Visualization of the Relations Between A Certain Paper and Its References and Citations in Acemap

***Abstract*—Nowadays, the number of academic papers grows at an incredible speed. Researchers have to face thousands of papers when trying to get involved in a new field of study. Traditional search engines like Google Scholar etc. provide query-based search service but the results may be fuzzy, which means the results are relevant but not meaningful to the query. Existing map-based method like Acemap is able to mine structure information and provide macroscopic information about a study field or topic, failing to study the in-depth information on a certain paper. To address this problem, I design a new paper map in Acemap to visualize the relations between a certain paper and its references and its citations. Also, a KDP (Keyword based Double-Damping PageRank) algorithm is introduced to identify papers with underlying principles. This algorithm will rank all the papers according to their PageRank value. The newly developed map contains much structure information and can help researchers acquire knowledge more efficiently.**

I. Introduction

Academic papers usually contain the most up-to-date technologies and extremely rich information about the knowledge structure in a specific field. Nowadays, as the number of papers grows in an incredible speed, researchers have to face thousands even millions of papers when they try to get involved in a new field of study. Existing scholar search engines like Google Scholar[1], Microsoft Academic[2] and IEEE Xplore[3] provide useful and convenient search tools to help researchers based on their query and return a long list of relevant papers. However, those results are often chaotic and not meaningful in content.

When using search engines like Google Scholar, etc., one will get papers that are relevant to the query. Those papers are listed in a descent order in terms of their "rank". Such "rank" is obtained through a certain kind of algorithm, which can represent the strength of correlation with the topic. But what we really want is papers that are "meaningful", which is different from "relevant". More specificly, if researchers want to understand knowledge graph and they search it on Google Scholar, the top three results returned from the search engine is about TransH [1], TransR [2] (two state-of-the-art methods for knowledge representation) and Trinity Graph Engine [3] (a powerful graph system developed by Microsoft Research Asia). Of course they are all about the topic we discuss—knowledge graph, but what the researchers really need is the concept, function or mathematical representation of knowledge graph, which we call "underlying principles". Researchers have to read numerous papers to get the meaningful ones. To issue this problem, several map-based systems like Acamap [4], Paperscape[4], Metro Maps of Science [5] and Aminer [6] etc. have designed several kinds of paper maps with regard to authors, topics or affiliations to reveal the structure information of papers. Those maps have very large scale and provide macroscopic information about every study field, while lacking the ability to provide specific information about a certain paper.

It can be observed that some underlying principles about one paper can be obtained from its references. But one has to read through all the references to get papers that really meaningful, considering that one paper may have lots of references. We call references that contain underlying principles of the original paper as "guiding papers", and "guiding intensity" refers to the guiding significance a cited paper to its citing paper. The larger guiding intensity is, the more meaningful a reference paper is.

We can also observe that citations of papers can help researchers understand the development trend of the topic, especially when they want to get involved into a new study field. Citations usually contain new ideas based on the original one from the cited paper, which may be a hint for new study. But just like references, not all citations share the same underlying principle with the cited paper. For researchers who is not familiar with the topic, it will be difficult to address those meaningful citations.

To address those challenges, I will firstly introduce a new paper map which can visualize the direct relations between a certain paper and its references and citations in Acemap. This new map will show the structure information of the reference/citation network in an intuitive way. Then I will implement a KDP (Keyword based Double-Damping PageRank) algorithm to rank all the references and citations, which can exactly address those papers with underlying principles. In this way, researchers can easily identify the guiding papers and understand where the idea of the paper comes from and where it can be led to in the future. Maps generated from the Acemap data base have shown significant results in reference/citation network structure display and addressing papers with underlying principles of the original paper.

The main contributions of my project can be summarized as follows:

1. A new paper map is developed to visualize the relations between a certain paper and its references and citations in

[1]https://scholar.google.com/

[2]https://academic.microsoft.com/

[3]https://ieeexplore.ieee.org/

[4]http://www.paperscape.org/

Acemap.

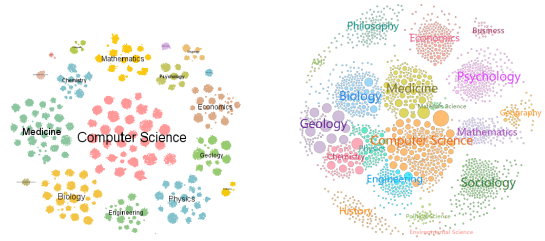
2. I introduce a KDP algorithm to the new paper map in order to identify the guiding papers from all references of the paper. Meaningful citations which share the same underlying principles can also be addressed in the map.

