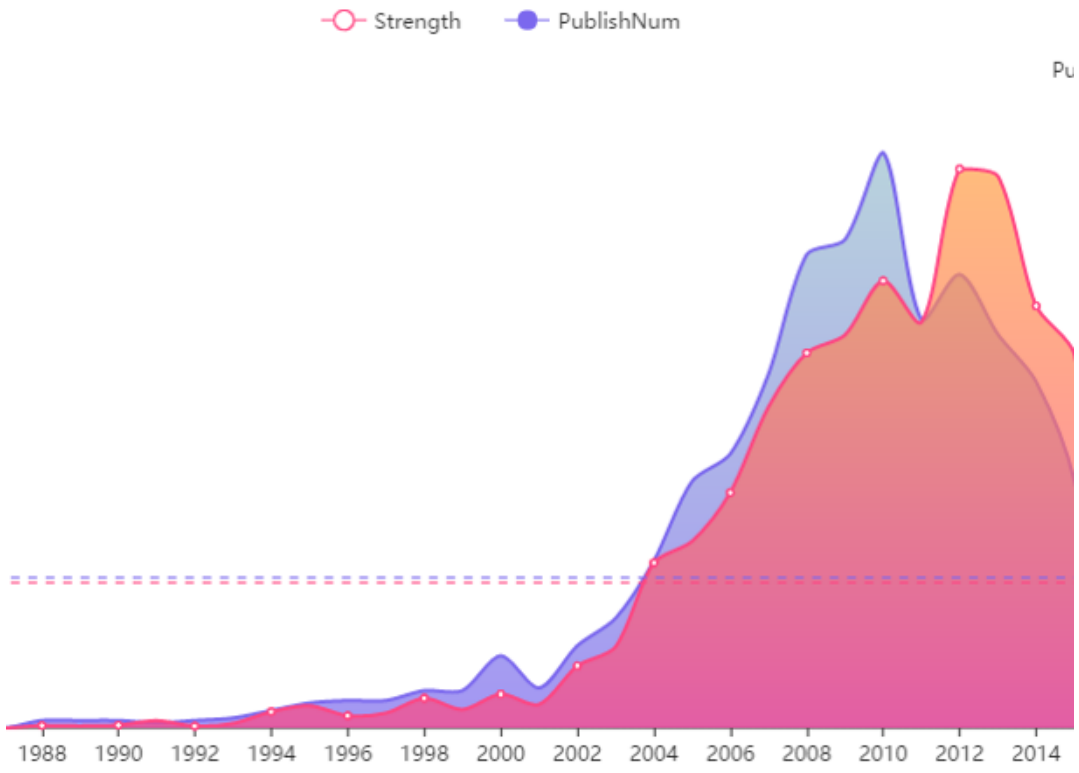
Topic strength, like heat, is defined as to reveal the relative popularity of topics at different times, which can help users to identify current and previous research topics as well as the rough topic life spans.

Within the topic distributions we defined above, we propose a metric to evaluate the topic temporal strength $TS(z_k, t)$ for year t, topic k. In our analysis, we cover the impact of publish numbers, paper rank, paper relations, citations for a particular publish year to make a comparatively all-round evaluation. The first term stands for the influence of all papers d_{kt} belonging to topic k published in year t to topic popularity and the second term represents the influence of papers cited by papers d_{kt}.

$$TS(z_k, t) = \sum_{d_{kt}^{(i)} \in z_k} r_k^{(i)} c_k^{(i)} + \alpha \sum_{t' \leq t} \sum_{d_{kt}^{(j)} \in z_k} \frac{D(d_{kt}^{(i)}, d_{kt}^{(j)}) r_k^{(j)} c_k^{(j)}}{\sum_{d^{(j')} \in C(d_{kt}^{(i)})} D(d_{kt}^{(i)} d^{(j')})}$$

where $d_{kt}^{(i)}$ is a paper belonging to topic k published in year t with its paper rank $r_k^{(i)}$ and confidence to the topic $c_k^{(i)}$. $D(d^{(i)}, d^{(j)})$ shows the dependence degree between paper $d^{(i)}$ and $d^{(j)}$. $C(d^{(i)})$ is the papers set cited by paper $d^{(i)}$.

