AceMap: A Novel Approach towards Displaying Relationship among Academic Literatures

Zhaowei Tan
Shanghai Jiao Tong University
dilevski_tan@sjtu.edu.cn

Changfeng Liu
Shanghai Jiao Tong University
stevenevets@sjtu.edu.cn

Yuning Mao
Shanghai Jiao Tong University
morningmoni@sjtu.edu.cn

Yunqi Guo
Shanghai Jiao Tong University
luckiday@sjtu.edu.cn

Jiaming Shen
Shanghai Jiao Tong University
sjm940622@sjtu.edu.cn

Xinbing Wang
Shanghai Jiao Tong University
xwang8@sjtu.edu.cn

ABSTRACT

A large number of papers are being published every year, which makes it difficult for researchers to grasp the relationship among the scientific literatures and the big picture of academic fields. The new challenges have thus been raised, such as analyzing the complicated citation and author network, mining valuable scientific knowledge, and visualizing big scholarly data. The existing academic systems, such as Google Scholar and DBLP have mainly adopted text-based methods, while some other systems make attempts to better navigate the literatures, for example, AMiner and Science Navigation Map. Although these systems show improvements, they fail to present the academic data in a holistic way, and also have limited functions. Therefore, we need to develop new tools which can realize more modules and further explore the academic literatures.

In this paper, we conceptualize and design a novel academic system, AceMap, to analyze the big scholarly data and present the results through a "map" approach. AceMap integrates several algorithms in the field of network analysis and data mining, and then displays the information in a clear and intuitive way, aiming to help the researchers facilitate their work. After describing the big picture, we present achieved results and our work in progress. By far, AceMap has implemented the following functions: dynamic citation network display, paper clustering, academic genealogy, author and conference homepage, etc. We have also designed and performed distributed network analysis algorithms in a cutting-edge Spark system and utilized modern visualization tools to present the results. Finally, we conclude our paper by proposing the future outlooks.

1. INTRODUCTION

Activities on scientific research play a strategic supporting role in improving the social productive forces as well as the comprehensive national strength. Countries around the world put great emphasis on the scientific research, investing money, people and other resources.

As researchers, we personally feel privileged to work in such a supportive academic environment. However, we sometimes also feel helpless – with limited ability and time, we have to face thousands of new papers in the field of study, as scientific research is so active nowadays. It is simply impossible that we read every publication and comprehend the relationship among them. Therefore, it's urgent to build systems capable of analyzing the attributes of the papers and the relationship among them exclusively for researchers. Using such systems, the scholars are able to clearly see the properties of one paper and know which one to pinpoint after seeing the big picture in a field.

Currently, several research entities and companies have already developed some systems to support the academic activities, such as Google Scholar [5] and dblp [11]. However, these systems primarily focus on textual contents of publications, namely the metadata of one specific paper, instead of displaying a global view of the whole academic area. In addition, they generally fail to provide the users with a straightforward way to comprehend the relationship among scientific literatures. In the meantime, some systems like AMiner [18], Metro Maps of Science [14] and Science Navigation Map [12] have been digging deep into the academic data and displaying them in a modern way. These systems propose methods to solve some problems like ranking or text mining. Meanwhile, some systems pay attention to knowledge visualization and want to build academic landscapes. Web of Science [8] provides the function of citation map, which can present citations and references of two generations at most without filtration. But it is still difficult for users to understand the internal relationship among papers through the depth-limited and raw links. Moreover, some other work tries to depict the overviews of academic domains. Map of Science [7] aims to portray the overall pictures of science and technology with fine-grained details and has provided several academic landscapes of different fields on their website. Eigenfactor.org [4] implements a field-level interactive academic map browser to reveal the relationship among various fields. However, the viewpoints of these systems are relatively inflexible and users can't explore deeper. Infobaleen [6] proposes a dynamic navigating method to explore a knowledge network in multiple levels and applies

WWW'16 Companion, April 11–15, 2016, Montréal, Québec, Canada.
ACM 978-1-4503-4144-8/16/04.
http://dx.doi.org/10.1145/2872518.2890514

it to presenting Wikipedia. In contrast, we focus more on multilayer interactive map of academic fields and network analysis and aim to be more comprehensive.

