Thinking and Programming on Academic Search Engine

WenTian Bao, School of Electronic, Info. & Electrical Engineering, Shanghai Jiao Tong University

***Abstract*—Nowadays, the academic search of some popular search engines such as Google, performs not as good as we expected in visualization, topics' development and the recommendation. Therefore, we want to develop a new search engine, which can give more proper recommendation according the users' educational background and research interest. We also want to give the development of topics and make knowledge map using big data and the visualization software. This paper will combine my own work and the research result of the other members in my group, to introduce the development of Acemap, our academic search system. These results not only contain the theory part, but also the program achievement.**

***Index Terms*—academic search engine, Acemap, knowledge map, topic development**

I. Introduction

WHAT makes a search engine become successful? Google has given us the answer: The speed, the concise interface, and a little luck of course. And what makes an academic search engine become successful? That's a little confusing, because it seems that no academic search engine nowadays can be called really successful or popular if we use a strict standard. We use Google scholar indeed, but do we really think it is a good enough product? Maybe it's just because we do not have an alternative.

Therefore, academic search engine has a large enough space to be improved, and the work based on needs both insight into recommendation algorithm and ability of web programming. So we have a team focusing on such great job, and I am honored to become one of them.

As for how we actually develop such a search engine, I probably have more words to say, because I take part in the engineering part in specific. This term our web team mainly focus on how can we present the search results and academic information more vividly. Visualization is a permanent topic for web programming, and how to visualize the content decides whether a web live or die, to a great extend. My job recently is modifying our homepage. And I am going to interpret how I actually change it and the reasons behind every little step.

II. Pre-knowledge for Acemap

Acemap is an academic search engine developed and maintained totally by the team of IIOT, whose search platform is based on solr, search algorithm is created originally by the team and web visualization is designed elaborately. Here I am going to show you the basic knowledge for the each part of our project.

A. Solr

Solr is the popular, blazing-fast, open source enterprise search platform built on Apache Lucene. Solr is highly reliable, scalable and fault tolerant, providing distributed indexing, replication and load-balanced querying, automated failover and recovery, centralized configuration and more. Solr powers the search and navigation features of many of the world's largest Internet sites. Solr's main features are generally used by all parts of our web, and here are some of the most thrilling features:

1. Adcanced Full-text Search capabilities. Powered by Lucene™, Solr enables powerful matching capabilities including phrases, wildcards, joins, grouping and much more across any data type.
2. Standards based open interfaces-XML, JSON and HTTP. Solr uses the tools you use to make application building a snap
3. Near Real-time Indexing. Want to see your updates now? Solr takes advantage of Lucene's Near Real-Time Indexing capabilities to make sure you see your content when you want to see it.
4. Comprehensive Administration Interfaces. Solr ships with a built-in, responsive administrative user interface to make it easy to control your Solr instances.
5. Highly Scalable and Fault Tolerance. Built on the battle-tested Apache Zookeeper, Solr makes it easy to scale up and down. Solr bakes in replication, distribution, rebalancing and fault tolerance out of the box.

However, some improvements also urgently needed for us to optimize the web search system. First and foremost, we still use the search and recommendation algorithm provided by Solr as default, and the search and recommendation results are still far from satisfying. After the research part finished by the theory group, we need to apply the algorithm to the web as soon as possible to improve our search results. And other problems let me discuss in the other parts of this paper.

B. LDA

In natural language processing, latent Dirichlet allocation

(LDA) is a generative statistical model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar. For example, if observations are words collected into documents, it posits that each document is a mixture of a small number of topics and that each word's creation is attributable to one of the document's topics. LDA is an example of a topic model and was first presented as a graphical model for topic discovery by David Blei, Andrew Ng, and Michael I. Jordan in 2003.[1] Essentially the same model was also proposed independently by J. K. Pritchard, M. Stephens, and P. Donnelly in the study of population genetics in 2000.[2] Both papers have been highly influential, with 13320 and 15857 citations respectively by January 2016.