II. RELATED WORKS

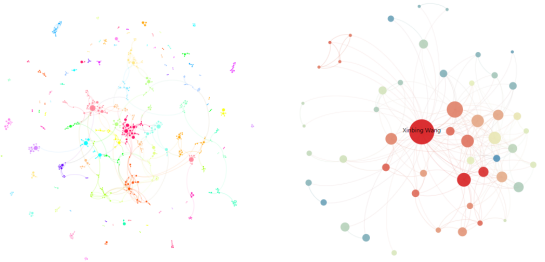
There are several classical methods to visualize the references/citations network, such as Histography, co-citation network, co-citation-author network and so on. Histography [7], brought up by E. Garfield in 1964, use sequential network of references to study the origin and the development of the field. Paperscape is inspired by this idea and draw references/citations network in terms of time sequence. Co-citation network [8] tends to study the relations between two papers based on the co-citation analysis. Co-citation relation exists between two different papers when they are co-cited by one paper. With more papers co-citing those two papers, the relation between them gets more close. Similarly, co-citation-author network [9] studies the relation between two authors based on the co-citation relation of their publications.

Efforts have been made to eliminate the limitations of existing academic search engines. AMiner [6] focuses on the evaluation of the influence of researchers by analyzing social network. Metro Maps of Science [5] tries to excavate the story line using Coherence, Coverage and Connectivity concepts. Acemap [4] uses several kinds of academic maps to describe the structure of knowledge. Academic Map (Figure 1(a)) is a large scaled map which contains millions of papers selected from hundreds of study fields. Papers that have significant influence in their field are clustered together, revealing the knowledge structure and study trends in this field. Other maps like Topic Map (Figure 1(b)), Affiliation Map (Figure 1(c)) and Co-author Map (Figure 1(d)) also contains large scale of information of study fields. Systems mentioned above are all coarse-grained to guide researchers to do in-depth study due to the lack of analysis on a single paper. HisCite[5] is inspired by the idea of Histography. It utilizes LCS (Local Citation Score), GCS (Global Citation Score), LCR (Local Cited References), and CR (Cited References), to address guiding papers for a paper with regard to cited times. HistCit recommends papers with high citation numbers, but such papers are not always the real guiding papers.

Much work has been done to address the guiding papers issue we've mentioned using both citation analysis and topic models. SimRank [10] is designed to explore the structural similarity between papers based on PageRank [11], but have little concern with the semantic information. [12] tries to recommend scientific articles by incorporating textual content into the traditional matrix factorization. Both methods ignored the overall structural information in the knowledge network. [13] introduces a novel model named RIDP (Reference Injection based Double-Damping PageRank) to mine the guiding papers out of massive academic papers. Experiment results of this model outperforms traditional Query Oriented method. The idea of this model mainly comes from PageRank [11], a web pages ranking algorithm used in Google search engine. However, the computational complexity of this model is relatively high and it requires text information of papers in the ranking process, which is difficult to obtain from existing data base. Considering that the paper map described in this project will finally be deployed in Acemap, low computation complexity is needed and information used in the process of generating maps must be obtained only from the data base. So a more appropriate method—KDP is designed in this project to address the guiding papers issue as well as fit the map algorithm into the existing Acemap system. Also, RIDP only studies the references of a paper, ignoring the citations which also play an important role in the structure of knowledge. The algorithm proposed in my project is able to rank all references and citations together, providing more information for researchers when they need inspiring ideas to keep going.



(a) Acacemic Map (b) Topic Map



(c) Affiliation Map (d) Co-author Map

Fig. 1. Paper Maps in Acemap

III. PAPER MAP VISUALIZATION

In Acemap, maps studying the relations between a certain paper and its references and citations have not been developed yet. Histograpy [7] is able to visualize the references or citations in terms of time sequence. Paperscape draws references/citations network based on this idea, but failed to bring all references and citations into one map. The map algorithm used in this project draw references and citations into one single map, and all reference/citation relations are illustrated on the map. Figure 2 shows a paper map generated by using the scheme mentioned in this project, where the seed paper is [15]. In this map, every paper is represented by a circle. The location of the circle represents the publish year of the paper. The positive direction on the x-axis represents the direction where time floats. The radius of the circle is

