From Citation Network to Study Map: A Novel Model to Reorganize Academic Literatures

Shibo Tao
SPCCTA, School of Electronics and Computer Engineering, Peking University
expo.tao@gmail.com

Xiaorong Wang
Technology and Strategy Research Center,China Electric Power Research Institute, Beijing, China
xrwang@epri.sgcc.com.cn

Weijing Huang
Key Laboratory of High Confidence Software Technologies (Ministry of Education), EECS, PKU
huangwaleking@gmail.com

Wei Chen*
Key Laboratory of High Confidence Software Technologies (Ministry of Education), EECS, PKU
pekingchenwei@pku.edu.cn

Tengjiao Wang
Key Laboratory of High Confidence Software Technologies (Ministry of Education), EECS, PKU
tjwang@pku.edu.cn

Kai Lei
SPCCTA, School of Electronics and Computer Engineering, Peking University
leik@pkusz.edu.cn

ABSTRACT

As the number of academic papers and new technologies soars, it has been increasingly difficult for researchers, especially beginners, to enter a new research field. Researchers often need to study a promising paper in depth to keep up with the forefront of technology. Traditional Query-Oriented study method is time-consuming and even tedious. For a given paper, existent academic search engines like Google Scholar tend to recommend relevant papers, failing to reveal the knowledge structure. The state-of-the-art Map-Oriented study methods such as AMiner and AceMap can structure scholar information, but they're too coarse-grained to dig into the underlying principles of a specific paper. To address this problem, we propose a Study-Map Oriented method and a novel model called RIDP (**R**eference **I**njection based **D**ouble-Damping **P**ageRank) to help researchers study a given paper more efficiently and thoroughly. RIDP integrates newly designed Reference Injection based Topic Analysis method and Double-Damping PageRank algorithm to mine a Study Map out of massive academic papers in order to guide researchers to dig into the underlying principles of a specific paper. Experiment results on real datasets and pilot user studies indicate that our method can help researchers acquire knowledge more efficiently, and grasp knowledge structure systematically.

Keywords

Reference Injection; Topic Analysis; Double-Damping PageRank; Study Map; Academic Papers

*Corresponding author.

WWW 2017 Companion, April 3–7, 2017, Perth, Australia.
ACM 978-1-4503-4914-7/17/04.
http://dx.doi.org/10.1145/3041021.3053059

1. INTRODUCTION

Academic papers usually represent the forefront of technology, and provide good approaches to grasp the knowledge structure in a specific field. Researchers often need to study promising papers in-depth to acquire new knowledge. However, relevant academic information is usually overloaded, and researchers often find themselves overwhelmed by the increasing number of publications. While existing academic search engines such as Google Scholar[1] (abbreviated as GS), Microsoft Academic[2] etc. are efficient in fuzzy query, and can recommend relevant papers to researchers in form of a long list, oftentimes these results are inadequate, owing to lacking of in-depth exploration on a specific field.

We have observed that "relevant" doesn't always mean "meaningful", especially for beginners. For instance, if researchers want to study the theoretical basis of cleansing big data, they search paper entitled "BigDansing: A System for Big Data Cleansing" [9] in GS. 26 results[3] are provided by GS, and most of them are indeed relevant to "BigDansing" on subject. But, what researchers really need are papers about Data Cleansing methods, Distributed Computing Framework etc.. Those wandering results provided by GS are of little use when researchers want to have a glimpse of the structure of a specific paper's underlying principles. Researchers have to rephrase the query and read numerous papers to identify the meaningful ones. By contrast, although Map-Oriented study methods like AceMap [13], Metro Maps of Science[12] and AMiner [14] etc. can structure academic information such as authors, affiliations and venues etc., these methods fail to focus on specific papers and therefore are too coarse-grained to guide researchers to do in-depth study. Generally, existing Query-Oriented study methods and Map-Oriented ones are more suitable for experienced researchers to find relevant papers in new research fields.

We have also observed that an academic paper's references are usually adequate to the understanding of its underlying principles, yet these references are not listed according to their significance to the paper. As is illustrated in Fig.1, the number of papers increases

[1] https://scholar.google.com/
[2] https://academic.microsoft.com/
[3] Retrieved on January 10, 2017.

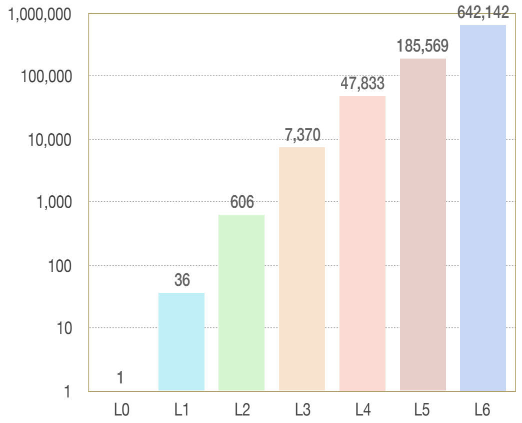


Figure 1: Level Count of paper "BigDansing" [9]; x-axis represents different levels, and y-axis represents the logarithmic number of papers.

exponentially as reference chain grows, it's increasingly difficult for researchers to identify meaningful papers, namely guiding papers, out of hundreds of thousands of papers. Guiding papers are the cited papers that value most in grasping a citing paper's underlying principles, and guiding intensity is used to measure the guiding significance a cited paper to its citing paper. For example, Fig.2 contains two learning routes: *data cleansing* and *big data processing*. Specifically, "MapReduce" is a guiding paper of "BigDansing" with guiding intensity 0.62, and "Google File System" is a guiding paper of "MapReduce" with guiding intensity 0.65. In other words, "MapReduce" and "Google File System" are the most meaningful papers to learn *big data processing*.