Here we show a graph in *Fig.*2 obtained by our model for topic 'Data Stream Mining' as below.



Compared with current graph on Acemap web page, since we add the influence of papers cited by current year papers and many other information to our matric of topic strength evaluation, obviously in our graph, time delay in academic paper publication is concluded and the cold boot problem has been promoted to some extent as well(Notice that this decrease can't be entirely eliminated since it is caused by data absence in database).

5 Visualization

Considering that statistic numbers are not attractive, we decide to demonstrate it through a more fascinating and visualized way.

Home page(Fig.3):

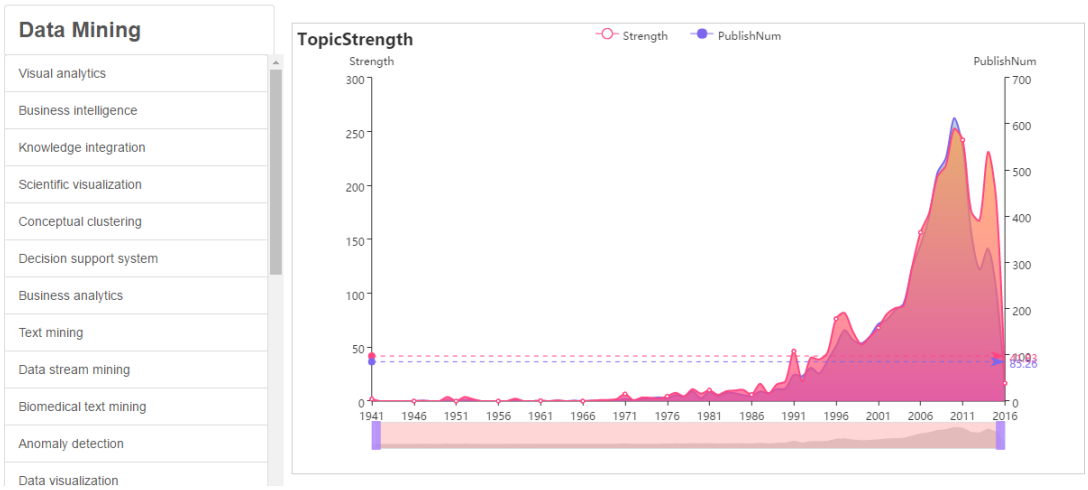
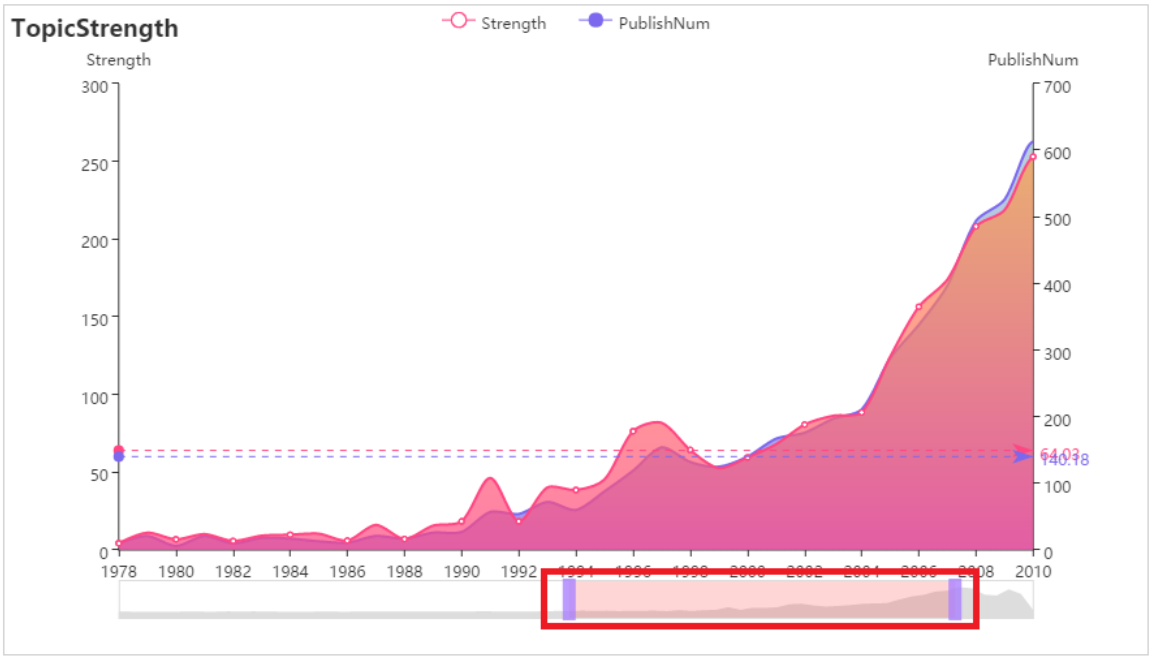
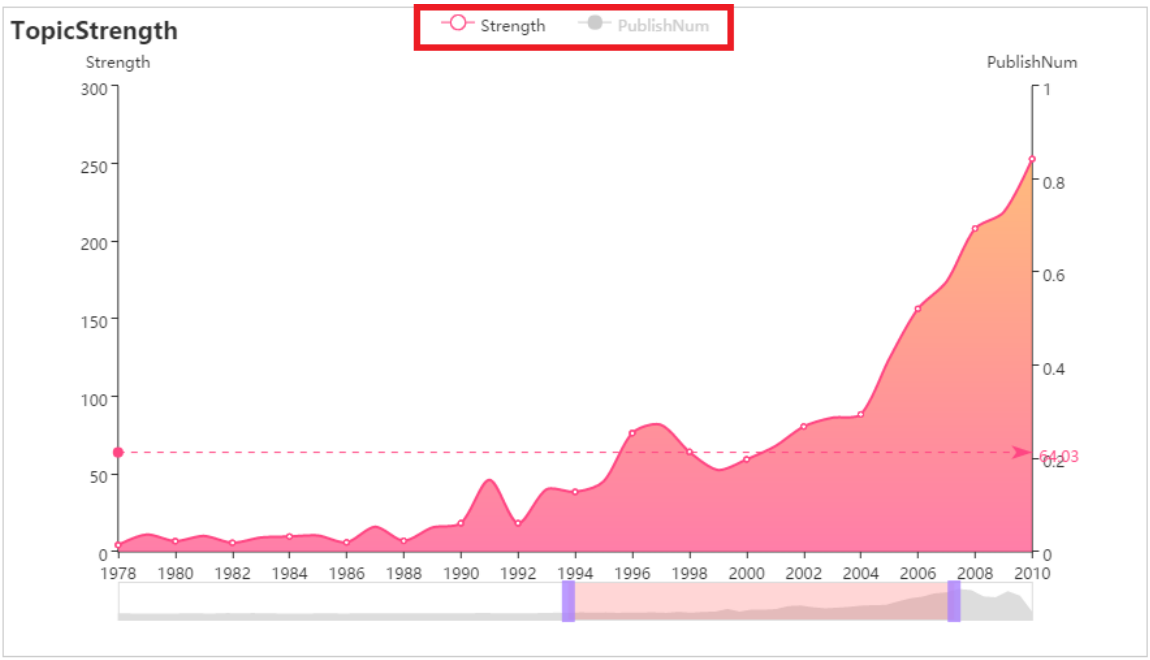


Fig.3 shows the topic strength of 'Data mining'. The subtopics of it are listed at the left column. We can click each of them to check corresponding subtopic popularity.

