Constructing Seminal Paper Genealogy

Duck-Ho Bae
Hanyang University
Seoul, Korea
dhbae@agape.hanyang.ac.kr

Se-Mi Hwang
Hanyang University
Seoul, Korea
semi@agape.hanyang.ac.kr

Sang-Wook Kim
Hanyang University
Seoul, Korea
wook@hanyang.ac.kr

Christos Faloutsos
Carnegie Mellon University
Pittsburgh, PA, USA
christos@cs.cmu.edu

ABSTRACT

When a researcher starts with a new topic, it would be very useful if seminal papers in the topic and their relationships are provided in advance. We propose an approach to construct seminal paper genealogy and show the effectiveness and efficiency of our approach.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Algorithms, Measurement, Design

Keywords

Seminal Paper Finding, Paper Genealogy Construction

1. INTRODUCTION

When a researcher starts with a new topic, she or he faces the problem of literature survey for the topic. This requires her or him to spend a considerable amount of time for searching for *high-quality* and *relevant* papers from conferences, journals, or scholar search engines. For doing literature survey, it is also important to understand a *research trend* which gives us insight on the topic. If one grasps the research trend in the topic by thorough literature survey, she or he will predict the future research direction well, and thus will perform successful researches. However, during the literature survey, it is extremely difficult to catch such a trend because the coverage of survey could be limited and/or the understanding of the topic could be insufficient.

Therefore, it is crucial to identify seminal papers in a research topic in question and the relationships among them. The construction of *seminal paper genealogy* will reduce the efforts of literature survey greatly and will enable to easily catch the trend of the research domain. More specifically, in this paper, we address two problems: (1) finding the seminal papers and (2) constructing their genealogy automatically from a set of papers belonging to the same topic. We first suggest the four goals for seminal paper finding and propose a novel method that makes them satisfied (Section 2). Second, we identify two requirements of an influence measure for paper genealogy and examine the characteristics of candidates for being used as our influence measure and their applicability to our seminal paper genealogy (Section 3). Finally, we verify the effectiveness and efficiency of our proposed approach through extensive experiments (Section 4).

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
CIKM'11, October 24–28, 2011, Glasgow, Scotland, UK.
Copyright 2011 ACM 978-1-4503-0717-8/11/10 ...$10.00.

2. SEMINAL PAPER FINDING

2.1 Goals for Finding Seminal Papers

A seminal paper could be an old paper that is highly original and thus is cited by a lot of papers for a long time. On the other hand, it also could be a root of a recent trend and recognized as a rising star. The following is our list of goals in those finding seminal papers.

- *Goal 1: Find such papers cited by many papers and/or highly cited papers.* If a paper p is cited by a number of papers and/or highly cited papers, it means that p is a high-quality paper and thus has a great impact on the topic.

- *Goal 2: Find such papers cited a lot by relevant papers in the same topic.* The papers in the same topic published after a paper p have responsibility to cite p if p is authoritative and relevant to them. If they did not cite p, it means that the quality or importance of p in the topic is not that significant.

- *Goal 3: Find such papers cited by the papers that are published much later.* Authors do not want to miss *recent* papers in the reference section of their papers if it is authoritative and relevant to their papers. In this situation, if authors cite an old paper at the time of their writing, this indicates that the old paper is regarded very important. Therefore, a paper is considered seminal if it is cited by many relevant papers that were published much later.

- *Goal 4: Include young seminal papers even though they do not get a lot of citations from other papers.* Normally, recent papers have a chance to get citations much less than old papers because only a small number of papers are published after them. So, they are unlikely to be included as seminal papers if a simple citation-based method is employed[7][9]. However, the inclusion of such recent papers in a set of seminal papers enables users to understand a recent research trend.

2.2 Seminality Score

A *seminality score* of a paper is a score that measures the degree of how much the paper satisfies the four goals mentioned in Section 2.1. We have a citation graph, where nodes represent papers belonging to a given topic and edges do citations among them: i.e., if a paper p_1 cites another paper p_2, a directed edge from p_1 to p_2 appears in the graph. By analyzing the graph, we get seminality scores of all the papers. Because we want interpretable results and thus prefer a probabilistic framework, our method is based on the *Random Walk with Restart* (RWR)[7].

