Citation Retrieval in Digital Libraries

Chen Ding, Chi-Hung Chi, Jing Deng, Chun-Lei Dong
School of Computing, National University of Singapore
Lower Kent Ridge Road, Singapore 119260

ABSTRACT

Currently more and more research papers are being published in the form of digital libraries on the web. How to search them efficiently and effectively is a big challenge for researchers. With a static subject tree to index the paper and with the traditional query mechanism, many problems appear. The detail-level topics can hardly be found, the emerging new area can not be identified, the non-semantically related papers can hardly be retrieved, and the key papers in the area can not be obviously pointed out. In order to solve these problems, this paper proposes a novel approach to map the citation retrieval problem into a graph-partitioning problem. All citations in a digital library will be mapped to a citation graph through their reference links. It is observed that the citation graph is not evenly connected. Highly connected sub-graphs will often emerge. Different sub-graph will represent different topic and to partition the graph to higher levels can reveal the detail topics. The different connectivity can also help to find the hot topics, related topics and key citations. Since all these can be done automatically and efficiently, the user's manual effort to search citations will be saved but the results will be more comprehensive and accurate.

1. INTRODUCTION

With the blooming of Internet, the World Wide Web is becoming an important medium for publishing research papers in the form of digital libraries. This kind of publishing is more up-to-date than the paper documents. And it is more convenient for researchers to access than the paper form journal or proceedings. So the web is also becoming an important source for researchers to get information. But owing to the vast volume of the online research papers and the high update rate, it is quite a challenge for the researchers to search for the relevant papers and keep up with the most recent development in their areas. A distinct property of the research paper, different from other web resources, is its bibliographic reference. A bibliographic reference (or citation) means that a document contains, in its bibliography, a reference to another document. This reference can be viewed as a relationship between documents, at least in the author's mind. So the citation is a semantic feature of a document. A citation index is based on the bibliographic references contained in a document, linking the document to the cited works. The citation index allows navigation backward in time (the cited documents) and forward in time (the citing documents). Thus it is a powerful tool for research paper retrieval.

Most of the existing citation-indexing systems are focused on building a large and integrated citation database. But regarding to the searching services they can provide, the functions are limited. Firstly, since the search is based on the keyword instead of concept, if the provided keyword does not match the exact word appearing in the paper, although they are both about the same topic, the paper can not be found as a match. Secondly, the key citations and current hot sub-topics in one area can not be reflected from the returned results. Thirdly, it is hard to find the papers about the related topics of the given area if the papers are not semantically related. Finally, although they update the citation libraries periodically and add the new research papers constantly. But it is implicit. And researchers are difficult to catch up with the recent development of their areas only from the retrieval. All these problems are some basic requirements. No matter the new comers or the knowledgeable researchers of a certain research area have these information requirements. This paper proposes a novel approach to tackle these problems.

The basis of the approach is to form a citation graph from the extracted references. Every vertex in the graph represents a citation paper and every directed edge is a citation occurrence. Obviously the degree of the connectivity for each vertex is different. As a consequence, the highly connected sub-graphs will emerge. The in-degree of a vertex can define the importance of the citation in a research area; the out-link set of the vertex will indicate related topics that might or might not be semantically similar to the current citation. The key procedure of the approach is the graph partitioning. The sub-graphs after partitioning will represent the sub-topics in the collection. The connectivity measurement can reveal the key citation, hot topics and related topics for a certain sub-graph. And if partitioning procedure goes on to the finer level, the topics can be divided into the more detailed level. In this way, the subject tree of the digital library can be formed automatically and dynamically, which can solve the problem of the existing static subject tree in most of the digital libraries. So this approach can help to better locate the user-desired information in many different ways.

The rest of the paper is organized as follows. First the related work is reviewed. Then the attributes of the citation graph are described and the terminology will be defined. The next section explains the algorithms and methods to solve the above mentioned problems. The last section makes the conclusion and comes up with the future directions.