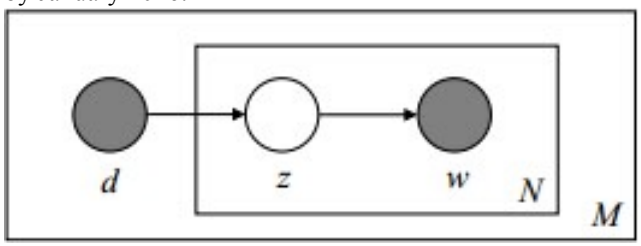

figure1. How to get a topic from the document.

As for the exact model of the LDA, here I have a figure to interpret.

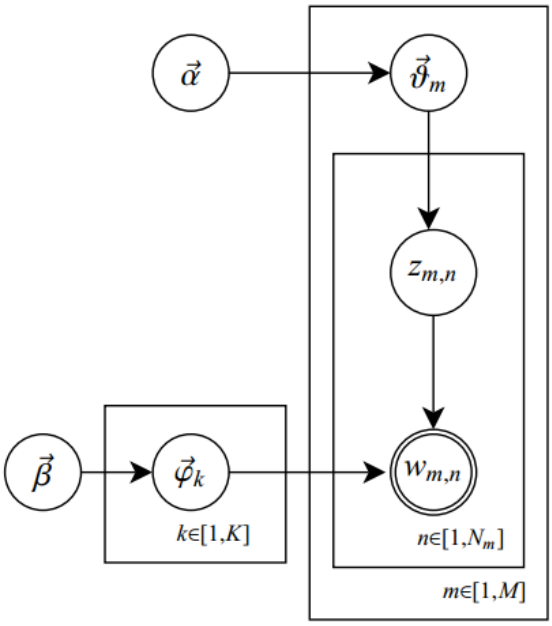

figure2. The model of LDA

Applying the model to our system, we can get topic distribution of a specific paper as follow:

figure3 the topic distribution of a specific paper

III. VISUALIZATION

This part I am going to talk about the visualization work of our academic search engine.

How to present the information more effectively and vividly is such a question that faced by any webs. Our team comes to realize that the academic search engine pay little attention to the visualization and make their webs hard to use except for searching for papers. We stress on the visualization just as we call our system Acemap, and we care much about how we show the results in a more effective way that the user can get more other than a list of papers. Here I am going to present some results of our project, which maybe contain both the work of mine, and the work of the other group members'.

A. The homepage

The homepage greatly decides the first impression the web leaves on the users. Should the homepage be as concise as possible just as what Google and Baidu is trying to do? Should it try to deliver more useful information that may attract the users? To be honest, that is really a problem that confuses our team a lot. We refer to Bing.com and get the idea finally. We can combine both the beautiful visualized effects and conciseness together, just as what Bing.com has done. Having a look into Bing.com, we find they provide the search blank with a beautiful background. What's more, every background picture has a story behind it. If the user is interested with the information about this picture, he/she can click the button at the bottom of the page and read the detailed introduction. So, As we call our project Acemap, why don't we also provide our search blank with a topic map or knowledge map? Our advantage is the map we generate contain far more profound and worthwhile information than the pictures on Bing, because each of these map stands for a changing statement of a specific academic field, from which the user can get direct impression of how the trend goes and the insight into this very field. To be more precise, every map in the homepage maybe the most useful content of the Acemap. And here I am going to show the result of my work.

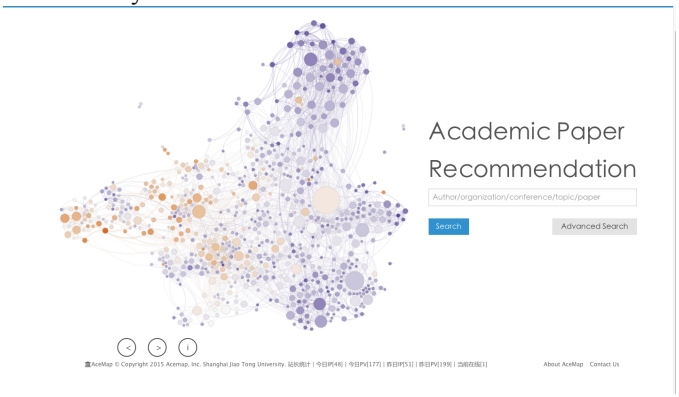

figure4. The Homepage of Acemap

The map behind is the author map of wireless network. From the map we can easily find the connection between authors, and the direct impression of how big the influence an author has. Of course, we can actually dig out more useful information behind such kind of maps, and what we aim to do

is leave this big space to our users, which means we generate the map, and let the user find what they want by themselves. Some basic information is provided in the slide blank of course. Just as figure5.