RWR successfully determines the authority of a paper in the citation graph. Therefore, this RWR-based computation for the seminality score makes Goal 1 of seminal paper finding satisfied.

Equation 1 shows the concept of RWR[7].

$$\vec{r} = (1-c)W\vec{r} + c\vec{e} \tag{1}$$

where W is an $n \times n$ row-normalized citation matrix, $\vec{e}$ is an $n \times 1$ restart vector with $e_i = 1/n$, c is a restart probability, and $\vec{r}$ is an $n \times 1$ seminality score vector.

In order to achieve Goal 2, we propose two solutions. First, we introduce a new notion of *customized RWR* (cRWR), which is a method of computing a seminality score for a *single* paper i. We change restart vector $\vec{e}^{(i)}$ in cRWR to be sensitive to the similarity of other papers to paper i instead of uniform distribution that is employed in the original RWR. Namely, for paper i, we employ a customized $\vec{e}^{(i)}$ in RWR-based computation. Here,

$$e_j^{(i)} = sim(i,j)/Z \tag{2}$$

where $sim(i,j)$ is a similarity between paper i and paper j, and Z ($= \sum_{k \in N} sim(i,k)$, *where N is a set of all the papers in the topic*) is a normalization factor.

As a result, owing to customized $\vec{e}^{(i)}$, cRWR accumulates a more seminality score (i.e., more honored) for paper i when it is cited by *relevant* papers rather than by non-relevant papers.

Second, we change citation matrix W based on the similarity between citing and cited papers as follows:

$$W_{i,j} = sim(i,j) \,/ \sum_{k \in CitedBy(i)} sim(i,k) \tag{3}$$

where $CitedBy(i)$ is a set of all the papers cited by paper i. With this modification, when a paper distributes its score to its citing papers, it gives a larger score to a more relevant paper. Thus, paper i gets a high score when it receives citations from many relevant papers. As a result, the two solutions make Goal 2 satisfied.

To achieve Goal 3, we reflect the difference of publication years between two papers in restart vector $\vec{e}^{(i)}$. We adopt an exponential growth function to give some incentive to a pair of papers of a large difference in publication years. The modified restart vector is as follows:

$$e_j^{(i)} = \left\{ sim(i,j) * e^{|yr(i)-yr(j)|/t} \right\} / Z \tag{4}$$

where $yr(i)$ is the publication year of paper i, and t is a scaling parameter that needs to be tuned to get a suitable value.

By using the exponential growth function, $\vec{e}_j^{(i)}$ becomes higher when the difference of publication years between papers i and j gets larger. Thus, paper i gets a high seminality score when it receives many citations from the papers having long temporal-distance from it, which implies Goal 3 is satisfied.

Finally, to find young seminal papers (Goal 4), we modify restart vector $\vec{e}^{(i)}$ in such a way that the restart probabilities of the papers published before paper i are set to 0 as in Equation 5. This makes all the papers take part in a fair competition regardless of their ages: (1) Old papers have a lot of following papers that are published after them and thus can cite them. Therefore, the customized restart vector $\vec{e}^{(i)}$ of old paper i has a lot of small probability values. As a result, old paper i cannot get a high seminality score unless it is cited by most papers published after it. (2) Recent papers, on the other hand, have a few papers published after them and can get a high seminality score if they are cited by a few papers published after them relatively. Thus, a seminality score of paper i is greatly affected by the *ratio* of the papers that are able to cite paper i to the papers that actually cite paper i instead of the $absolute$ *number* of papers that cite paper i. Therefore, Goal 4 can be satisfied.

$$\begin{aligned} e_j^{(i)} &= \left\{ sim(i,j) * e^{|yr(i)-yr(j)|/t} \right\} / Z \quad \text{if } yr(i) \le yr(j) \\ e_j^{(i)} &= 0 \qquad\qquad\qquad\qquad\qquad\qquad\quad \text{otherwise} \end{aligned} \tag{5}$$

3. CONSTRUCTING GENEALOGY

3.1 Overview

A paper is influenced by some papers that have been previously published, which is represented in the form of its citations. However, the degrees of influence of the cited papers on a citing paper are different[1]. In this paper, we model the influence relationship of two papers as a *parent-child relationship*: A parent paper p has a great influence on a child paper c. A *genealogy* is defined as a set of all parent-child relationships among seminal papers in a topic.

