Discovering Joint Research Topics based on Social Networks Using A Traversing Algorithm

Dongwook Shin, Joongmin Choi

Department of Computer Science and Engineering
Hanyang University, Ansan, KOREA
foremostdw@gmail.com, jmchoi@hanyang.ac.kr
doi:10.4156/jnit.vol1.issue3.6

Abstract

Researchers need to examine trends and novel technologies of their own research areas. With the rapid growth of the Web, however, large amounts of information are generated daily. Therefore, it is generally difficult for researchers to obtain information related to their own areas and novel technologies from huge data residing in the Web. Furthermore, they often try to apply the technologies of their own fields to other different areas to solve existing difficult problems or improve the performance of existing systems. Hence, it is important to discover joint research topics in which technologies of particular research areas are applied to other different areas in order to recognize and follow various current trends. In this paper, we propose a novel method to discover joint research topics using a traversing algorithm based on social networks representing the relations among the authors of papers, and describe some experimental results to show the effectiveness of the proposed method.

Keywords: Joint research topics, Social networks, Traversing algorithm, Digital library.

1. Introduction

Researchers have been devoting a lot of time and effort to discover new technologies and trends of research areas of their interests. In particular, much efforts are given to filter necessary information with increasing cases that the technologies in one area are used in other areas. For this reason, studies on searching for the necessary information for the researchers more easily and quickly have been actively progressed.

In this context, most previous studies focused on recommending papers to the users based on user’s profiles or user's history information having similar interests [1] and filtering meaningless information out of a vast amount of information [2]. However, each research area is rapidly developing based on active information interchange, so joint research topics are accordingly growing fast where a specific technique in an area is applied to the other areas.

For these reasons, it is important that users discover joint research topics. With understanding joint research topics, the researchers could easily acquire the information about the novel technologies and determine the direction of novel studies. Furthermore, it could solve the existing difficult problems or improve the performance of their systems. In the view of a company, there is clearly an advantage that it is possible for the technologies to be transferred to various areas by discovering joint research topics.

For discovering joint research topics, we use a traversing algorithm based on social networks. Social networks apply the relations among humans existing in the real world to the cyber-space. It is possible to detect relevant information or the characteristics of communities through the relations among humans on the cyber-space. Using the advantages of social networks, we can understand the relations among authors and discover joint research topics.

In this paper, we propose the method of discovering joint research topics using a traversing algorithm based on social networks. Our method is referred to as DJRT – Discovering Joint Research Topics.

The rest of this paper is organized as follows. Section 2 discusses and analyzes related work with the DJRT system. Section 3 describes the system architecture of DJRT. Section 4 explains the process of constructing social networks. Section 5 describes the method of discovering joint research topics. Section 6 describes the experimental results. Finally, in section 7 concludes with some future works.
2. Related Work

In the web, people face with an enormous amount of information. As many people publish more research papers, it increasingly becomes difficult to find necessary information quickly. To resolve these problems, studies on the method of efficient research paper search and research paper recommendation have been proposed [3, 4, 5, 6].

Bollacker developed the CiteSeer digital library system to provide up-to-date information on relevant research data for users [3]. Scientific literatures on the Web exist in a form of disorganized database with massive noise data, which makes it difficult to discover knowledge from them through the Web. CiteSeer performs information-filtering and knowledge-discovery functions that automatically extract only relevant records and keep the users up-to-date on relevant researches.

Watanabe [5] proposed the Papits system that is a research support system that shares research information, such as PDF files of research papers, in computers on the networks and classifies the information according to its research types. To develop the Papits system, Watanabe needed to design a mechanism to identify a user’s interest and then he proposed a user’s model using the scale-free network in order to construct effective recommendation system. The scale-free network had vertices, edges and preference information. This method applied paper viewing history to construct a scale free-network based on the word co-occurrence. He additionally defined the topic weight, defined by using two elements; the topic frequency and the topic recency. Based on this information, the Papits system effectively recommended relevant documents to users.

Studies on efficient research paper search and research paper recommendation have been proposed before. But studies of each area have been rapidly improving, and not only their techniques are applied to their own areas but also some joint research topics are emerging widely. For these reasons, it is important for researchers to grasp trends about joint research topics as well as their interesting areas.