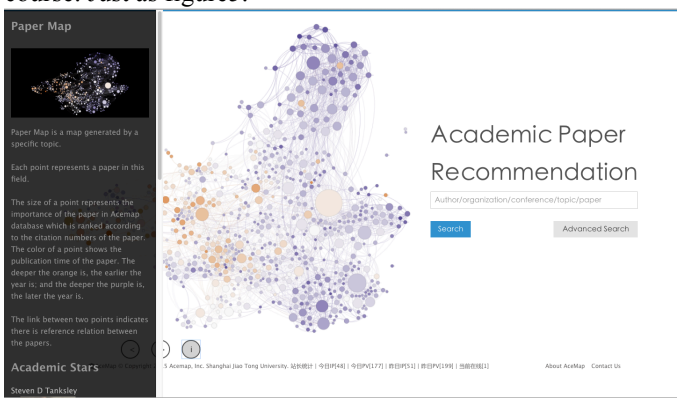

figure5. The basic information of topic map in the homepage

As we planed, the background figure will be changed automatically as long as we have generated enough maps. Now we can change the map by ourselves through clicking the button at the bottom of the page. And the result showed as follow.

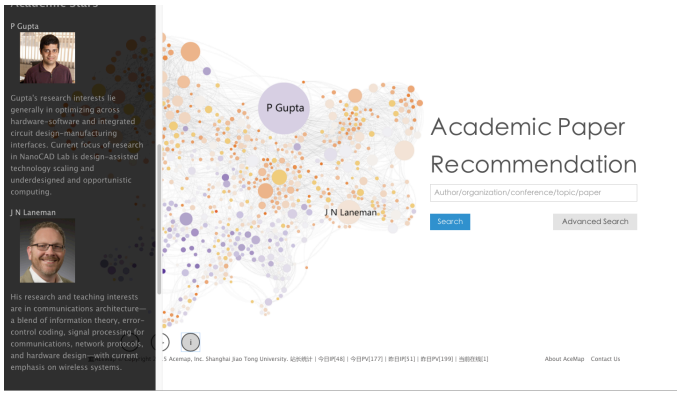

figure6. Change the background map

B. *Other results relating to visualization*

Here, I am going to present some results which are achieved by other group members.

First is the topic map. The topic map aims to reflect the trend of a specific topic, and the result we want to achieve is, through having a look on the map, the user can get insight into what's happening and what will happen in this filed. The trick points to generate such maps are how do we clean the big data we have already gathered and in which method do we use to cluster them. An application called Gephi is frequently used during the procedure that we generate the maps. And what we have to do is import the data we have already clean into excel or CSV file, then put it into Gephi for further processing. By using Gephi, we can generally transform a group of points, in which each point denotes a topic, into some regular graph. And in such graph we can find more useful information relating to the whole topic or even the field. One of the result is shown as follow.