We adopt the *Markovian assumption* which is employed in [1]. The authors are sure to cite the papers having a significant influence on their paper. Therefore, in finding the parent-child relationship, it is sufficient to consider those pairs of papers having a citation relationship only, rather than all the possible pairs of papers.

The procedure of constructing genealogy is sketched as follows: For every paper c, we do Step 1 through Step 3.

- **Step 1.** For a paper c, we measure an influence score from every paper p cited by paper c. The paper p is considered as a candidate parent paper.
- **Step 2.** Among candidate parent papers, we select the top m papers having the largest influence scores as the final parent papers of child paper c.
- **Step 3.** we draw directed links from the parent papers to child paper c.

3.2 Influence Score

The *influence score* of a parent paper p to a child paper c indicates the degree of paper c is influenced by paper p. The influence score can be measured in diverse ways. However, to understand the influence between citing and cited papers, the influence measure should satisfy the following requirements.

- *Requirement 1: It should reflect the strength of the influence correctly.* If a cited paper has a great influence on a citing paper, the score between the two papers should appear to be high. On the contrary, it should be low if the two papers are unrelated or not influenced by each other.
- *Requirement 2: It should consider the temporal-distance between citing and cited papers.* If the difference of their publication years is large, the degree of influence between the two papers should be small because there are other more influencial papers published during the time window between the two years. So, it is desirable to penalize the influence score when the difference of publication years is large.

We decided to use text-based[2] or link-based[3][8][10] similarity as our candidate influence measure. The reasoning behind this decision is that if a child paper c was influenced by a parent paper p, then their texts and/or citations should be similar to each other to some degree.

Here, we discuss which of them is appropriate for being used as our influence score. The degree of influence between two papers is well evaluated in all the similarity measures. If a paper c was influenced by another paper p, these two papers are meant to deal with the same issue. So, their texts have a number of common terms. Similarly, they also have common citations. Therefore, Requirement 1 is satisfied with all the similarity measures.

In Requirement 2, text-based similarity measures do not take the temporal-distance between papers into account at all, and thus cannot satisfy Requirement 2. On the other hand, link-based similarity measures consider the temporal-distance. Because authors tend to cite recent quality papers in a topic, old and young papers have different sets of citations even though they deal with the same issue. The similarity between two papers becomes small in this case. Thus, link-based similarity measures satisfy Requirement 2.

4. EXPERIMENTS

4.1 Environment for Experiments

For experiments, we used 1,071,973 papers in DBLP dataset[5] downloaded on March, 2009 and extracted their citation information from Libra[1]. We first found seminal papers and then constructed their genealogy on three well-known topics of clustering, association rules, and graph mining. The ground truth for seminal papers was extracted from the reference section of each topic in a well-known data mining textbook[2].

4.2 Finding Seminal Papers

We measured the *Mean Average Precision* (MAP) of three different paper ranking methods, RWR[7], PopRank[6], and our Seminal Paper Finding (SPF). We conducted 3-fold cross validation for parameter tuning of previous ranking methods. As a result, the parameter c was set to 0.5 for previous ranking methods which was shown to provide the best performance. The parameters for our SPF method were set to $c = 0.5$ and $t = 2.6$.

The results are shown in Table 1. Our proposed method performs better than any other methods. This is because SPF method performs much better in finding recent seminal papers (Table 2 clearly shows this result).

Table 1: MAP scores of three different methods

	RWR	PopRank	SPF
MAP	0.589	0.627	0.670

We also measured the recall at top 30 to evaluate the performance of each method in finding recent papers. Table 2 shows the results. Our SPF method finds both old and recent seminal papers very well and its recall is higher than those of any other methods. Although the recall of SPF in old papers is slightly lower than those of RWR, it boosts recall levels a lot in recent papers.

We did qualitative evaluation for our SPF method. We present only the result of the SPF method on a clustering topic due to the limited space. Table 3 shows the list of the top 20 seminal papers sorted by the publication year. Our method found not only old seminal papers such as CLARANS (#1), DBSCAN (#3), and BIRCH (#4), but also relatively recently-published papers such as Chameleon (#13), LOF (#16), and Data Streams (#19 and #20).