To address these challenges, we propose a novel model named RIDP (**R**eference **I**njection based **D**ouble-Damping **P**ageRank) to mine the Study Map that consists of a specific paper's guiding papers out of massive academic papers. In this way, researchers can immediately identify a given paper's guiding papers and look through the structure of its underlying principles. It's worth mentioning that RIDP model can be applied to different domains such as Big Data Processing, Software Engineering, Electric Power Network etc.. Experiment results indicate that our Study-Map Oriented method outperforms traditional Query Oriented method. The main contributions of our study can be summarized as follows:

1. We propose an original academic literatures organization schema —Study Map, to provide proper learning routes for researchers to understand a specific paper.

2. We propose a novel RIDP model to mine fine-grained Study Map to reveal the structure of a specific paper's underlying principles from massive academic papers.

3. We propose a Reference Injection based Topic Analysis method to reveal the guiding intensity between citing papers and cited papers, and a Double-Damping PageRank algorithm to identify the guiding papers.

2. RELATED WORKS

Many scholars have done much work on academic search engines and recommender systems by using both citation analysis and topic models. SimRank [7], inspired by PageRank [11] aims to analyze structural similarity between two papers, but have little concern with their semantic information. [16] tries to recommend scientific articles by incorporating textual content into the traditional matrix factorization. All of the above methods fail to reveal the overall knowledge structure. The recent work [15] classifies papers into important ones and non-important ones via a supervised approach, but neglects the overall interconnection between papers.

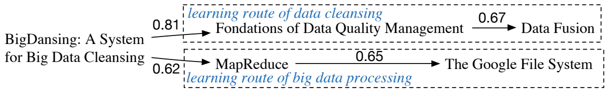

learning route of data cleansing
BigDansing: A System
for Big Data Cleansing
0.81
Fondations of Data Quality Management
0.67
Data Fusion
0.62
MapReduce
0.65
The Google File System
learning route of big data processing

Figure 2: A portion of the Study Map (Fig.8) of paper "BigDansing" [9]. Edges represent citing relations, and the weight of each edge represents guiding intensity.

Some researchers have noticed the limitations of academic search engines and recommender systems. AceMap[13] tries to display relationship among academic entities. Eigenfactor.org [4] builds a field-level interactive academic landscape to present the internal interrelation. AMiner[14] focuses on the evaluation of the influence of researchers by analyzing social network. AKMiner [5] aims to mine useful terminologies and to present them in a knowledge graph by using MLN (Markov Logic Network). Metro Maps of Science[12] tries to excavate the story line using Coherence, Coverage and Connectivity concepts. All of the above methods are too coarse-grained to guide researchers to do in-depth study. Metacademy[5] graphically organizes topics and their prerequisite relationships: each topic is a node in a DAG (Directed Acyclic Graph) study map and contains several key literatures. The map clearly illustrates the learning routes and is very helpful for beginners to acquire knowledge systematically, rather than learn some fragmented knowledge. But such a study map is manually edited, which is a difficult and time-consuming task—only experienced experts in related fields can do it well. HistCite[6] (History of Cite) utilizes LCS (Local Citation Score), GCS (Global Citation Score), LCR (Local Cited References), and CR (Cited References), to locate key papers and authors in a specific field according to cited times and bibliographic information. HistCite tends to recommend frequently cited papers to users. However, such papers are not always guiding papers.

Our goal is to automatically generate a fine-grained Study Map which contains guiding papers for a specific paper by using both semantic and structure information.

3. PROPOSED METHOD

3.1 Definition

DEFINITION 1 (SEED PAPER). *is a promising paper that researchers want to study in-depth. E.g., P0 in Fig.3 is a seed paper.*

DEFINITION 2 (GUIDING PAPER). *is a cited paper which is helpful to understand the underlying principles of its citing papers.*

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

[4] http://www.eigenfactor.org/
[5] https://metacademy.org/
[6] https://en.wikipedia.org/wiki/Histcite

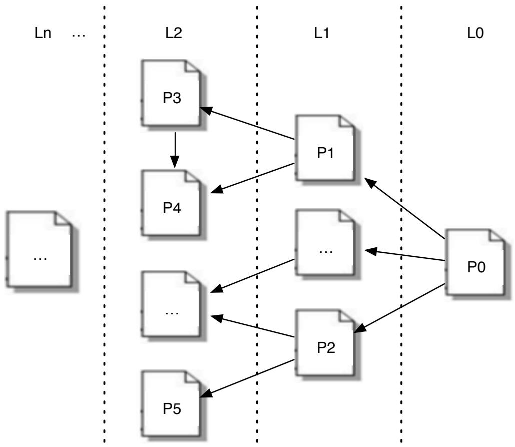


Figure 3: N-level SCDAG Diagram.