As the bibliographic information includes the features such as the title of a paper, author names, author’s affiliations, author’s emails, co-author information and etc, many issues in searching and manipulating research paper information such as paper recommendation [1, 7], paper retrieval [6] and author name disambiguation [8] have been tackled by using social networks or the relations among authors in the digital library context.

Gori proposed a research paper recommending algorithm based on the Citation Graph and random-walker properties [1]. The PaperRank algorithm suggests in this paper is able to assign a preference score to a set of documents contained in a digital library and linked to each other by bibliographic references. PaperRank supports the resource filtering process, in fact it requires a user to select an initial small subset of documents relevant for the topic he/she is writing about. Then the algorithm can spread its boosting effect based on selected papers, through the citation graph in order to discover other interesting and useful resources.

Baghi proposed a semantic web search engine called ConnectA! [6]. This helps the users to find the publications of a community of authors on a specific subject, while they only know a very limited number of the authors in that community. The user sends the list of the known authors and the keywords to the ConnectA!. The engine extracts the related co-authors to the user-entered authors hierarchically and searches for the documents that are written by the extracted authors and contain the keywords.

Zaiane constructed social networks using DBLP data, then found the sequences of vertices using random walk algorithm for detecting relations of vertices based on information of authors [8]. He constructed communities based on detected sequences of vertices and then tried to resolve the name disambiguation problem by considering relationships among authors.

We know the relation of authors through the co-author information of papers and these relations are useful information to discover joint research topics. Co-authors of papers are familiar with each other. Furthermore, most researches have their interesting areas and usually write papers for these areas. For these reasons, papers written by authors contain important terms representing the interesting areas of each author. Hence, social networks that are
constructed by using the features of the bibliographic data are suitable for discovering joint research topics.

For discovering joint research topics in various research areas, we construct social networks by using the relations among authors. We propose the DJRT system that considers both the direct relations and the indirect relations among authors in social networks.

3. System Architecture

The system architecture of DJRT is shown in Figure 1.

The DJRT system performs the process of discovering joint research topics as follows. The Web Robot collects the bibliographic data from the digital library. The Research paper crawler collects the necessary information of papers from the digital library. After that, the Research paper cleaner removes unnecessary contents such as HTML tags.

The Information Extractor extracts information about joint research topics from the collected bibliographic data. The Bibliographic data extractor extracts bibliographic information such as the title of a paper, author names, the publisher, the date of publication and the abstract of a paper. The Topic extractor extracts topic candidates from the abstract of a paper for representing the topic of the paper. Note that we only extract noun phrases as topic candidates using the morphemic analysis and then store extracted information to the Bibliographic Database.

The Social Network Constructor constructs social networks by using the author information as index based on extracted bibliographic information. The Joint Research Topic Detector traverses the social networks and discovers joint research topics based on the similarity measures among the topic candidates of network vertices.

4. Constructing Social Networks based on Bibliographic Data

We collected bibliographic data from the ACM Portal and constructed social networks based on the collected data.

4.1. Bibliographic Data Extraction

Figure 2 represents an example of bibliographic data extraction from the ACM Portal. A regular expression is defined to extract the bibliographic information when the web page content matches with defined patterns in the expression. The extracted information includes the title of a paper, the author names, the author’s affiliation, the publisher, the date of publication, the abstract of the paper and more as shown in Figure 2.
4.2. Topic Candidates Extraction

We extract topic candidates that are the representative of a paper using the morphemic analysis from the abstract of the paper. In the abstract, noun phrases are considered to represent topic candidates of a paper. Because the abstract of a paper represents the summary of the whole content of the paper, we assume that its noun phrases represent important subject terms.

If we consider only nouns as topic candidates, it is difficult for extracted topic candidates to represent topics of the paper because of noun’s ambiguity or meaningless nouns. For example, ‘machine learning’ is a topic candidate representing machine learning fields, but ‘machine’ and ‘learning’ as a separate term are ambiguous to represent machine learning fields. Hence, we extract noun phrases of the abstract as topic candidates.

Since the similarity measure between topic candidates is calculated through the comparison of strings, noun phrases of similar meanings might be judged inharmoniously because of modifiers in the noun phrases. For solving these problems, we first eliminate unnecessary components of a sentence before extracting noun phrases. After that, we perform stop word removal and stemming [9] for the extracted noun phrases. Finally, we regard extracted noun phrases as topic candidates.

