Mining Landmark Papers

by

Annu Tuli, Vikram Pudi

in

SIGAI Workshop on Emerging Research Trends in AI
(ERTAI)

Mumbai, India

Report No: IIIT/TR/2010/32

Centre for Data Engineering
International Institute of Information Technology
Hyderabad - 500 032, INDIA
April 2010
Mining Landmark Papers

Annu Tuli and Vikram Pudi
Center for Data Engineering
IIIT-H
Hyderabad, India
{annu@research., vikram}@iiit.ac.in

Abstract—In recent years, the number of electronic journal articles is growing faster than ever before; information is generated faster than people can deal with it. In order to handle this problem, many electronic periodical databases have proposed keyword search methods to decrease the effort and time spent by users in searching the journal’s archives. However, the users still have to deal with a huge number of search results. In this paper, we present the problem of mining landmark papers. We treat papers that introduce important keyphrases for the first time as landmark papers. Our approach combines simple ideas from text mining, information extraction and information retrieval to identify landmark papers. We show that existing related techniques such as first story detection, mining hot topics and theme mining do not effectively handle the landmark paper mining problem. Our approach is simpler and more direct for this task. We experimentally evaluate our approach on a large dataset of papers in the database or data mining areas downloaded using DBLP.

Keywords—Landmark papers, Text Mining, Information Extraction and Retrieval, First-Story Detection.

I. INTRODUCTION

Text Data Mining (TDM) can be considered a field of its own, containing a number of applications. It has also been known as text analysis, text mining or knowledge discovery in text. In general, TDM applications are used to extract non-trivial and useful information from large corpora of text data, which are available in unstructured or structured format. Text mining applications require the use and application of many related fields such as Information Retrieval, Machine Learning, Statistics, and Linguistics. There are various applications of TDM, such as in bioinformatics, market research, consumer trend studies, and scientific research [1].

The internet today is going through a rapid phase of growth and development. With the growth of the internet, information contained in electronic documents is increasingly widespread, with the World Wide Web as its primary repository. The convenience of electronic documents has motivated their more efficient application in information management and knowledge discovery [2].

In many application domains, we encounter a stream of text, in which each text document has some meaningful time stamp [3]. For example, a collection of new articles about a topic and research papers in a subject area can both be viewed as natural text streams with publication dates as time stamps. In such text data streams, there often exist some interesting and meaningful keywords. For example, an event covered in news articles generally has some meaningful keywords consisting of themes (i.e., subtopics) characterizing the beginning, progression, and impact of the event, among others. Similarly, in research papers, some important and meaningful keywords may also exhibit similar patterns. For example, the study of one topic specified by some keyphrases in some time period may have influenced or stimulated the study of another topic associated with same keyphrases after the time period. In all these cases, it would be very useful if we can discover and extract these important keyphrases and also identify the first corresponding paper automatically from text to get knowledge about the keyphrases from where they originate with respect to time stamp. Indeed, such research papers are not only are useful by themselves, but also would facilitate organization and navigation of the information stream according to the underlying keywords.

Consider, for example, there are often hundreds of research papers published annually in a research area. A researcher, especially a beginning researcher, often wants to understand how the research topics in the literature have been evolving. For example, if a researcher wants to know about data mining, both the historical milestones and the recent research trends of data mining would be valuable for him/her. Identifying the origins of important and new keyphrases will also make it much easier for the researcher to selectively choose appropriate new field of research. Also, the corresponding first document (i.e. landmark paper) for that keyphrase will also help the researcher to read only those papers based on his/her research interests.

Mining landmark papers is not only useful for the beginning researcher, but for anyone keeping track of important developments in a particular area. This is important today due to the large numbers of researchers and published research papers. Keeping track of landmark papers is especially useful to track key research developments not necessarily in the specific area of a researcher, but in its numerous related areas, which tends to be voluminous.

This paper is organized as follows: In Section 2, we introduce the related work. In Section 3, we first describe our problem definition and then in section 4 we describe methodology of our work. Detailed experimental results are presented in Section 5. In Section 6, we draw conclusions and present future work.