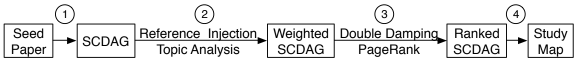


Figure 4: Framework of RIDP model.

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

DEFINITION 5 (RANKED SCDAG). *is a SCDAG in which nodes are ranked.*

DEFINITION 6 (STUDY MAP). *is a SCDAG extracted from Ranked SCDAG, and it contains guiding papers.*

3.2 Framework of RIDP Model

Fig.4 illustrates the framework of RIDP model that includes four phases:

1. For a given seed paper, we use breadth-first algorithm to download its ancestors[7] from Microsoft Academic level by level. Then we can get a SCDAG if we ignore self-reference.

2. Run Reference Injection based Topic Analysis on the SCDAG, and then we can get a Weighted SCDAG.

3. Run Double-Damping PageRank algorithm on Weighted SCDAG, and then we can get a Ranked SCDAG.

4. Generate a Study Map including guiding papers from the Ranked SCDAG.

3.3 Double-Damping PageRank

In this subsection, we detail how to compute the Ranked SCDAG. Although Classic PageRank is effective, it neglects an important fact: some nodes are more important than others. WPR (Weighted PageRank) [19] can assign larger PR values to the important pages rather than divide PR value evenly, and outperforms classic PageRank for queries on focused topics. The classic PageRank and WPR can be denoted as Eq.(1), Eq.(2) respectively.

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

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

In the above equations, $PR(p_u)$ 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 $link(p_u, p_v)$. However, there're still some problems:

1. A paper can't be updated and can only cite older papers, so Citation Graph should be a DAG if we ignore self-reference. The acyclic property will make dangling nodes gain more PR value than usual.

2. SCDAG contains a larger proportion of dangling nodes than the Internet. This will promote dangling nodes to accumulate more PR value.

3. Academic papers are more specialized and more difficult to understand than web pages. Researchers usually have clear purposes when reading a paper, and often read and reread it. While surfers on the Internet tend to browse numerous intelligible fragmented news, blogs etc., and they rarely go back to the viewed pages.

We propose Double-Damping PageRank algorithm to solve the one-way propagation problem of classic PageRank and WPR on SCDAG. The modified PageRank can be denoted as Eq.(3) and Eq.(4):

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

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

Where α is similar to $1-d$ in Eq.(1). β is forward damping factor and similar to d in Eq.(1), γ is the introduced backward damping factor, and reflects the real reading process of researchers. $O(p_u)$ is the set of pages that page p_u links to. $L^{'}(p_v)$ is the in-degree of page p_v. W_1 and W_2 are weights calculated by using Reference Injection based Topic Analysis.

Fig.5 demonstrates that our Double-Damping PageRank algorithm is effective. In order to compare with classic PageRank, we set $\gamma = 0.15$, s.t. $\alpha + \beta = 0.85$. We conduct several experiments on β in range $[0, 0.85]$ with step 0.15. For each β we run Double-Damping PageRank on SCDAG shown in Fig.7(a). When $\beta = 0.85$, the ranking sequence is {*P4,P3,P5,P1,P2,P0*}, it's equivalent to running classic PageRank. The rankings of paper *P1,P2,P0* move forward as β decreases, e.g., the ranking sequence becomes {*P0,P2,P1,P4,P3,P5*} when $\beta = 0.35$, and *P0,P2,P1* have occupy the first three positions respectively. Fig.6 demonstrates the efficiency of Double-Damping PageRank algorithm. Convergence rate slows down as β decreases, because the introduced damping factor γ propagates PR value backwards. Taking both performance and accuracy of the results into account, we empirically set $\alpha = 0.05, \beta = 0.6, \gamma = 0.35$ in the following experiments.

$$\arg\max_{\beta}[\lambda_1 \log p(\boldsymbol{w}|\alpha, \beta) + (1-\lambda_1) \log p(\boldsymbol{w'}|\alpha, \beta)] \tag{5}$$

[7] A paper's references, references' references and so on.

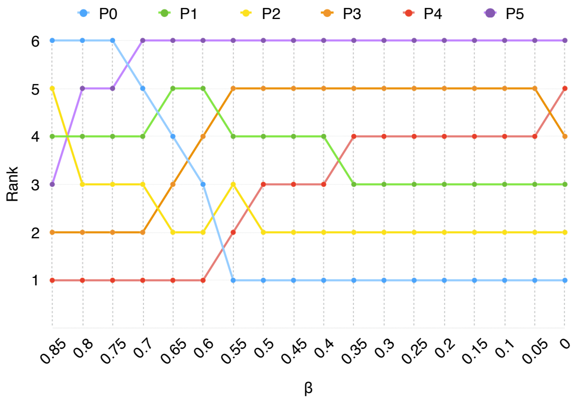


Figure 5: Effectiveness of Double-Damping PageRank algorithm.