Figure 3 is an example of extracting topic candidates in the abstract of a paper. As shown in Figure 3, noun phrases are extracted from an abstract using the morphemic analysis. Topic candidates consist of extracted noun phrases and their frequencies. The morphemic analysis is performed for each sentence as a unit, and for a sentence $S$ a tree structure with $S$ as root is formed.

Only the information about NP (noun phrase) is extracted from this structure and the unnecessary components of the sentences such as PRP (personal pronoun), DT (determiner), RB (adverb), CD (cardinal number), JJR (adjective, comparative) are removed.
4.3. Social Network Construction

The social network is defined as $SN = \{V, E\}$. $V$ is a set of vertices and $E$ is a set of links where a link connects a pair of vertices. A vertex of an author is created based on the extracted author name and the affiliation of the author as shown in Figure 4.

After creating a vertex, we connect co-authors and the author of reference papers to construct social networks. We denote the relations between co-authors by using $rE$ (real edge) links and the relations between the authors of references by using $vE$ (virtual edge) links to allow different weights according to the type of a relation.

Figure 4(c) is a social network constructed based on the information in Figure 4(a) and (b). Vertex 1, 2 and 3 representing the co-authors of the paper are linked by $rE$ and vertex 4 and 5 representing the co-authors of the reference of the paper written by the authors 1, 2 and 3(vertex 1, 2 and 3) are also linked by $rE$. Vertex 1, 2 and 3 are linked to vertex 4 and 5 by $vE$. 

Figure 3. An example of topic candidate extraction

The social network is defined as $SN = \{V, E\}$. $V$ is a set of vertices and $E$ is a set of links where a link connects a pair of vertices. A vertex of an author is created based on the extracted author name and the affiliation of the author as shown in Figure 4.

After creating a vertex, we connect co-authors and the author of reference papers to construct social networks. We denote the relations between co-authors by using $rE$ (real edge) links and the relations between the authors of references by using $vE$ (virtual edge) links to allow different weights according to the type of a relation.

Figure 4(c) is a social network constructed based on the information in Figure 4(a) and (b). Vertex 1, 2 and 3 representing the co-authors of the paper are linked by $rE$ and vertex 4 and 5 representing the co-authors of the reference of the paper written by the authors 1, 2 and 3(vertex 1, 2 and 3) are also linked by $rE$. Vertex 1, 2 and 3 are linked to vertex 4 and 5 by $vE$. 

Figure 4. Constructing social networks
5. Detecting Joint Research Topics

In order to discover joint research topics, the DJRT system traverses the social networks and detects joint research topics based on the similarity measures of interesting areas among authors.

5.1. Traversing Social Networks

Social networks represent the relations between vertices. If an edge exists between vertices, we assume that there is a relationship between them. Therefore, if an edge exists between vertices belonging to different areas; it is highly possible that there is some collaboration between the two research areas.

In this paper, we detect relations between vertices belonging to different areas using a traversing algorithm. The traversing algorithm is described in Algorithm 1.

In Algorithm 1, \texttt{seed} is a start vertex to traverse the social network. \texttt{Neighbors} and \texttt{colV} are lists of vertices and \texttt{colFields} is a list of research areas. \texttt{getNeighbors()} is a function that returns all the vertices linked to the start vertex. \texttt{fields()} is also a function that returns a list of research areas related to the start vertex.

The DJRT system measures the similarity between a start vertex and each linked vertex during traversing the social network (The method of similarity measure method is described in section 5.2). After estimating the similarity measure, we evaluate the similarity \(w\) between vertices.

In case when the similarity \(w\) is greater than threshold \(\alpha\), the system regards them as the same research area, otherwise as a joint research area (We determined threshold \(\alpha\) through experiments described in Section 6.).

An example of discovering joint research topics is shown in Figure 5. If the similarity measure between vertex \(v_1\) and \(v_2\) is larger than threshold \(\alpha\), we decide that both vertex \(v_1\) and \(v_2\) are the same research area.

We assume that vertices of the same research area contain similar topic candidates. Based on this assumption, we consider that both vertex \(v_1\) and \(v_2\) contain similar topic candidates. Hence,
the DJRT system re-evaluates the similarity measure between $v_2$ which is now set up as the start vertex and each of all the vertices linked to $v_2$ while it repetitively traverses the social network.