II. RELATED WORK

Information Retrieval (IR) and Information Extraction (IE) areas are associated with text mining. IE has the goal of transforming a collection of documents into information that is more readily digested and analyzed with the help of an IR system. IE extracts relevant facts from the
documents, while IR selects relevant documents. IE is a kind of pre-processing stage in the text mining process, which is the step after the IR process and before data mining techniques are performed.

A typical information retrieval problem is to locate relevant documents in a document collection based on a user's query, which is often some keywords describing an information need, although it could also be an example relevant document. In such a search problem, a user takes the initiative to “pull” the relevant information out from the collection; this is most appropriate when a user has some ad hoc (i.e. short-term) information need, such as finding information to buy a used car. When a user has a long-term information need (e.g. a researcher's interests), a retrieval system may also take the initiative to “push” any newly arrived information item to a user if the item is judged as being relevant to the user's information need.

Given the avalanche of electronic documents, the pervasive use of search engines helps to minimize the time required to extract information [2]. In the most popular form of search, the search criteria are keywords, or concepts that may be contained in the electronic documents [4]. However, the users still have to deal with the overabundance of search results in some way. During the last decade, the question of how best to filter the results of search engines has become an important issue. Topic detection is an experimental method for automatically organizing search results. It could help users save time in identifying useful information from large scale electronic documents.

A topic is defined to be a seminal event or activity along with all directly related events and activities [5]. Today, many different data mining methods are employed to recognize topics, for instance, the naive bayes classifier [6], hierarchical clustering algorithms (HCA) [7][8][9], paragraph relationship maps [10], formal concept analysis (FCCA) [11] and lexicon chains [12][13][14]. These methods use the frequencies of words to calculate the similarity between two documents. Therefore, their accuracy is greatly hindered by the presence of synonyms.

Halliday and Hasan [13] proposed a semantics-based lexical chain method that can be used to identify the central theme of a document. Based on the lexical chain method, combined with the electronic WordNet database, the proposed method clusters electronic documents by semantic similarity and extracts the important topics for each cluster. Ultimately, the method provides more user-friendly search results that are relevant to the topics.

Shewhart and Wasson [17] described a process that monitors newsfeeds for topics that receive unexpectedly high amounts of coverage (i.e. hot topics) on a given day. They performed trend analysis in order to find hot topics, except that they used controlled vocabulary terms rather than phrases extracted from text. The purpose of the study is to monitor newsfeeds in order to identify when any topic from a predefined list of topics is a hot topic.

In [1], Indro De, presented First Story Detection (FSD) whose task requires identifying those stories within a large set of data that discuss an event that has not already been reported in earlier stories. In this FSD approach, algorithms look for keywords in a news story and compare the story with earlier stories. FSD is defined as the process to find all stories within a corpus of text data that are the first stories describing a certain event [15]. An event is a topic that is described or reported in a number of stories. Examples can be governmental elections, natural disasters, sports events, etc. The First Story Detection process runs sequentially, looking at a time-stamped stream of stories and making the decision based on a comparison of key terms to previous stories. FSD is closely linked to the Topic Detection task, a process that builds clusters of stories that discuss the same topic area or event [16]. Comparable to this, FSD evaluates the corpus and finds stories that are discussing a new event. FSD is a more specialized version of Topic Detection, because in Topic Detection the system has to determine when a new topic is being discussed and the resulting stories will be the “first-stories”.

In 2005, Mei and Zhai [3] discovered evolutionary theme patterns from text information collected over time. Temporal Text Mining (TTM) has many applications in multiple domains, such as summarizing events in news articles and revealing research trends in scientific literature. In this paper, TTM task is discovering and summarizing the evolutionary patterns of themes in a text stream. They define this new text mining problem and present general probabilistic methods for solving this problem through (1) discovering latent themes from text; (2) constructing an evolution graph of themes; (3) analyzing life cycles of themes.