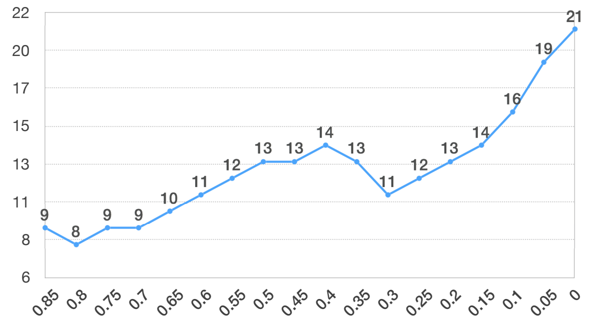


Figure 6: Convergence rate of Double-Damping PageRank. X-axis is β in range from 0.85 to 0 with step -0.05, and y-axis is the number of iterations until algorithm converges.

3.4 Reference Injection based Topic Analysis

In this subsection, we detail how to compute the weights on the edges of SCDAG to generate the Weighted SCDAG. The weight indicates the guiding intensity between citing papers and cited papers. Topic analysis is designed for this task—not only the text content but the inherent topic in the paper should be considered.

We propose Reference Injection based Topic Analysis, which is rooted in the fact that the topic can be better learned when "injecting" the references' content into the paper as a supplement. Take the paper "BigDansing" [9] as an example, it's about big data cleansing, and cites such papers as "Foundations of Data Quality Management" and "MapReduce: simplified data processing on large clusters", both of which are also related to this topic. Therefore, paper d's content $\boldsymbol{w}_d$ and its cited papers' ensemble content $\boldsymbol{w}'_d$ are assumed to share the same inherent K-dimension topic θ_d. Each single word w_{dn} or $w'_{dn'}$ is controlled by the hidden topic assignment z_{dn} or $z'_{dn'}$ (which is sampled from the multinomial distribution $Multi(\theta_d)$), then generated from the word distribution $\beta_{z_{dn}}$ or $\beta_{z_{dn'}}$. Under this assumption, the optimization is as expression(5), where λ_1 $(0 \leq \lambda_1 \leq 1)$ is the parameter to trade off the two targets and α is the hyper parameter in the distribution $\theta_d \sim Dirichlet(\alpha)$. Once λ_1 is set to be 0 or 1, the reference injection based topic analysis would be degenerated as the typical LDA [1]. The smaller the parameter λ_1, the more the references' effect on the topic θ_d is taken into account.

The above optimization target is infeasible to obtain a closed form solution, but still can be solved by variational inference as [1]. We omit the derivation process and retain the result shown in the Eq.(6)(7) as the E-step in the inference and Eq.(8) as the M-step, where γ, ϕ, ϕ' are the variational parameters introduced for θ, z and z' respectively.

$$\phi_{dnk} = \frac{\beta_{k,w_{dn}} \exp \Psi(\gamma_{dk})}{\sum_{k=1}^{K} [\beta_{k,w_{dn}} \exp \Psi(\gamma_{dk})]} \tag{6a}$$

$$\phi'_{dn'k} = \frac{\beta_{k,w'_{dn'}} \exp \Psi(\gamma_{dk})}{\sum_{k=1}^{K} [\beta_{k,w'_{dn'}} \exp \Psi(\gamma_{dk})]} \tag{6b}$$

$$\gamma_{dk} = \alpha_k + \lambda_1 \sum_{n=1}^{N_d} \phi_{dnk} + (1-\lambda_1) \sum_{n'=1}^{N'_d} \phi'_{dn'k} \tag{7}$$

$$\beta_{kv} \propto \lambda_1 \sum_{n=1}^{N_d} \phi_{dnk} I(w_{dn} = v + (1-\lambda_1) \sum_{n'=1}^{N'_d} \phi'_{dn'k} I(w'_{dn'} = v) \tag{8}$$

After the necessary EM iterations, we approximate the optimal solution of the target expression (5), then learn the papers' topic as $E_{q(\theta_d|\gamma_d)}\theta_{dk} = \frac{\gamma_{dk}}{\sum_{k=1}^{K}\gamma_{dk}}$. The above process is shown in Algorithm 1. We empirically set λ_1 to be 0.6 to inject references' content rationally. Finally the weight on the edge (d,d') is set to be the cosine distance between θ_d and θ'_d.

Algorithm 1: Variational Inference Learning Algorithm for Reference Injection Based Topic Analysis

```
Input: Papers' words $\boldsymbol{w}_{1:D}$, words in reference papers $\boldsymbol{w}'_{1:D}$, the parameters
       $\alpha$, $\lambda_1$, $K$
Output: Papers' topics $\theta_{1:D}$ and word distributions on topics $\beta$
while not convergence do
    /* E-step                                                     */
    for $d = 1$ to $D$ do
        while not convergence do
            Given $\boldsymbol{w}_d$ and $\boldsymbol{w}'_d$, update $\phi_d$ and $\phi'_d$ by Eq.(6), $\gamma_d$ by Eq.(7).
    /* M-step                                                     */
    Update $\beta$ by using Eq.(8) and the normalization.
for $d = 1$ to $D$ do
    Update paper's topic $\theta_d$ as $\gamma_{dk}/(\sum_{k=1}^{K}\gamma_{dk})$.
return $\theta_{1:D}$, $\beta$
```