[1]http://academic.research.microsoft.com/

Table 2: Recall scores of three different methods

	RWR	PopRank	SPF
Old	0.726	0.660	0.711
Recent	0.239	0.336	0.435
Average	0.483	0.498	0.573

4.3 Constructing Paper Genealogy

Due to the space limitations, we show only the genealogy on the clustering topic. First, we found 20 seminal papers, shown in Table 3, and then constructed genealogy using Inter-Connection[10], the state-of-the-art link-based similarity method for scientific literature, with the most significant parent-child relationship only as shown in Figure 1.

Figure 1(a) shows the genealogy using Inter-Connection, which satisfies all our requirements. The genealogy of the clustering topic originates from Paper #1 in our database. As time goes by, the topic was developed into several sub-topics such as density-based clustering (#3 and #11), hierarchical clustering (#4, #8, and #18), high dimensional clustering (#7 and #12), outlier detection (#9 and #16), and data stream clustering (#14, #20, #18, and #19).

Because Inter-Connection can reflect the temporal-distance, two papers could be in parent-child relationship when their publication years are close to each other. This contributes to producing a nice genealogy. Even when the temporal-distance between two papers is large, however, Inter-Connection could connect them by a parent-child relationship if they have a strong influence relationship. The connection between Papers #3 and #11 (both address density-based clustering) shows this example.

Figure 1(b) shows the genealogy using Cosine similarity[2]. Different from the genealogy using Inter-Connection, the genealogy originated from two papers (#1 and #2). Because Cosine similarity does not consider the temporal-distance between two papers, it directly connects Paper #11 to Paper #1 (the origin) instead of Paper #3 that has a great influence on Paper #11 and was published later than Paper #1.

In summary, Inter-Connection (link-based similarity method) performs better than Cosine similarity (text-based similarity method) in constructing seminal paper genealogy. On the other two topics, we performed the same experiments and obtained similar results.

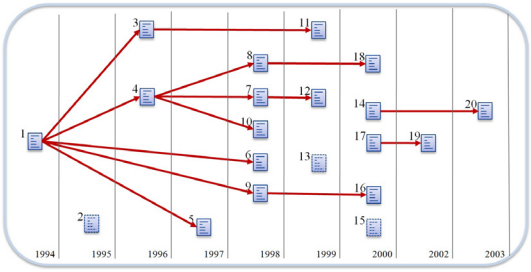
5. RELATED WORK

Ranking Ranking is one of the most important issues in the scientific literature domain. Unlike the ranking algorithms for web pages such as PageRank[7] and HITS[4] that only use hyperlinks, ranking algorithms for academic and scientific literature try to use information such as authors and publication venues in addition to the citation information in order to measure the authority of a paper more accurately. Browsing-based model[9] and PopRank[6] utilize the author-paper relationships in ranking in addition to the citation relationships. Furthermore, PopRank considers additional publication venue-paper relationships.

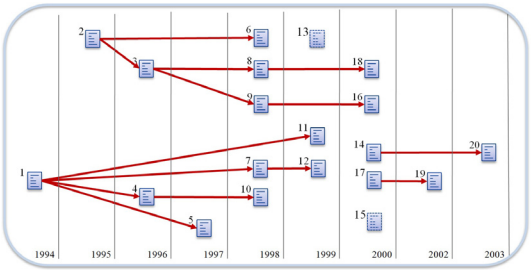
Similarity measure The similarity measures can be classified into two categories: (1) text-based measures that consider the terms contained in the title, abstract, keywords, and body of a paper as features of the paper in similarity computation[2], and (2) link-based measures that consider the papers commonly *cited by* or *citing* a pair of papers in computing similarity of the two papers[3][8][10]