Inspired by geographic maps, we build this novel academic map, AceMap, to provide better services for researchers. In proposed system AceMap, we realize these functions among papers accordingly: (1) dynamically unfold and fold a paper's citations, checking what "the region around this publication" is; (2) cluster the papers in different levels, using algorithms to find communities or fields; (3) calculate paths between two papers and display the paths; (4) see the detailed information about a publication and its most relevant or idea-generating papers; (5) check the home pages for AACT (author, affiliation, conference and topic).

Our contributions in this paper are as follows:

- Design a novel approach to display the relationship among academic entities
- Build a prototype system that partly realize the functions and present the results

The rest of the paper is organized as follows. First, we start from giving a system overview in Section 2, which gives a brief description of AceMap. Then we introduce what the system looks like right now in Section 3. In Section 4, we provide the outlook of our system, after which some related work is shown in Section 5. Finally, we conclude this paper in Section 6.

2. ACEMAP DESCRIPTION

As stated in the Introduction, we borrow the ideas from a geographic map to build an academic map. We now list and elaborate the basic functions of our system. We also show their significance and the ways/philosophies to realize them.

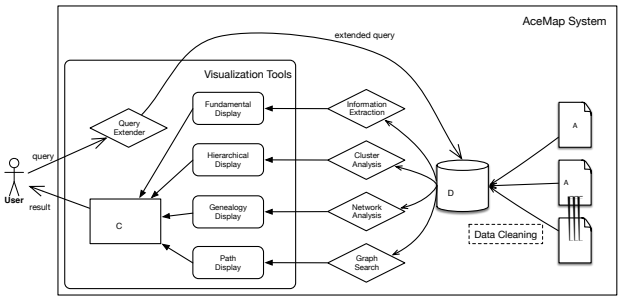

Figure 1: The Architecture of AceMap System.

2.1 Fundamental Display

Fundamental Display is just like the very first view a user can see when entering an online map. In this interactive graph, we display a network centered on the paper that the user just searched and chose. Each node represents a paper, and each link represents a reference with an arrow mark to indicate the direction. When the user clicks on one node, it will be expanded to show its references. Similarly, the user can collapse a node in the same way. In addition, the user can hover on one paper to see the specifics of it such as title, publication venue, and publication year.

Some papers might be too distant from the center one. Rather, there are too many papers between them on the shortest path. We thus offer the function of limiting the maximum distance between the papers displayed and the center one. Then the users can focus on the papers which are closer in the reference network. In addition, by zooming in and out, the users are able to view the network at various scopes. They can either focus on the relationship among a few papers or have a broad view of the local clusters. Moreover, we provide the users with the function of saving the network graph into SVG file, so the users can save it conveniently for further use.

In this view, the users can have an understanding of the topology of the reference network. With the help of the links between papers, the users may have a deeper understanding of the specific paper and the papers nearby. By clicking nodes, the users can see a dynamic network of the papers, during which he is able to explore the history and the process of the development of the papers. To sum up, the users can get an intuitive impression of the paper he is interested in and have a general idea of its "neighbors" and "ancestors", just like checking a place on a geographic map.

2.2 Hierarchical Display

If we use a geographic map and keep zooming out, we will be able to see which city/state/country a place is located at. Similarly, zooming out from Fundamental Display in AceMap will lead to a hierarchical display which indicates the location of the specified paper in the field of study.

Academic papers accumulate every year; the citation network among papers is becoming larger and more complicated. Here our focus is clustering these papers into several categories according to this network. Clustering this large network can be seen as a community detection problem.

We use the Label Propagation Algorithm (LPA) [13] to cluster the papers. LPA is an extremely fast algorithm because it has the advantage of nearly-linear running time. Another strength of LPA is that it needs small amount of a priori information about the network structure. Surely, there are some accuracy problems in LPA. However, since the graph to cluster has 3 million vertices and most vertices only have no more than 10 edges, the graph is a typical sparse graph and LPA's performance is relatively better in clustering sparse graph. Considering the size of input data and the resource we have, we finally choose LPA to cluster the graph. We then analyze the titles of papers in each cluster and choose the most common words in titles as keywords to represent this cluster.

The interactive display of clusters provides the users with a straightforward way to have an intuitive idea of what the area he is focusing on is, and what the whole field and the sub-fields are like. Moreover, he can realize the importance of one field and the relationship between fields by looking at the clustering result.