[5]https://en.wikipedia.org/wiki/Histcite

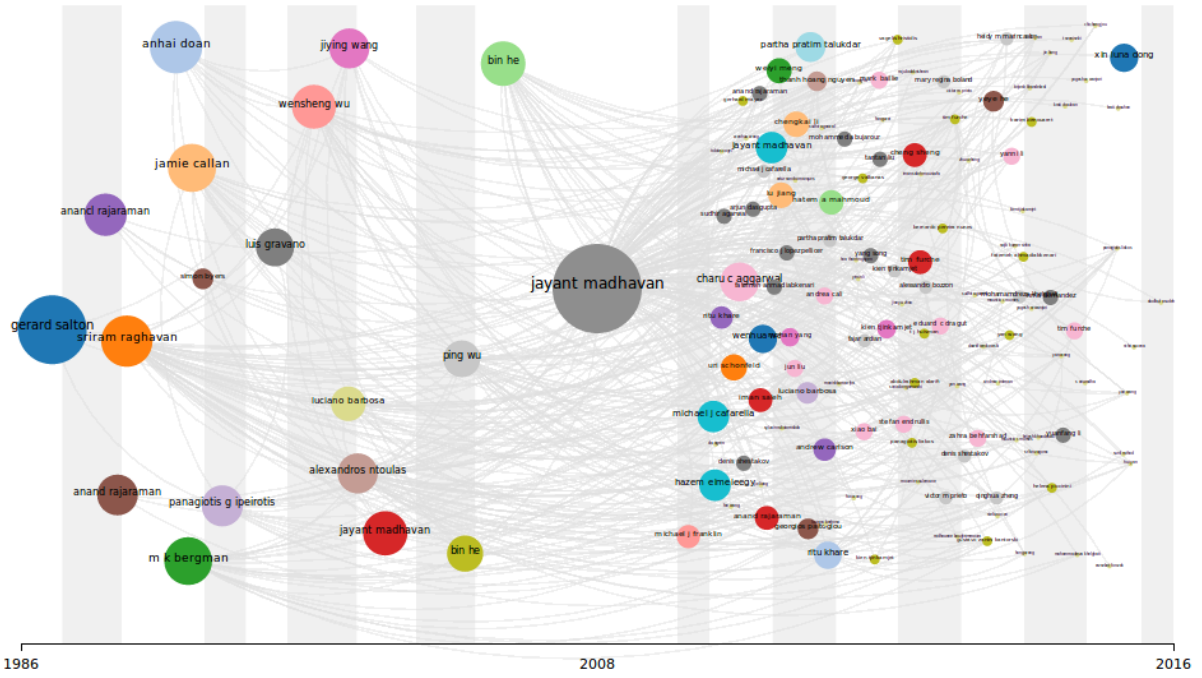


Fig. 2. Paper Map Generated from the Layout Algorithm

obtained according to the global citation number of the paper; the more citations, the bigger the radius. Curves between two circles reveal that there is a reference/citation relation between those two papers. The color of the circles are set completely randomly. The central circle which have connections with all circles in this map is the original paper we are interested in. $D3.js$[6] is a very useful javascript library used for data visualization. The tools and layout algorithms used in this project is mainly from $D3.js$.

Algorithm 1: MAP-LAYOUT: Layout Algorithm for the Map

Input: references set α; citations set β;
Output: paper map;
Initialize: $\Theta_{\mathcal{W}} \leftarrow \alpha + \beta$;
while $\Theta \neq \emptyset$ **do**
  locate the circle according to its publishment year;
  generalize the most appropriate location (x, y) based on d3.quadtree(); link it with its references circles;
Return: a paper map;

As we can see, the relations between references and citations are very complicated. The information this map provides is very large in scale, including not only the reference/citation relations from the original paper, but also the complex relations between references and citations. The original paper, its references and citations altogether forms a cluster, from which we can get some interesting results and conclusions.

[6]https://d3js.org/

1. In this map, circles on the left is usually bigger than the ones on the right. This means that references of one paper usually have more citations than the cited ones. Papers published earlier tends to have more citations.
2. One paper is more likely to be cited in one year or two after its publishment.
3. Papers with large global citation numbers don't always have large local citation number. This means that references with high GCS may not contain the underlying principles.