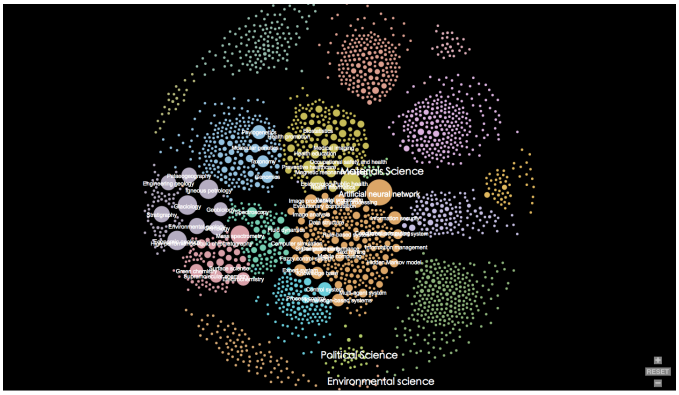

figure7. The topic map

Then what I want to talk is the paper map. We design a two layers map system, that you can have a look of the bigger picture at the topic map, and what's more you can dig out a specific topic and see what's going on in this topic through the second layer—the paper map. Paper map, which has the different function from topic map, aims to reflect what exactly compose a topic. To be more precise, we aim to display the papers in such topic in a more logical and vivid way. Now we can present the results in a circle and show the connection between each node with a line. Such outcome of course is far from thrilling and satisfying, which means we have a lot of work to do in the near future. An example of paper map shown as follow.

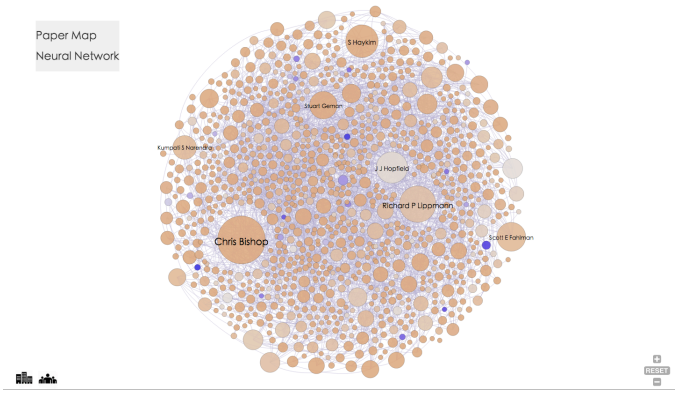

figure8. The paper map

IV. ACEMAP DEVELOPMENT BLOG

Straight to the point, I am going to build a development blog for our web. Such kind of idea does not come from nowhere. In the old version homepage of our web, we have a news block at the bottom of the page, whose job is to push some update news and new functions as well. Now the new homepage has already given up such news block for the sake of conciseness. However, the web does have the need to inform its users of some important changes as well as amazing new functions. Could we have another way to achieve the same goal? Just as some web also maintaining their official blog to present more information about the web and communicate with the users as well, we can also develop such blog to tell the stories about out team and every student

member behind such a big project. This idea has already consented by the group leader, and will be put into practice in the near future.

V. Future work

As for a member of engineering group, my work for the next stage will be still concentrated on the front and back end of our web. There are some goals I want to achieve in the near future with the help of our group members.

First, Put the algorithm designed by the theory group into practice. The algorithm functioning in our web now is still the default algorithm of Solr. And as long as we do not use our own search and recommendation algorithm, our system has no fundamental difference with the other ones. Therefore, I think that's our priority one task.

Second, we have not acquired the technic to generate all the same maps once for all. It's inevitable to produce more maps so that we can apply them to our web, and we cannot spend so much time on generating the maps one by one. Therefore, we have to acquire the technic that input all the nodes and output a map automatically.

Third, more visualization technics need to be found and applied to our system. It's a big difference between Acemap and other academic search engine that we focus more on how to present the information vividly. Therefore, we need to find more advanced visualization technics to achieve this goal and make the system more outstanding and special.

The foregoing opinions, just as I stated in the class, are the very work I plan to focus on in the near future. Of course my own ability is very limited when it comes to such a big project. Fortunately, our team has all the talented people we need to achieve such a great project, and I do believe that finally we can accomplish something amazing.

It's all because, just as I said at the end of presentation, we will never ever stop pursuing to be perfect.

References

[1] Thomas Hofmann. International Computer Science Institute, Berkeley. Probabilistic Latent Semantic Indexing. the 22nd International Conference on Research and Development in Information Retrieval (SIGIR'99):1999 04/2004,

[2] Blei, David M.; Ng, Andrew Y.; Jordan, Michael I. Latent Dirichlet allocation. Journal of Machine Learning Research 3 (2003) 993-1022