2.3 Path Display

When users are using the Fundamental Display, the back end programs can be executed to find the paths between two arbitrary papers in the AceMap, and return the answer to the visualization tool. Therefore, the paths from one paper to the specified one and related papers on the path will be highlighted in different colors, while the opacity of other nodes are increased so as to avoid distraction. The users can thus clearly see the relationship between the two papers.

Through learning the academic paths between two papers, we can learn the intuitive relationship between them, and thus draw an outline of the development in the related fields and find some important papers among the field. Imagine that in future if we can successfully classify papers into categories like "groundbreaking paper", "great follow paper", when we use the Path Display, we will be able to see a more clear relationship between publications, e.g. two papers are both great publications which follow a paper that generates a new research area, etc.

The paths between papers are easy to find when the data scale is small. We take two classic methods, BFS and DFS respectively, to find the academic paths. When the scale goes larger, some parallel variations [10, 19] can be conducted on distributed platform to accelerate the process.

2.4 Paper View

In the Paper View, we make an analogy to Google Maps StreetView – In StreetView, we can see the real scene of a place, as well as go for a direction and see the pictures nearby. Paper View aims at achieving similar view. When we search for a paper, we can enter the Paper View, where we see the details of a paper and several arrows pointing to different directions, such as citations, references, most relevant papers, other papers from this author, etc. By clicking one paper in the view, we enter the Paper View of that chosen paper.

The citations, references and relevant papers are easy to find and gather. Additionally, we also design algorithms to find the "genealogy" of a paper and put that into the Paper View, i.e. given a single publication and its references, we try to answer this question: which ones among these references are of the most importance for generating this paper? We try to traverse the whole citation network and generate a "family tree" of the paper, finding the ancestors with the original idea, and present that in the Paper View. In a given citation network, when a root paper is chosen, we calculate the relative importance scores of related papers. The relative importance score of a paper reflects the influence this paper has on the root paper. A simplified algorithm is presented in Algorithm 1. Starting from the root paper, we distribute each paper's score to those cited by it as an increment to update scores layer by layer. Since groundbreaking papers are cited by papers from different layers, they get higher scores. The citation paths from high score papers to low score ones display the development and the evolvement of a single paper, and the generation of various paths in a field provides a panorama of the whole field. We omit the distributed version of the algorithm, which is easy to derive.

Algorithm 1 Academic genealogy scoring algorithm

Require: $G(V, E)$, v_0
1: $Q \leftarrow \emptyset, S \leftarrow \emptyset$
2: $v_0.score = 1$
3: add $v_0 \rightarrow Q$, add $v_0 \rightarrow S$
4: **while** $Q \neq \emptyset$ and $K < K_0$ **do**
5: $v = Q.DeQueue()$
6: update K as the length of current path from u to v
7: $n = RefNum(v)$
8: **for** $\forall v'$ satisfies $e(v', v) \in E$ **do**
9: $v'.score = v'.score + v.score/n$
10: **if** $v' \notin S$ **then**
11: add $v' \rightarrow Q$, add $v' \rightarrow S$
12: **end if**
13: **end for**
14: **end while**

Using Paper View, the users can start from the paper of interest and find the relevant publications/ similar papers/ development skeleton of this paper. Paper View is also scalable – simply adding new "directions" we can present another set of papers that are related to a paper in some context.

2.5 AACT Home Pages

Our AACT (Author, Affiliation, Conference and Topic) Home Pages is a platform to succinctly display our analysis on AACT entities. For example, on the home page for each author, we present the basic information of the scholar, including the related papers, organizations he/she works for and the related fields that the professor is interested in. Moreover, we show the related author list on this page. Furthermore, if we click 'see more' button, the site will be directed to a more complex relationship network, which displays a personal network with the original author as the focal node.

3. PRELIMINARY RESULTS

To fulfill the above visions, we build a team of around 80 students to implement the system. Currently, we have preliminarily realized most of the functions, and the prototype system will have a huge leap after this winter holiday. In this Section, we briefly introduce the status of the system and demonstrate some results with the url links where you can have a more direct impression.

3.1 System Architecture

The overall architecture of AceMap system is shown in Figure 1.