Table 3: Top 20 seminal papers found on the clustering topic

No.	Title	Publisher	Year
1	Efficient and Effective Clustering Methods for Spatial Data Mining	VLDB	1994
2	A Database Interface for Clustering in Large Spatial Databases	KDD	1995
3	A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise	KDD	1996
4	BIRCH: An Efficient Data Clustering Method for Very Large Databases	SIGMOD	1996
5	STING: A Statistical Information Grid Approach to Spatial Data Mining	VLDB	1997
6	Scaling Clustering Algorithms to Large Databases	KDD	1998
7	Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications	SIGMOD	1998
8	CURE: An Efficient Clustering Algorithm for Large Databases	SIGMOD	1998
9	Algorithms for Mining Distance-Based Outliers in Large Datasets	VLDB	1998
10	Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values	DMKD	1998
11	OPTICS: Ordering Points To Identify the Clustering Structure	SIGMOD	1999
12	Fast Algorithms for Projected Clustering	SIGMOD	1999
13	Chameleon: Hierarchical Clustering using Dynamic Modeling	IEEE Computer	1999
14	Clustering Data Streams: Theory and Practice	IEEE TKDE	2000
15	Biclustering of Expression Data	ISMD	2000
16	LOF: Identifying Density-Based Local Outliers	SIGMOD	2000
17	NiagaraCQ: A Scalable Continuous Query System for Internet Databases	SIGMOD	2000
18	ROCK: A Robust Clustering Algorithm for Categorical Attributes	Inf. Syst.	2000
19	Models and Issues in Data Stream Systems	PODS	2002
20	A Framework for Clustering Evolving Data Streams	VLDB	2003



(a) Inter-Connection.



(b) Cosine similarity.

Figure 1: Seminal paper genealogy on the clustering topic.

6. CONCLUSIONS

In this paper, we proposed effective and efficient solutions to the problem of constructing seminal paper genealogy: (1) We formulated the problem of constructing seminal paper genealogy. (2) We proposed a novel algorithm for finding seminal papers in the framework of RWR. (3) We addressed influence scores among papers and proposed a method for genealogy construction by using them. (4) We verified the effectiveness and efficiency of the proposed methods by conducting extensive experiments.

7. ACKNOWLEDGE

This research was supported by the MKE(The Ministry of Knowledge Economy), Korea, under the Convergence-ITRC(Convergence Information Technology Research Center) support program (NIPA-2011-C6150-1101-0001) supervised by the NIPA(National IT Industry Promotion Agency), the National Research Foundation of Korea(NRF) grant funded by the Korea government(MEST) (No. 2008-0061006), and NHN Corp. Any opinions, findings, and conclusions or recommendations expressed in this material are the authors' and do not necessarily reflect those of the sponsors.

8. REFERENCES

[1] L. Dietz, S. Bickel, and T. Scheffer, "Unsupervised Prediction of Citation Influences," In *Proc. Int'l Conf. on Machine Learning*, ICML, pp. 233-244, 2007.

[2] J. Han and M. Kamber, *Data Mining: Concepts and Techniques(2nd Edition)*, Morgan Kaufmann, 2006.

[3] M. Kessler, "Bibliographic Coupling between Scientific Papers," *Journal of the American Documentation*, Vol. 14, No. 1, pp. 10-25, 1963.

[4] J. Kleinberg, "Authoritative Sources in a Hyperlinked Environment," *Journal of the ACM*, Vol. 46, No. 5, pp. 604-632, 1999.

[5] M. Ley, "DBLP: Some Lessons Learned," In *Proc. Int'l Conf. on Very Large DataBases*, VLDB, pp. 1493-1500, 2009.

[6] Z. Nie et al., "Object-Level Ranking: Bringing Order to Web Objects," In *Proc. Int'l Conf. of World Wide Web*, WWW, pp. 567-574, 2005.

[7] L. Page et al., The PageRank Citation Ranking: Bringing Order to the Web, *Technical Report*, Standford University, 1999.

[8] H. Small, "Co-citation in the Scientific Literature: A New Measure of the Relationship between Two Documents," *Journal of the American Society for Information Science*, Vol. 24, No. 4, pp. 265-269, 1973.

[9] S. Yan and D. Lee, "Toward Alternative Measures for Ranking Venues: a Case of Database Research Community," In *Proc. Int'l Conf. on Digital Libraries*, JCDL, 2007.

[10] S. Yoon, S. Kim, and S. Park, "A Link-based Similarity Measure for Scientific Literature," In *Proc. Int'l Conf. of World Wide Web*, WWW, pp. 1213-1214, 2010.