3.5 Study Map Generation

The motivation we choose Weighted PageRank as our base model to identify guiding papers is inspired by three key observations:

1. PageRank has been widely used and proved beneficial in extracting meaningful terms.

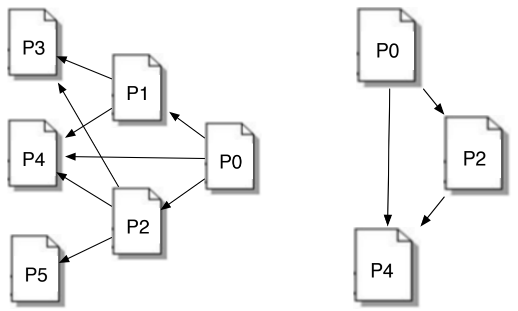
2. The more frequently a paper is cited, the bigger guiding significance it tends to have.

3. Some cited papers have more guiding intensity than others.

For these observations, our proposed Reference Injection based (Weighted) Double-Damping PageRank is reasonable. Finally, we extract a Study Map from the Ranked SCDAG using the following steps:

1. Sort all papers in Ranked SCDAG in descending order.

2. Take the first N papers, denoted as $topK = N$, and then construct a subgraph using these N papers.



(a) Citation Graph (b) Study Map

Figure 7: (a) A SCDAG containing six papers. (b) A Study Map extracted from Fig.7(a).

3. Remove isolated papers that cannot be reached from seed paper, out of the subgraph we got in the last step. Then we get Study Map for the seed paper.

Take Fig.7(a) as an example, assume that all papers in this toy dataset have the same topic distribution. Given $\beta = 0.6$ and $topK = 3$, we can get a ranking sequence {*P0,P4,P2,P1,P3,P5*} and a corresponding Study Map shown in Fig.7(b).

4. EXPERIMENTS AND RESULTS

4.1 Dataset Description

We have observed that guiding intensity between a seed paper and its ancestors will weaken as the reference chain grows, which is an analogy of the Small World Problem [10]. Based on this observation, we crawl up to L6 at most for each seed paper. Table 1 contains the seed papers we used in experiments including their publication year, field, and paper count on each level. In our experiments, only title and abstract sections are used, because these sections have contained academic papers' subject matter; while the other text sections—such as introduction, approach and experiments—often contain much noise, such as equations, figures and tables.

4.2 Effectiveness

Study Map Example. As is previously mentioned, taking both performance and accuracy of the results into account, we set $\alpha = 0.05, \beta = 0.6, \gamma = 0.35, \lambda_1 = 0.6$ in the following experiments.

Fig.8 shows the Study Map automatically generated for paper "BigDansing" [9] with $topK = 20$. In order to show the advantages of Study Map vividly, we draw its papers in red, green, and blue; we draw the other non-study-map papers in grey and yellow, and compare it with the former. Due to the limited space, we only draw a portion of non-study-map nodes. To avoid the suspicion of selecting not-good papers intentionally, we require that the compared papers should obey the following criteria: (1) it's an isolated paper—pathetically, it ranks in the top 20 list but does not appear in the study map, owing to its isolation from the seed paper; or (2) it's a bridging paper, which connects the isolated papers and the study map papers.

In Fig.8, *Paper-{0∼14}* are study-map papers, specifically, *paper-0* is the seed paper, *paper-{1∼7}* are *L1-Papers* which are cited by seed paper directly, *paper-{8∼14}* are *L2-Papers* which are cited by the seed paper's references directly. *Paper-{15∼19}* are isolated papers, and *paper-{20∼22}* are bridging papers. As we can see, the seed paper is about big data cleansing, *L1-Papers* are the guiding papers for the seed paper. Specifically, *paper-1* and *paper-2* will help researchers to understand the underlying principles of data cleansing, and *paper-3*, an illustrious Distributed Computing Framework will help to understand the state-of-the-art "Hadoop"[8] and "Spark"[20] which are necessary in big data processing. "Shark"[18] is a SQL data analysis tool based on "Spark". Obviously, "MapReduce"[2] is more instructive than "Shark". *paper-{5∼7}* are artifices about data cleaning, researchers should be more willing to read fundamental *paper-1* and easy-to-use *paper-2* rather than tricky papers. *L2-Papers* are the selected guiding papers for *L1-Papers* from all 606 papers in L2. E.g., *paper-8* is about data fusion, and *paper-10* is about Database Repairing and Consistent Query. Obviously, They are helpful to understand the underlying principles of *paper-2* which is about Data Quality Management. The bold lines in Fig.8 are the main study routes. The reason why isolated papers appear in top 20 is they are important in their sub-SCDAG respectively. But they have little guiding relevance to the seed paper. E.g., *paper-15* and *paper-17* are all about large-scale graph computation which have limited relation with data cleansing, they establish connections with the seed paper via bridging *paper-20* (GraphX[17]) which is a graph computing tool on "Spark". "GraphX" might be helpful in processing large-scale graph, but researchers seldom use them in learning stage unless they have plenty of time and energy. As previously mentioned, top-cited papers are not always guiding papers. For instance, among the top 5 cited papers of "BigDansing" [9] in SCDAG, as shown in Table 2, only "MapReduce" appears in Study Map Fig.8. *Paper-B* focuses on repairing constraints. *Paper-C* is "Shell" built on "MapReduce", and it's harder than "Shark" to learn. *Paper-D* aims to elaborate problems in "Data Cleaning", which is particularly difficult for beginners. *Paper-E* focuses on capturing data inconsistencies. Although these four papers are relevant, they have low guiding intensity, and Reference Injection can handle such seeming inconsistence.