When a user comes, he interacts with our system using a website. The server handles the input query using the unified database and sends the result back to the visualization tools, which are responsible for displaying the relationship among the academic literatures in a clear and intuitive way.

Different from the existing text-based methods, we implement a new method to display the topology of papers using D3.js [3], a JavaScript library for web-based data visualization. We use the JSON files to store the data and load it dynamically into the front end. We also leverage Apache Spark [2] with 6 executors each with a memory of 30GB to execute novel algorithms in the back end. By this approach,

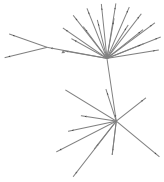

Figure 2: The graph shows the "ancestor" of the center paper. By clicking the nodes, networks of the references are expanded and collapsed. You can find a simple demo at http://acemap.sjtu.edu.cn/map.

we present the relationship among academic literatures in a vivid and interactive way.

3.2 Data Collection

It is very kind of Microsoft Research to release their Microsoft Academic Graph (MAG) [16] dataset. We use this open source dataset to construct the AACT pages, while realizing the other functions using a crawled IEEE dataset of around 3.5 million papers. We use this smaller dataset in order to save computational resources and will later perform all the functions based on MAG.

3.3 Demos

Currently, we have already built a search engine as shown in Figure 3. The website takes the advantage of Solr [1] framework and can return the results after entering a query. We use this search engine as an entrance to our map.

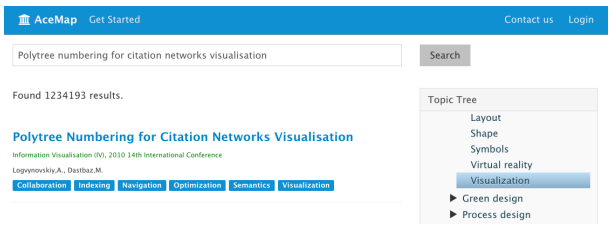

Figure 3: A Snapshot of the Website. The system is hosted at http://acemap.sjtu.edu.cn.

A dynamic process of Fundamental Display is illustrated in Figure 2. This view is implemented based on a layout of D3.js named force. We add many functions to make the interface more user-friendly. Some major techniques we use are described as follows.

We load the whole data in the JSON file into some variables, say nodes and links, but do not display them at once. By using such variables we can calculate whether some nodes are connected or not. Whenever the users click on one node, the corresponding nodes will be added to the displayed network. We also judge whether the references of that clicked paper is already in the network or not, and handle the two situations separately.

We use asynchronous techniques since loading too much data at once is not realistic for the browser to handle. When the users have navigated to a node on the margin of the graph, which means no reference information of it is stored in the currently-loaded JSON file, the system can fetch the data automatically from database and merge it into the existing network.

The implementation of the dynamic cluster graph is based on another basic layouts of D3.js named pack. We use predefined format in the JSON file so as to load the data into the browser efficiently. By using the cluster algorithm LPA, we store this relationship between a cluster and its sub-clusters as parent and children in advance. We also pre-calculate the size of each cluster to save the running time. In addition, we add a callback function to navigate to the next detailed view, which will be introduced when the system detects that the users have explored to the deepest level.

We implement a parallel version of LPA method. For our experiments, the system clusters papers approximately into 100 communities within linear time. The result is shown in Figure 4. The path finding can be executed within seconds when using small datasets under Spark settings. The result can be seen in Figure 5.

The Paper View module is under construction, while the AACT Home Page can be viewed. Figure 6 is an example of an Author Page, where users can check the author's publications, related authors and research interests.

4. OUTLOOKS

After we build this prototype system, the tasks right now can be roughly divided into three categories. The first one is to fully realize our current ideas, the second one is to get feedbacks and prove the effectiveness of the system, and the last one is to conduct new functions.

In addition to finishing the basic version of AceMap, many ideas are worth a try to upgrade the map. For example, we can classify different papers into several categories, and mark these classes with different colors or shapes in our map. When we look at a map, we can recognize which place is a bridge and which place is a tourist attraction. Different places hold different functions and statues inside a city. Similarly, papers play different roles. We are eager to distinguish whether a paper is a groundbreaking pioneer, a great follower, or just a useless one that can be simply ignored. This conception is depicted in the Figure 7. Upon successful

Figure 4: The cluster graph shows fields in computer science. `http://acemap.sjtu.edu.cn/map/show_circle`**.**

Figure 5: Finding multiple paths to the center paper.