IV. KEYWORD BASED DOUBLE-DAMPING PAGERANK

A. Definition

Definition 1 (GUIDING PAPER). *is a cited paper which is helpful to understand the underlying principles of its citing papers.*

Definition 2 (SCDAG). *is a Single-Source Citation Directed Acyclic Graph which has only one node with* 0 *in-degree, i.e., seed paper. Figure 3 is a SCDAG diagram. Each paper in Figure 3 belongs to a level which indicates its shortest path from the seed paper that is the only paper on level 0 (L0).*

Definition 3 (MCDAG). *is a Multiple-Source Citation Directed Acyclic Graph which has many nodes with* 0 *in-degree, i.e., papers that have no citations in the references/citations network.*

Definition 4 (WEIGHTED MCDAG). *is a MCDAG in which each edge is weighted in accordance with guiding intensity.*

Definition 5 (RANKED MCDAG). *is a MCDAG where nodes have a "rank" according to their PageRank (PR) values, and it contains guiding papers.*

B. Major Steps of KDP

1. For a given paper, we get its references and citations from data base, and form a MCDAG according to the reference/citation relations. The direction of each link is from the citing paper to the cited paper.
2. Run keyword analysis on this MCDAG and get weights for each edge, then we get a weighted MCDAG.
3. Run Double-Damping PageRank algorithm on the weighted MCDAG, then we get a ranked MCDAG according to the PR value.
4. Papers have higher rank share similar underlying principles with the original paper.

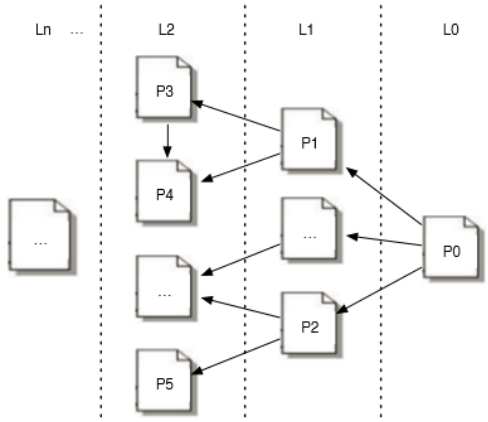


Fig. 3. N-level SCDAG Diagram

C. Double-Damping PageRank

In this part, I am going to show how to get PageRank value for each paper based on Double-Damping PageRank. PageRank algorithm is designed by Google, aiming to rank relevant web pages and return papers in a descent order with regard to their page rank. PageRank algorithm assumes that user will randomly browse several web pages which may contain url links to each other. At any time, user may jump to another web page or enter some page which doesn't have link to the current page. As time goes by, the probability of user browsing each page will converge to a certain value, which is exactly the PageRank value we are talking about. However, PageRank algorithm assumes that each link on the page will be clicked equally probably, which is not possible because some pages may be more popular and more likely to be clicked. WPR (Weighted PageRank) [14] solves this problem by giving each link a "weight". Links with higher weight get access more frequently than those with lower wight. Equation (1) and Equation (2) denote the PageRank and WPR algorithm respectively.

$$PR(p_u) = \frac{1-d}{N} + d \times \sum_{p_v \in I(p_u)} \frac{PR(p_v)}{L(p_v)} \tag{1}$$

$$PR(p_u) = \frac{1-d}{N} + d \times \sum_{p_v \in I(p_u)} PR(p_v) \cdot W \tag{2}$$

$PR(p_u)$ is the PR value of page p_u, $I(p_u)$ is the set of pages that link to p_u, $L(p_v)$ is the out-degree of page p_v, $d(0 \leq d \leq 1)$ is the damping factor (usually set to 0.85), N is the total number of all pages, and W is the weight of $link(p_u, p_v)$.

When using WPR to rank all the papers, some problems still exist. When a researcher finished reading a reference paper, he may go back to the paper he has read to re-understand some important concepts, which is quite rare in surfing web pages, by contrast. That often happens because of the difficulty of understanding an academic paper. So a Double-Damping PageRank [13] is proposed to solve this problem. In Double-Damping PageRank, each one-direction link is added with a link in the reverse direction. The added link, of course, has a corresponding weight. The basic algorithm is described as follows:

$$\begin{aligned} PR(p_u) = \frac{\alpha}{N} + \beta \times \sum_{p_v \in I(p_u)} PR(p_v) \cdot W_1 \\ + \gamma \times \sum_{p_v \in O(p_u)} PR(p_v) \cdot W_2 \end{aligned} \tag{3}$$

$$\alpha + \beta + \gamma = 1 \tag{4}$$

Where α is similar to $1-d$ in Equation (1). β is forward damping factor and similar to d in Equation (1), γ is the introduced backward damping factor. $O(p_u)$ is the set of pages that page p_u links to. W_1 and W_2 are weights calculated by using Keyword Analysis.