In addition, we can generate a bigger Study Map by simply increasing $topK$. Guiding papers on higher levels will gradually appear as $topK$ increases, and users can conduct more in-depth exploration according to their individual needs.

User Study. In our user study, we evaluated the effectiveness of Study Map in aiding users to navigate, integrate, and understand the whole structure of the underlying principles for a specific paper. User Study is an evaluation method borrowed from Metro Maps of Science [12], since there is no previous work to compare with, and evaluating Study Map has certain big challenges.

We conduct several experiments on paper "BigDansing" [9]. We choose this paper because many researchers have data cleaning requirements at work, and want to learn some basic knowledge about Distributed System in the era of big data. In this way, our participants willingly carry out the experiment, which will ensure the reliability of the results. In order to generate a Study Map big enough for the subsequent experiments, we set $topK = 50$. Finally, we generate a Study Map containing 41 papers (12 *L1-Papers*, 27 *L2-Papers*, and 1 *L3-Papers*) and 96 edges.

We recruited 15 participants from our school. All participants were graduate students with background in Computer Science, and they meet the following two conditions: (1) have passed CET6; (2) not knowing about Data Cleansing or Distributed System well in advance. The 15 participants are divided into three groups: GS, GS+KG, and GS+SM. In GS+KG, participants are given a Knowledge Graph generated from the Wikipedia entries: Data cleansing[9]

[8] http://hadoop.apache.org/

[9] https://en.wikipedia.org/wiki/Data_cleansing

Table 1: Dataset

Paper	Year	Field	L1	L2	L3	L4	L5	L6
[9]	2015	Data Cleansing	36	606	7,370	47,833	185,569	642,142
[8]	2016	Software Engineering	41	972	10,257	63,266	246,000	714,037
[6]	2016	Event Detection	12	225	3,154	29,669	194,698	861,527
[3]	2016	SparQL Optimization	30	564	5,183	30,759	132,233	493,516
[4]	2016	Violence detection	20	469	6,219	46,834	237,707	876,002