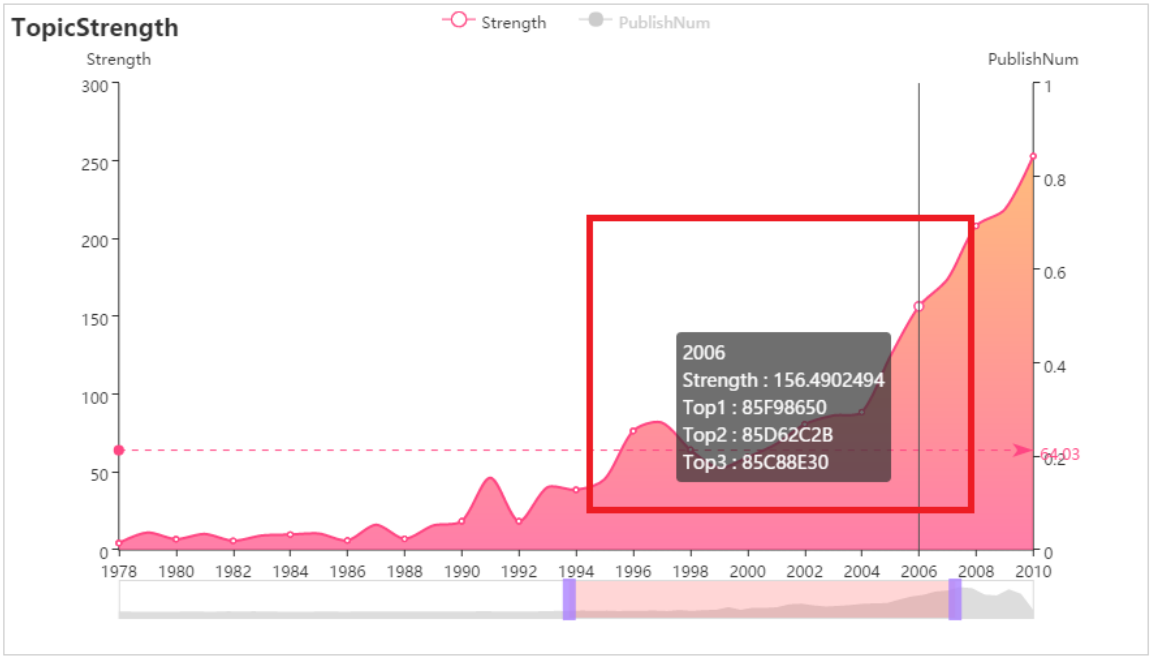
In the strength graph, we also add some user-friendly widgets. Users can focus on part of strength graph in a specific year range by scrolling the time bar at the bottom of the graph(shown in Fig.4).



What's more, users can also choose to only shown the line he/she would like to focus on by clicking corresponding legend.(shown in Fig.5)



Finally, when users click the data point of a specific year, we will show him/her the top three papers published in that year. This function is useful to help users have a further knowledge about one topic.(shown in Fig.6)



6 Future works

We still face some problems in our work. First, since our database is too large, it cost us a lot of time to get the distribution, thus it is hard to be real time. A second problem is that we lack an authoritative evaluation metric to measure whether the topic strength

we obtained is consist with the true condition. Thus we can only ask professors to help do a approximate manual check.

So there is still many future works waiting for us.

7 Reference

1.*Understanding Evolution of Research Themes: a Probabilistic Generative Model for Citations, Xiaolong Wang, Chengxiang Zhai and Dan Roth, Department of Computer Science University of Illinois, Urbana-Champaign Urbana*

2. Mapping the Topic Evolution Using Citation-topic Model and Social Network Analysis, Chunlei Ye, Dongmei Liu, Na Chen, Li Lin, Information Consultation Department Library of Beijing University