2. RELATED WORK

Citation indexes were originally designed mainly for information retrieval ([3]). And they can also be used in many other ways, e.g. helping to find other publications that may be of interest, finding out the importance of the paper from its cited frequency, identifying research trends and emerging areas. Because of the large volume of the online papers, some researches in citation indexing area are focused on building a universal citation database and providing an autonomous indexing facility. CiteSeer ([4]) is such an effort. There are many

0-7803-5731-0/99$10.00 ©1999 IEEE

ways in which CiteSeer can locate papers to index. CiteSeer locates papers on the web using search engines, heuristics, and web crawling. Other means of locating papers including indexing existing archives, agreements with publishers, and user submission. It also updates the citation library regularly. After the paper is located and downloaded, it is parsed to extract the citations and the context in which the citations are made. Then the citation is indexed and stored in database. Given a paper of interest, it can also find the related papers using various measure of similarity based on term occurrence or citation information.

Except the citation indexing systems, there are also other methods to solve the research paper retrieval problem. WebBase ([8]) is a new project in Stanford University and it is based on the research efforts from Google ([1]) activity. It aims to provide a storage infrastructure for web-like content, store a sizable portion of the web, enable researchers to easily build indexes of the page features and distribute WebBase content via multicast channels. The smart crawling technology is developed from Google. The system provides a feature extraction engine and this engine can be customized to different researchers.

[2] presents the analysis and modeling of the research paper literature. It visualizes a domain-specific information space. The content-similarity analysis is performed to the whole collection and then fed into a structuring and visualizing framework. And author co-citation analysis can also be incorporated into their Generalized Similarity Analysis (GSA) framework. The author co-citation analysis can not only find the interrelationships between pairs of authors, but also easily identify the active sub-fields of research.

3. CITATION GRAPH

The first step to represent the research paper library as a graph is to extract the citations from reference lists of papers. Then the citation graph can be built from citation links. Every vertex v in citation graph G will represent a citation document. A direct edge e_{ij} (v_i, v_j) in G will represent a link reference of v_j by v_i. The set of all the vertices in graph G is represented as $V(G)$. Any vertex v in $V(G)$ has the following attributes:

- *name(v)* - the citation name of v that appears in the document reference list, it is unique and every vertex has only one name. In order to avoid ambiguity, the name of the first appearance of the citation v is taken as *name(v)*.
- *alias(v)* - the set of alias names of citation v, this is in the case when the same citation has different names that appear in different document bibliography. The names after the first appearance are included in this set.
- *title(v)* - the title of the citation document v.
- *source(v)* - the publishing source of the citation document v, it can be the conference name, journal name, technical report name or others.
- *date(v)* - the publishing date of the citation document v.
- *in(v)* - the set of vertices that have directed edges to the vertex v, $\{v_i \mid \forall\, v_i, (v_i, v) \in G\}$.
- $|\, in(v) \,|$ - the in-degree of the vertex v (the cardinality of *in(v)*), it will define the importance of the citation v in the whole collection.
- *out(v)* - the set of vertices that have directed edges from the vertex v, $\{v_i \mid \forall\, v_i, (v, v_i) \in G\}$. It will indicate related topics that might or might not be semantically similar to the citation v.
- $|\, out(v) \,|$ - the out-degree of the vertex v (the cardinality of *out(v)*).
- *weight(v)* - the weight of vertex v, it indicates the relative importance of the citation v in G. The computation of the vertex weight is derived from the PageRank algorithm ([1]). A simplified version is taken here. Let c be a factor used for normalization so that the total weight of all vertices is constant. The formula is,

$$weight(v) = c \sum_{u \in in(v)} \frac{weight(u)}{|\, out(u) \,|}$$

The citation graph G itself also has some characteristics:

- It is a directed graph.
- There is no cycle in the graph, because in research literature, one document can only refer to the publications before its own publishing date. That means between any two documents, all the paths are in one direction.
- The vertices are not evenly distributed in the graph. Some are tightly connected to form the sub-graphs. By adjusting the tightness level, the sub-graphs can be formed in different levels.

Based on the attributes and characteristics of the citation graph, it is possible to find the key citation, hot sub-topics and related topics for a given subset of the graph, and it is also possible to build the subject tree for the digital library dynamically and automatically. To form the subject tree, the content analysis is necessary. Since in this approach, content analysis is only performed on the sub-graph whose size has been largely reduced comparing to the size of the whole collection, the efficiency is highly improved and the accuracy is also improved when collection size is small.