At this time, the DJRT system traverses from $v_2$ to its neighboring vertices for discovering joint research topics. If the similarity measure between $v_1$ and $v_2$ is smaller than threshold $\alpha$, we decide that they have joint research topics.

![Diagram of discovering joint research topics]

**Figure 5.** An example of discovering joint research topics

### 5.2. Similarity Measures between Research Areas

The DJRT system detects joint research topics using the similarity measures between research areas. In a social network, each vertex is assigned to an author. We use the equation 2 based on cosine similarity [10] to measure the similarity between research areas.
\[
\text{sim}(v_1, v_2) = \frac{v_1 \cdot v_2}{\| v_1 \| \ast \| v_2 \|}
\]

\(v_1\): a vector of the topic candidates of \(v_1\) vertex and directly linked vertices
\(v_2\): a vector of the topic candidates of \(v_2\) vertex and directly linked vertices

\textit{sim()} function takes as inputs the vector of topic candidates in each vertex and returns the similarity \(w\) as a result. The vector of topic candidates in each vertex consists of the topic candidates extracted by using methods describe in Section 4.2. The similarity measure considers the topic candidates of directly linked vertices as well as the topic candidates of each vertex.

If an edge exists between vertices, it means that there is relationship between them. Therefore, there is a high chance of the topic candidates of linked vertices being related with the topic candidates of the vertex. We consider the topic candidates of linked vertices because it is possible that the topic candidates of a vertex (an author) are biased about that vertex.

At this time, we set up the weight value of 0.8 to the topic candidates for the co-author relation linked by \(rE\) (real Edge) and the weight value of 0.5 to the topic candidates for the reference relation linked by \(vE\) (virtual Edge).

6. Experimental Results

To determine the threshold for the similarity measure between research areas, we apply random sampling method [11] to select 100 documents and 218 related vertices from the whole data and we measure the precision, recall and F-measure by using equation 3.

\[
\text{Precision} = \frac{\text{The number of joint research topics that the system detects correctly}}{\text{The number of joint research topics found by the system}}
\]

\[
\text{Recall} = \frac{\text{The number of joint research topics that the system detects correctly}}{\text{The number of correct joint research topic}}
\]

\[
F - \text{measure} = \frac{2 \ast (\text{Precision} \ast \text{Recall})}{\text{Precision} + \text{Recall}}
\]

Figure 6 shows the result of evaluating the similarity measures by varying the threshold values.

![Figure 6](image-url)
Discovering Joint Research Topics based on Social Networks Using A Traversing Algorithm
Dongwook Shin, Joongmin Choi

We set the threshold of the similarity measure between research areas as 0.4, since the performance is maximized at this threshold value. Hence, we eventually evaluated the effectiveness of the DJRT system based on threshold 0.4.

To evaluate the performance of the DJRT system, we randomly selected another 200 documents and 398 related vertices. We measured the precision, recall and F-measure by using equation 3, and the experiment result is shown in Table 1.

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>75.6%</td>
</tr>
<tr>
<td>Recall</td>
<td>78.2%</td>
</tr>
<tr>
<td>F-measure</td>
<td>76.9%</td>
</tr>
</tbody>
</table>

Table 1. The result of performance evaluation

We evaluated the DJRT system performance based on the real world data. As a result, we claim that we obtained satisfactory performance results.

7. Conclusions and Future Work

We have proposed the DJRT system for discovering joint research topics using social networks. It is a tool that helps user’s research as it presents the latest trend or the direction of new studies by discovering joint research topics which the users are interested in.

The DJRT system collects papers in the digital library and extracts necessary information from them and then constructs social networks using the relations between authors and extracted information. Social networks are constructed by considering the reference relation referred to as \( vE \) (virtual edge) as well as the co-author relation referred to as \( rE \) (real edge). Through the detection of relations among authors on the social networks, it is possible to effectively discover joint research topics, and our experimental results showed satisfactory performance.

As future works, we will consider the relations among research areas by clustering authors through the formation of hierarchical relations in social networks. Also we refine the similarity measure methods and the traversing algorithm in order to construct more robust system and then we will evaluate the DJRT system’s performance based on a vast amount of data collection.

8. References


[9] Porter’s Stemming Algorithm
http://tartarus.org/~martin/PorterStemmer/