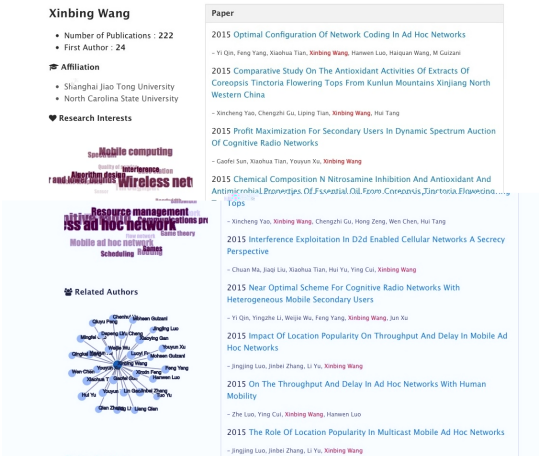

Figure 6: The Home Page for Prof. Xinbing Wang. `http://acemap.sjtu.edu.cn/authorpage?AuthorID=41327514`**.**

classification, all of the Fundamental Display, Hierarchical Display and Path Display will make more sense to the users.

However, we cannot claim our system "useful" before the users agree so. Therefore, when the beta version is released, we will design surveys, ask researchers to use the system and let them provide insightful feedback. Using such information, we can correctly evaluate our system and modify it towards a right direction.

Besides visualization, our group's efforts have reached other promising areas. For example, academic recommendation system, academic topic modeling, AACT analysis, etc. We want to build a system that can help researchers from different aspects.

5. RELATED WORK

There are existing academic searching systems such as Google Scholar [5], Web of Science [8] and dblp [11]. Google Scholar defines itself as the way to find scholarly articles across disciplines. Web of science aims to be today's premier research platform for information in the sciences, social sciences, arts, and humanities. The dblp computer science bibliography is the online reference for open bibliographic information on computer science journals and proceedings. However, all these systems are text-based and focus more on the information of a specific paper. Though displaying some static figures as auxiliaries, they fail to provide the users with a straight-forward way to comprehend the relationship among academic literatures and to have a global view of the whole academic field.

There are also systems like VEGAS [15], AMiner [18], Metro Maps of Science [14] and Science Navigation Map [12], which are all important efforts towards better academic systems. VEGAS concentrates on summarizing the large citation graph according to the user's interest and finding the impact of a highly influential paper. AMiner focuses on the evaluation of the influence of researchers by analyzing social network. Metro Maps of Science tries to excavate the story line using text mining, while SNM is a tool for interactive data mining.

There is another kind of systems which concentrate on portraying the academic maps. Woon et al. [21] leverage bibliometrics to visualize the science and technology landscape. Vilhena et al. [20] visualize the cultural holes in the form of a topographical map. Skupin et al. [17] use over two million publications to depict a detailed landscape of the medical knowledge field. Boyack and Klavans [9] use cocitation-based techniques to implement an article-level model and

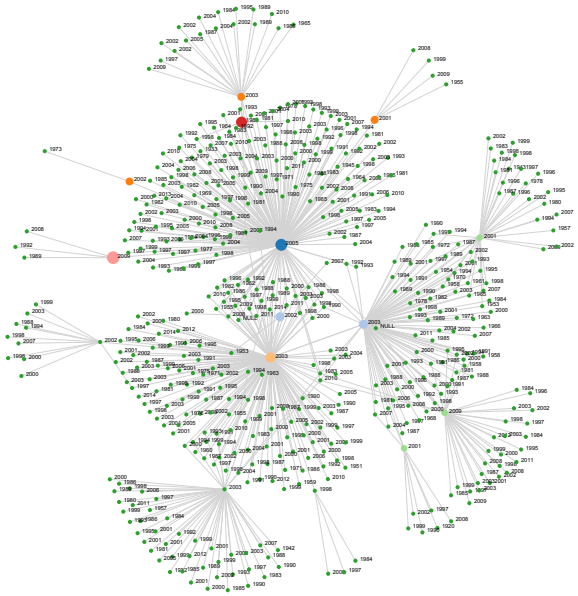

Figure 7: An example of the graph. Each node has a different size, indicating its importance. The nodes are shown in different colors to indicate their status in the network. `http://acemap.sjtu.edu.cn/Academic_map/main.html`.

map of science. Eigenfactor.org [4] builds a field-level interactive academic landscape to present the internal interrelation. In contrast, we pay more attention to the the dynamic navigation over different levels of the academic map and network analysis.