4. CITATION RETRIEVAL PROCEDURE

4.1 Forming The Sub-Graph For The Given Paper

After papers are collected in the repository and citation indexing is finished, the citation graph G can be formed from the citation links. Given a vertex v in graph G, assuming the sub-graph it belongs to is S, if the given vertex is taken as the initial vertex in $V(S)$, then from the citation link expansion and by the control of the tightness of the connectivity for every vertex in S, $V(S)$ can be obtained as the result in specified granularity.

Initially, only one citation v is included in the vertex set $V(S)$ - the starting citation. From its citation links, the set can be expanded to both directions - the citations it references and the

citations it is referenced. This expanding procedure can be continued to several levels. After that, the sub-graph S including the starting citation v can be identified. In this expansion procedure, not all the vertices connected to v directly or indirectly will be added into S. Only those vertices that have tight connectivity with v are included. The starting citation is v and represented as v_{11} in the following formulas. And v_{ij} represents the jth citation in the ith level (the level does not consider the citation link direction, no matter a vertex has an in-link to the existing vertex or out-link from the existing vertex, it is considered as in the same level). The tightness factor is defined by the following formulas:

$$tightness(v_{11}) = 1$$

$$tightness(v_{ij}) = \alpha_1 * \sum_{\substack{(v_{i-1,k}, v_{ij}) \in G \\ or (v_{ij}, v_{i-1,k}) \in G}} \frac{tightness(v_{i-1,k})}{|\, in(v_{i-1,k}) \,| + |\, out(v_{i-1,k}) \,|} + (1 - \alpha_1) * path_{ij}$$

where $path_{ij}$ is the number of edges between v_{ij} and v_{ik} ($k<j$) and α_1 is the fading factor

When the tightness factor of a vertex is larger than a threshold, it is considered to be tight. Then it can be retained in $V(S)$. At the end of this step, all the tight vertices expanded from the starting citation are added into $V(S)$. The initial citation is chosen by the researcher himself, so usually it is highly related to the desired topic and taking it as the starting point can ensure the relevance of the final sub-graph with the topic. Since the vertex in the citation graph usually has not the high connectivity, after imposing the tightness restraint, the vertices in the sub-graph will converge at the small expanding level.

4.2 Finding The Key Citations

The weight of a vertex can determine the relative importance of this citation in the whole citation graph. Every vertex has a weight value after the citation graph is formed. This principle can also be used to determine the important citations in the sub-graph. The equation to compute the weight is the same but the graph is now S instead of G.

The intuitive description of PageRank ([1]) is that a page has high rank if the sum of the ranks of its back-links is high. This intuition is just similar to citation link - if there are a large number of papers citing to a paper, then this paper is very important and thus its weight should be high. So the PageRank algorithm is taken in the approach to compute the weight of the vertex. There are two methods to compute the PageRank in Google system. The reason of taking the simplified version is because the more complex method is to tackle the rank sink problem and there is no such a problem in citation retrieval environment. The equation is recursive but it can be computed by starting with any set of weights and iterating the computation until it converges (its convergence property has been proved in Google). After computing the weights for all the vertices in $V(S)$, the top-ranked citations can be selected as the key citations. If the researcher prefers the recently published papers as the key citations, the *date* attribute of the vertex can be considered and the formula can be adjusted to reflect this requirement.

4.3 Finding The Current Hot Topics

After obtaining the key citations, the researchers can have some fundamental understandings of the area. To help to find the currently hot sub-topics or identify the research trends in the area is another requirement from the researchers. There can be multiple such topics in the area. Assume the set of the papers on the current hot topic t is represented as $HS(t)$ and the vertex in the set is v_t. There are several characteristics of $HS(t)$ and v_t.

- The publishing date of v_t is very late. This is determined by $date(v_t)$.
- There are few papers citing v_t. This is determined by in-degree of the vertex $|in(v_t)|$.
- Every v_t has the similar out-link set $out(v_t)$.

  $$|\bigcap_{v_t \in HS(t)} out(v_t)| > threshold_{HS}$$

- The number of papers on topic t $|HS(t)|$ should be larger than a threshold because the topic is hot and there should be an enough number of papers talking about it.