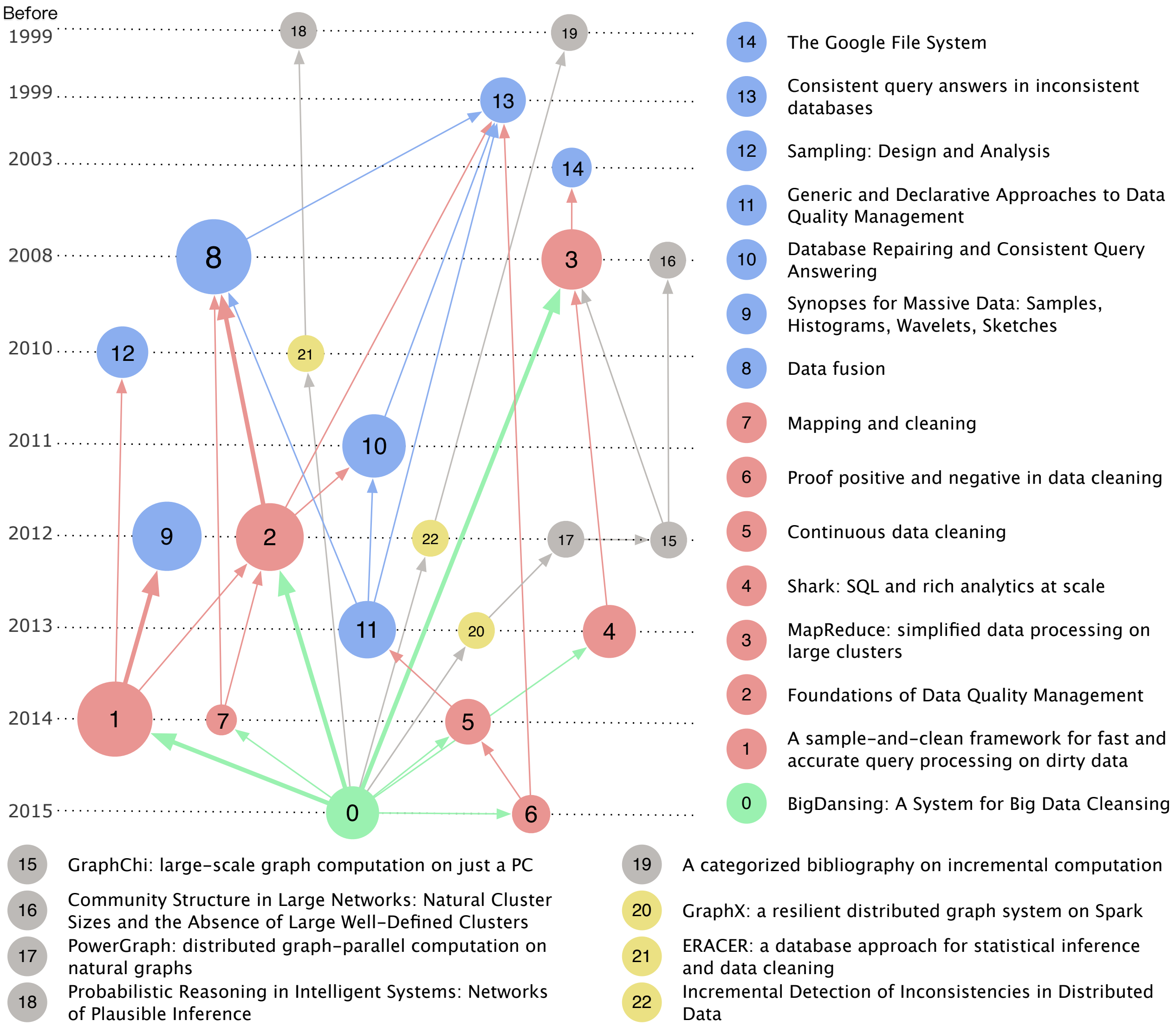

Before 1999
1999
2003
2008
2010
2011
2012
2013
2014
2015
18
19
13
14
8
3
16
12
21
10
9
2
22
17
15
11
20
4
1
7
5
0
6
14 The Google File System
13 Consistent query answers in inconsistent databases
12 Sampling: Design and Analysis
11 Generic and Declarative Approaches to Data Quality Management
10 Database Repairing and Consistent Query Answering
9 Synopses for Massive Data: Samples, Histograms, Wavelets, Sketches
8 Data fusion
7 Mapping and cleaning
6 Proof positive and negative in data cleaning
5 Continuous data cleaning
4 Shark: SQL and rich analytics at scale
3 MapReduce: simplified data processing on large clusters
2 Foundations of Data Quality Management
1 A sample-and-clean framework for fast and accurate query processing on dirty data
0 BigDansing: A System for Big Data Cleansing
15 GraphChi: large-scale graph computation on just a PC
16 Community Structure in Large Networks: Natural Cluster Sizes and the Absence of Large Well-Defined Clusters
17 PowerGraph: distributed graph-parallel computation on natural graphs
18 Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference
19 A categorized bibliography on incremental computation
20 GraphX: a resilient distributed graph system on Spark
21 ERACER: a database approach for statistical inference and data cleaning
22 Incremental Detection of Inconsistencies in Distributed Data

Figure 8: A Study Map of paper "BigDansing" [9]. Each node is a paper, and node size indicates its importance.

Table 2: Top 5 Cited Papers of "BigDansing" [9]

ID	Title	Cited Count
A	MapReduce: simplified data processing on large clusters	70
B	A cost-based model and effective heuristic for repairing constraints by value modification	50
C	Pig latin: a not-so-foreign language for data processing	37
D	Data Cleaning: Problems and Current Approaches	32
E	Conditional functional dependencies for capturing data inconsistencies	31

Table 3: User Study on [9] (Unit: Minute)

Group	T1	T2	T3	T4	T5	Avg
GS (Google Scholar)	116	141	148	167	194	153
GS+KG (Google Scholar + Knowledge Graph)	98	129	136	143	181	137
GS+SM (Google Scholar + Study Map)	56	77	85	91	109	83

and Big data[10] (We choose Wikipedia because we can't find any appropriate maps from existing map tools.). In GS+SM, participants are given a pre-computed Study Map, but they were only informed the meaning of nodes and edges in it. All participants in the three groups are allowed to use Google Scholar.

We asked all the participants to imagine themselves interested in big data cleansing, intending to dig into the underlying principles of "BigDansing" [9]. In particular, they were asked to write a brief report including the guiding papers and their understanding of the underlying principles of "BigDansing" [9]. We have an expert to assess their learning outcomes and point out their flaws. The expert doesn't know which group a participant belongs to. Participants will keep adjusting their performance until they meet the expert's requirements. We recorded the time required to write a satisfactory report of all participants, and the expert's assessment time is excluded. The results are summarized in Table 3.

Table 3 shows that group GS+SM takes an average of 83 minutes, and group GS and group GS+KG take an average of 153 minutes, 137 minutes respectively. In other words, the map users are much more efficient to write a satisfactory report. To our surprise, Knowledge Graph can improve efficiency slightly. We interviewed participants in group GS+KG, and found that concepts in Knowledge Graph can only give an intuitive feeling of knowledge structure, but fail to reveal how to study; therefore, researchers still have to do numerous searches to find guiding papers on each concept.