6. CONCLUSION

In this paper, we present a novel academic system, AceMap, which aims to process the big scholarly data, analyze the citation network and visualize the relationship among papers to help researchers grasp the academic big picture more conveniently and more intuitively. First we give the full conceptualization of AceMap. We discuss each part of AceMap and its (possible) implementation and corresponding significance. Next, we present our prototype system and show several screenshots of the current AceMap. Last but not least, we describe a clear and attainable blueprint of our future system. You are more than welcomed to check AceMap at anytime, as we sincerely want to conduct a project that could help the scholars and thus will keep updating this promising project.

7. ACKNOWLEDGMENTS

This work was partially supported by NSF China (No. 61532012, 61325012, 61271219, and 61428205).

8. REFERENCES

[1] Apache solr. `http://lucene.apache.org/solr/`. The Apache Software Foundation.

[2] Apache spark. `http://spark.apache.org/`. The Apache Software Foundation.

[3] D3.js. `http://d3js.org/`. Mike Bostock.

[4] Eigenfactor.org. `http://eigenfactor.org/`. Eigenfactor.org.

[5] Google scholar. `https://scholar.google.com/`. Google.

[6] Infobaleen. `http://www.infobaleen.com`. Infobaleen.

[7] Map of science. `http://www.mapofscience.com/`. SciTech Strategies.

[8] Web of science. `webofknowledge.com`. Thomson Reuters.

[9] K. W. Boyack and R. Klavans. Creation of a highly detailed, dynamic, global model and map of science. *Journal of the Association for Information Science and Technology*, 65(4):670–685, 2014.

[10] J. Freeman. Parallel algorithms for depth-first search. 1991.

[11] M. Ley. The dblp computer science bibliography: Evolution, research issues, perspectives. In *String Processing and Information Retrieval*, pages 1–10. Springer, 2002.

[12] Y. Liu, Z. Huang, Y. Yan, and Y. Chen. Science navigation map: an interactive data mining tool for literature analysis. In *Proceedings of the 24th International Conference on World Wide Web Companion*, pages 591–596. International World Wide Web Conferences Steering Committee, 2015.

[13] U. N. Raghavan, R. Albert, and S. Kumara. Near linear time algorithm to detect community structures in large-scale networks. *Physical Review E*, 76(3):036106, 2007.

[14] D. Shahaf, C. Guestrin, and E. Horvitz. Metro maps of science. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1122–1130. ACM, 2012.

[15] L. Shi, H. Tong, J. Tang, and C. Lin. Vegas: Visual influence graph summarization on citation networks.

[16] A. Sinha, Z. Shen, Y. Song, H. Ma, D. Eide, B.-j. P. Hsu, and K. Wang. An overview of microsoft academic service (mas) and applications. In *Proceedings of the 24th International Conference on World Wide Web Companion*, pages 243–246. International World Wide Web Conferences Steering Committee, 2015.

[17] A. Skupin, J. R. Biberstine, and K. Börner. Visualizing the topical structure of the medical sciences: a self-organizing map approach. *PloS one*, 8(3):e58779, 2013.

[18] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su. Arnetminer: extraction and mining of academic social networks. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 990–998. ACM, 2008.

[19] J. L. Träff. A note on (parallel) depth-and breadth-first search by arc elimination. *arXiv preprint arXiv:1305.1222*, 2013.

[20] D. A. Vilhena, J. G. Foster, M. Rosvall, J. D. West, J. Evans, C. T. Bergstrom, J. Sørensen, and D. Baldassarri. Finding cultural holes: how structure and culture diverge in networks of scholarly communication. *Sociological Science*, 1:221–238, 2014.

[21] W. L. Woon and S. Madnick. Semantic distances for technology landscape visualization. *Journal of Intelligent Information Systems*, 39(1):29–58, 2012.