The third property is also known as bibliographic coupling ([5], [9]). The underlying hypothesis is that if two papers have a similar bibliography, they must have a similar content, and thus deal with similar subjects. This measure of similarity between two documents v_i and v_j is defined as $|out(v_i) \cap out(v_j)|$. The formula above is taken from it. These properties can also be described in Figure 1.

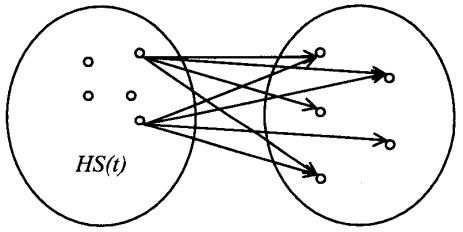

Figure 1: To find the hot topics

From these characteristics, the papers about all the current hot sub-topics can be found from the sub-graph. To get what the hot topics are, the further content analysis (which can be referenced from many information retrieval literatures) is needed. When forming the set $HS(t)$, the content analysis can also be incorporated to ensure the relevance of the vertex v_t with topic t.

4.4 Finding the related topics

In CiteSeer, the related papers of a given paper can be obtained by the term similarity measurement. In this case, usually the two papers are discussing the same topic. Sometimes there is an edge between two vertices in the citation graph, but from the content analysis, they are not on the same topic. So these two citations are related but not semantically related. To know about the related topics of an area can help to better understand the area and broaden the vision of the researchers. There also can be

multiple related topics with the given one. Assume the given paper v_0 is on topic t and the related topic is t_1. The sub-graph including v_0 is represented as S and sub-graph on t_1 is S_1. The vertex in S is represented as v and the vertex in S_1 is v_1. There are some characteristics of the relationship between S and S_1. They can also be described as in Figure 2 (the links within S or S_1 are not drawn).

- There are only 1 or 2 edges linking v to S_1.
- There are only 1 or 2 edges linking v_1 to S.
- $|V(S)|$ and $|V(S_1)|$ should be larger than a threshold to make sure the related topics worthy of further research.
- The total number of edges between S and S_1 is high.

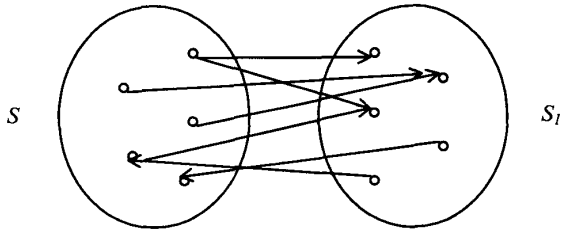

Figure 2. How to find the related topics

When forming the sub-graph, a tightness check is performed to determine whether to add the vertex into the sub-graph S. In usual case, the vertices in S_1 will not be included in S because of the first and second point mentioned above. But they do have connections to vertices in S. Therefore those vertices that can not pass the tightness check are probably talking about the related topics. The connectivity between vertices can help to differentiate the different related topics. Then according to the above characteristics, S_1 can be identified.

4.5 Forming The Subject Tree

The size of the complete citation graph will be very large. If the content analysis is performed on this set to cluster papers and build the subject tree, it is too tough a task to handle either by manual work or by machine. So the first step is to downsize the collection. As mentioned before, the sub-graphs will emerge because of the different connectivity of vertices. Therefore the downsizing problem is actually converted to sub-graph forming. And this graph partitioning procedure can be continued until the size of sub-graphs can be handled. In this approach, the graph partitioning problem can be simplified to the clustering problem and the clustering is only based on the citation link information of the graph.

In citation retrieval area, Small ([7]) proposed the co-citation scheme to measure the relationships between documents. It is computed by:

$$CC(D_i, D_j) = | TO(D_i) \cap TO(D_j) |$$

where $TO(D_i)$ represents the set of documents that refer to document i or its citation set (the set of documents that make reference to a given document). Thus, to be strongly co-cited, two documents must appear together in a large number of documents. In this case, the underlying hypothesis is that co-citation measures the subject similarity established by author group. Of course, this approach favors older documents. In CiteSeer, they use this measurement to compute the similarity between documents. *CCIDF* (Common Citation * Inverse Document Frequency) is the measure they take for the computation.