Expert Evaluation. In order to further verify that papers in our Study Map are guiding papers. we invite an expert for each paper in Table 1 to evaluate the Study Map we generated automatically. First, we show experts our Study Map generated with $topK = 50$, and asked them to (1) find out guiding papers in the map; (2) recommend guiding papers that are not included in the map. It is worth mentioning that we can't ask experts to manually generate a Study Map as a ground truth, because identifying guiding papers out of thousands of papers is really time-consuming. The evaluation results are summarized in Table 4. N_{map} is the number of papers in Study Map, N_{hit} is the number of papers labeled by the experts, and $N_{missing}$ is the number of papers that the experts recommend. As we can see, the overall precision of our generated map is above 75%, with the missing rate below 15%. After in-depth analysis, we found that there are several reasons that might affect precision and missing rate:

1. Reference information in Microsoft Academic is incomplete—there are about 30% missing data, and this might lead to low precision or high missing rate;

2. some papers cover a wide range of disciplines, and evaluation on this kind of papers are relatively more subjective, even to experts.

However, these flaws are acceptable. The experts speak highly of our Study Map, because they think our Study Map help them structure knowledge pieces during evaluation, and study routes provided by our Study Map can guide researchers to dig into the underlying principles of a specific paper, especially for beginners.

Table 4: Expert Evaluation

Paper	N$_{map}$	N$_{hit}$	N$_{missing}$	Precision	Missing Rate
[9]	36	27	2	75%	5.6%
[8]	35	28	3	80%	8.6%
[6]	49	37	6	75.5%	12.2%
[3]	50	38	4	76%	8%
[4]	39	30	2	76.9%	5.1%

5. CONCLUSIONS AND FUTURE WORK

In this paper, we propose a novel RIDP model to mine fine-grained Study Map which can reveal the structure of a specific paper's underlying principles out of massive academic papers. Pilot user studies and expert evaluation demonstrate that our automatically-generated Study Map can help researchers acquire knowledge efficiently and systematically. It is worth mentioning that our RIDP model is also efficient, because it only relies on the data earlier than the seed paper's publication time. In future work, we will utilize a paper's descendants[11] to generate an Exploration Map which contains influential papers for researchers to understand the future works based on a specific paper. In this way, we can not only indicate to researchers the direction of future research, but help them understand the original paper better.

6. ACKNOWLEDGMENTS

This research is supported by Natural Science Foundation of China (Grant No.61572043), State Grid Basic Research Program (DZ71-15-004) and National Key Research and Development Program (2016YFB1000700).

7. REFERENCES

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[2] J. Dean and S. Ghemawat. MapReduce: simplified data processing on large clusters. *Communications of the ACM*, 51(1):107–113, 2008.

[3] L. Gai, W. Chen, and T. Wang. ROSIE: Runtime Optimization of SPARQL Queries Using Incremental Evaluation. *arXiv preprint arXiv:1605.06865*, 2016.

[4] Y. Gao, H. Liu, X. Sun, C. Wang, and Y. Liu. Violence detection using Oriented VIolent Flows. *Image and Vision Computing*, 48:37–41, 2016.

[5] S. Huang and X. Wan. AKMiner: Domain-specific knowledge graph mining from academic literatures. In *International Conference on Web Information Systems Engineering*, pages 241–255. Springer, 2013.

[6] W. Huang, W. Chen, L. Zhang, and T. Wang. An Efficient Online Event Detection Method for Microblogs via User Modeling. In *Asia-Pacific Web Conference*, pages 329–341. Springer, 2016.

[7] 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*, pages 538–543. ACM, 2002.

[8] S. Jiang, C. McMillan, and R. Santelices. Do Programmers do Change Impact Analysis in Debugging? *Empirical Software Engineering*, pages 1–39, 2016.

[10] https://en.wikipedia.org/wiki/Big_data

[11] A paper's citing papers, citing papers' citing papers and so on.

[9] Z. Khayyat, I. F. Ilyas, A. Jindal, S. Madden, M. Ouzzani, P. Papotti, J.-A. Quiané-Ruiz, N. Tang, and S. Yin. Bigdansing: A system for big data cleansing. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, pages 1215–1230. ACM, 2015.

[10] S. Milgram. The small world problem. *Psychology today*, 2(1):60–67, 1967.

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

[12] 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.

[13] 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*, pages 437–442. International World Wide Web Conferences Steering Committee, 2016.

[14] 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.

[15] M. Valenzuela, V. Ha, and O. Etzioni. Identifying meaningful citations. In *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[16] 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*, pages 448–456. ACM, 2011.

[17] R. S. Xin, J. E. Gonzalez, M. J. Franklin, and I. Stoica. Graphx: A resilient distributed graph system on spark. In *First International Workshop on Graph Data Management Experiences and Systems*, page 2. ACM, 2013.

[18] R. S. Xin, J. Rosen, M. Zaharia, M. J. Franklin, S. Shenker, and I. Stoica. Shark: SQL and rich analytics at scale. In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of data*, pages 13–24. ACM, 2013.

[19] W. Xing and A. Ghorbani. Weighted pagerank algorithm. In *Communication Networks and Services Research, 2004. Proceedings. Second Annual Conference on*, pages 305–314. IEEE, 2004.

[20] M. Zaharia, M. Chowdhury, M. J. Franklin, S. Shenker, and I. Stoica. Spark: Cluster Computing with Working Sets. *HotCloud*, 10(10-10):95, 2010.