In order to keep consistent with classical PageRank, we set $\alpha = 0.15$. Experiment results have shown that when $\beta = 0.5, \gamma = 0.35$, both time performance and accuracy can be achieved. In next part, I am going to show how to get weights of the links by using Keyword analysis.

D. Keyword analysis

As mentioned above, the weights on the links represent how important the pointed papers are. We assume that papers with similar topic or sharing same underlying principles are important to each other. The best way to analysis the similarity between papers is to compare the content of both papers. However, this method is time-consuming and cannot be implemented in existing Acemap system. However, the keyword information of each paper in data base can be used to analysis the similarity. The weight on a certain link can be denoted as:

$$W_{ij} = \frac{same(i,j)}{\sum_k same(k,j)} \tag{5}$$

Where $same(i,j)$ is the number of keywords that i and j have in common. Obviously, a link will get a higher weight if the linked papers share more keywords. The overall KDP algorithm is shown as follows.

V. Results

Once the KDP algorithm is implemented on the existing map, we can get a ranked paper map. Figure 4 is the paper map generated from Figure 2 after KDP process. The ranking information is represent by the color of each circle. If the color

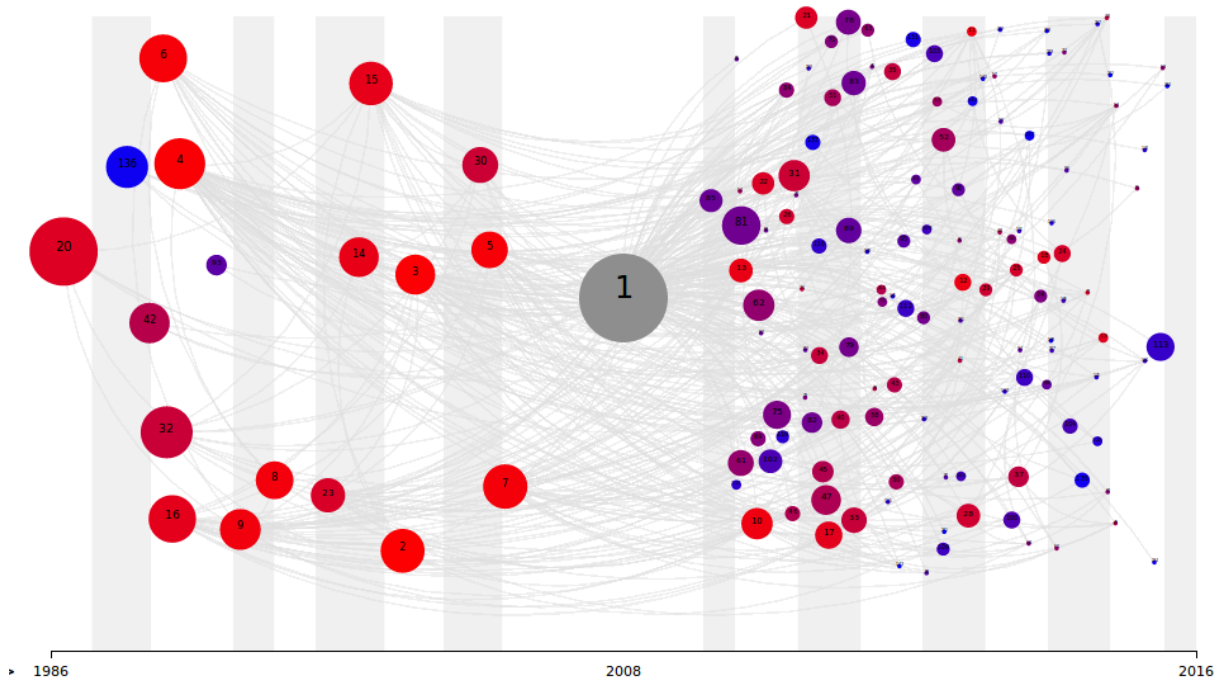


Fig. 4. Paper Map with PageRank

TABLE I
TABLE OF PAPERS WITH TOP5 PAGERANK.

Rank	Title
1	Google's Deep Web Crawl [15]
2	Corpus Based Schema Matching [16]
3	Downloading Textual Hidden Web Content Through Keyword Queries [17]
4	Crawling the Hidden Web [18]
5	Query Selection Techniques for Efficient Crawling of Structured Web Sources [19]

Algorithm 2: KDP: Keyword based Double-Damping PageRank

Input: keyword set α; reference set β **Output:** PageRank Θ;
Initialize: $\Theta[i] = 1/n$;
while $\Theta - \Theta' \geq \epsilon$ **do**
  update all the PageRank value according to Equation(3);
Return: Θ;

is more close to red, it obtains a higher rank; if the color is more close to blue, it obtains a lower rank. Papers with Top 5 PageRank in this map have been listed in Table 1. As we can see, all Top 5 papers (except the seed paper) have very close relation with the original paper.