Citation link is similar to the hypertext link. It is also the link to indicate the relations between the linked two vertices. In hypertext area, there are several different measurements for the hyperlink similarity. HyPursuit ([10]) captures three important notions about certain hyperlink structures that imply semantic relations: a path between two documents, the number of ancestor documents that refer to both documents in question, and the number of descendant documents that both documents refer to. The final hyperlink similarity is a linear combination of the three components.

In this approach, in order to measure the similarity between two citations based on citation links, the similar three components are considered as in HyPursuit. But owing to the distinct aspects of citation links, slightly different formulas are taken to calculate these three parts of similarity:

$$Sim_{ij}^{path} = \frac{1}{2^{spath_i(j)}}$$

$$Sim_{ij}^{an} = \sum_{\substack{x \in common \\ ancestors}} \frac{1}{2^{(spath_i(x)+spath_j(x))}}$$

$$Sim_{ij}^{de} = \sum_{\substack{x \in common \\ descendants}} \frac{1}{2^{(spath_x(i)+spath_x(j))}}$$

where $spath_i(j)$ is the length of the shortest path between v_i and v_j.

The first equation calculates the similarity between two vertices with the measure of the shortest path. The hypothesis is that the similarity between two documents varies inversely with the length of the shortest path between these two. Although the citation graph is directional, the citation link between two documents can be only in one direction because of the backward citation property. So only one direction path is taken into account. The similarity between two documents is proportional to the number of ancestors that the two have in common. This hypothesis is just another kind of description for co-citation scheme. The ancestor here means the document citing the given document. The second equation is the representation for the co-citation scheme. The similarity between two documents is also proportional to the number of descendants that the two have in common. The descendant means the document cited by the given document. This hypothesis is represented in the third equation. When the complete similarity is calculated between two vertices, the linear combination is the solution of our choice in the current stage. Owing to the importance of the co-citation scheme ([6]), the second part of similarity is given a higher weighting factor.

After computing the similarity between all pairs of vertices, any standard clustering algorithm can be executed to get the

document cluster. By adjusting the clustering threshold, the sub-graphs can be formed in different granularity. After the size of the sub-graph is small enough, the content analysis can be performed. In this way, dynamic subject tree can be built. All these steps after citation similarity computation have been comprehensively researched previously, so they are not the focus here.

6. CONCLUSION

Citation retrieval system can help to find research papers on the web. Many of the existing systems have implemented the basic functions for the citation retrieval requirements. This approach is to complement them and to better serve the researchers. From the different citation attributes (mainly the connectivity information), the key citations, the hot topics and related topics of a certain research area can be found out. By citation graph partitioning, the dynamic subject tree can be formed in a detailed level. The initial experimental data can prove the correctness of the approach. But how to better form the sub-graphs from the citation link information is yet to be explored. And the complete system to build the subject tree for a digital library will be implemented in the future work.

REFERENCES

[1] S. Brin and L. Page, "The Anatomy of a Large-scale Hypertextual Web Search Engine", http://google.stanford.edu/~backrub/google.html.

[2] Chaomei Chen and Les Carr, "Trailblazing the Literature of Hypertext: Author Co-Citation Analysis (1989-1998)", *Hypertext99*, 1999.

[3] E. Garfield, "The concept of citation indexing: A unique and innovative tool for navigating the research literature", *Current Contents*, Jan. 3, 1994

[4] C. Lee Giles, Kurt D. Bollacker and Steve Lawrence, "CiteSeer: An Automatic Citation Indexing System", *Digital Libraries 98*

[5] M. M. Kessler, "Bibliographic Coupling between Scientific Papers", *American Documentation*, 24, pp. 123-131, 1963

[6] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*, New York, NY, McGraw Hill, 1986

[7] 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*, 24, pp. 265-269, 1973

[8] *WebBase project* in Stanford university, http://www-diglib.stanford.edu:8080/~testbed/WebBaseDoc/webbaseGoals1.htm

[9] B. H. Weinberg, "Bibliographic Coupling: A Review", *Information Storage and Retrieval*, 10, pp. 189-196, 1974

[10] R. Weiss, B. Velez, M. A. Sheldon, C. Namprempre, P. Szilagyi, A. Duda and D. K. Gifford, "HyPursuit: A Hierarchical Network Search Engine that Exploits Content-Link Hypertext Clustering", *Hypertext96*, 1996.