VI. CONCLUSIONS AND FUTURE WORK

In this project, I design a new paper map to visualize the relations between a certain paper and its references and citations. To address the guiding papers or papers sharing similar underlying principles with the original paper, I introduce a KDP algorithm to rank all the references and citations. This project successfully draw all the references and citations along with all the relations into one map, which will help researchers to have a better vision about the structure of the reference/citation network. With the help of KDP ranking, one can easily address the papers with underlying principles. In reference papers, researchers can directly obtain the most important papers according to the rank; in citation papers, researchers can easily find the papers whose ideas are mainly inspired by the original paper, which will provide some useful hints for future study in this topic. In future work, I will focus on the relation between two papers and bring up some layout algorithm to visualize the relation. Possible schemes include LCR analysis or LCS analysis over all papers in the map.

REFERENCES

[1] Z. Wang, J. Zhang, J. Feng, and Z. Chen, "Knowledge Graph Embedding by Translating on Hyperplanes," in *AAAI*, 2014, pp. 1112-1119.
[2] Y. Lin, Z. Liu, M. Sun, Y. Liu, and X. Zhu, "Learning Entity and Relation Embeddings for Knowledge Graph Completion," in *AAAI*, 2015, pp. 2181-2187.
[3] B. Shao, H. Wang, and Y. Li, "The trinity graph engine," in *Microsoft Research*, 2012, 54.

[4] Z. Tan, C. Liu, Y. Mao, Y. Guo, J. Shen, and X. Wang, "AceMap: A Novel Approach towards Displaying Relationship among Academic Literatures," in *Proceedings of the 25th International Conference Companion on World Wide Web*, 2016, pp. 437-442.

[5] 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*, 2012, pp. 1122-1130.

[6] 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*, 2008, pp. 990-998.

[7] E. Garfield, and R. K. Merton, "Citation indexing: Its theory and application in science, technology, and humanities," 1979.

[8] H. Small, "Co-citation in the scientific literature: a new measure of the realtionship between two documents," in *Journal of the American Society of Information Science*, 1973, 24: pp. 265-269.

[9] C. Chen, "Visualizing semantic spaces and author co-citation networks in digital libraries," in *Information Processing and Management*, 1999, 35(2): pp. 401-420.

[10] G. Jeh, and J. Widom, "SimRank: a measure of structural-context similarity," in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2002, pp. 538-543.

[11] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank citation ranking: bringing order to the web," 1999.

[12] C. Wang, and D. M. Blei, "Collaborative topic modeling for recommending scientific articles," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2011, pp. 448C456.

[13] S. Tao, X. Wang, W. Huang, W. Chen, T. Wang, and K. Lei, "From Citation Network to Study Map: A Novel Model to Reorganize Academic Literatures," in *Proceedings of the 26th International Conference on World Wide Web Companion*, 2017, pp. 1225-1232.

[14] W. Xing, and A. Ghorbani, "Weighted pagerank algorithm," in *Proceedings. Second Annual Conference on Communication Networks and Services Research*, 2004, pp. 305-314.

[15] J. Madhavan, D. Ko, L. Kot, V. Ganapathy, A. Rasmussen, and A. Halevy, "Google's deep web crawl," in *Proceedings of the VLDB Endowment*, 2008, pp. 1241-1252.

[16] J. Madhavan, P. A. Bernstein, A. Doan, and A. Halevy, "Corpus-based schema matching," in *Proceedings. 21st International Conference on Data Engineering, 2005. ICDE 2005.*, 2005, pp. 57-68.

[17] A. Ntoulas, P. Zerfos, and J. Cho, "Downloading textual hidden web content through keyword queries," in *Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries*, 2005, pp. 100-109.

[18] S. Raghavan, and H. Garcia-Molina, "Crawling the hidden web," Stanford, 2000.

[19] P. Wu, J. R. Wen, H. Liu, and W. Y. Ma, "Query selection techniques for efficient crawling of structured web sources," in *Proceedings of the 22nd International Conference on Data Engineering, 2006. ICDE'06*, 2006, pp. 47-47.