A. Differences from Landmark Paper Mining

We now show that existing related techniques, specifically first story detection, hot topic mining and theme mining do not effectively handle the landmark paper mining problem. Our approach is simpler and more direct. We cannot reduce our requirements to the first story detection, hot topic mining and theme mining effectively.

In first story detection (FSD) [1], algorithms look for keyword in a first news story and compare the story with earlier stories. FSD is the process to find all stories within a corpus of text data that are the first stories describing a certain event [15]. The FSD process runs sequentially looking at a time-stamped stream of stories and making the decision based on a comparison of key terms to previous stories. FSD is closely linked to the topic detection task [16], a process that builds clusters of stories that discuss the same topic area or event.

Landmark paper mining differs significantly from FSD in the following ways. In FSD, a new story is detected as being a first story if it has a significant vocabulary shift from recent papers. First, a vocabulary shift could occur even without the introduction of new key terms if the frequencies of existing key terms are significantly altered. Second, a document can be flagged as a first story, even when there is an earlier document with the same key terms and frequencies. For example, even if there was an earthquake last year, the first story describing a more recent earthquake will be detected as a first story.

In hot topic mining [17], a topic is known as hot when it receives an unusually high amount of coverage in the news on a given day because of the importance of the events involving that topic. They used trend analysis in order to find hot topics, except that they are using controlled vocabulary terms rather than phrases extracted from text. Landmark paper mining is clearly a different problem as it
seeks to mine interesting papers, instead of interesting topics.

Finally, the Temporal Text Mining (TTM) [3] task is to discover, extract and summarize the evolutionary patterns of themes in a text stream. In this paper, the authors identify when a theme starts, reaches its peak, and then deteriorates, as well as which subsequent themes it influences. A timeline based theme structure is a very informative summary of the event, which also facilitates navigation through themes.

Theme Mining can be considered as an approach to mine interesting papers that originate themes. A new theme containing only existing keyphrases with altered frequencies does not have necessarily represented a new concept. However, in landmark paper mining, we follow a simpler and more direct approach. We identify papers that originate important keyphrases instead of themes (which can contain a collection of keyphrases). Our approach is simpler because it avoids the notion of themes - so there is no need to decide which collection of keyphrases form a theme. By avoiding this unnecessary step, our approach is more direct.

III. PROBLEM DEFINITION

In this paper we focus on the problem of finding documents (research papers) from a text corpus based on user specified keyphrases and also identifying the first document from the corpus where important keyphrases are introduced for the first time. We present the problem of mining landmark papers. This problem requires simultaneously understanding what keyphrases/topics are new or important and which documents drive these keyphrases. The following definition formally captures the problem statement:

Landmark Paper Mining: Given a collection of time indexed documents \( C = \{d_1, d_2, ..., d_r\} \), where \( d_i \) refers to a document with stamp \( i \), each document is a sequence of words from a vocabulary set \( V = \{w_1, w_2, ..., w_L\} \), the problem is to identify the first document that introduces important keyphrases for the first time known as landmark papers.

This can be broken into two sub-problems:

1) **Find the right key phrases / topics in a collection of documents.**

2) **Identify the originating documents of important key phrases or which documents introduced new keyphrases that had large impact?**

Here, we used the notion of keyphrases instead of keywords to avoid meaningless terms. A keyphrase consist of the set of keywords. In our experiments, we set a minimum length of keyphrase is 2 and maximum is 3.

IV. METHODOLOGY

The methodology we describe is a general approach that can be applied to text corpuses of varying complexity. The results of the mining are a set of landmark documents that match a query supplied by the user. Our methodology has the following major steps:

- **a)** Extract \( n \) keyphrases from each full text document by using Keyphrase Extraction Algorithm (KEA) [18]. In our experiments we set \( n = 10 \).

- **b)** Find the number of occurrences of each keyphrase within a document and within the corpus.

- **c)** For each keyphrase sort the documents containing it in increasing order of time-stamp, e.g. conference year.

- **d)** Remove keyphrases that are not contained in at least \( minsup \) documents. In our experiments, we set \( minsup = 3 \).

- **e)** In sorted list of documents, the first document for a keyphrase is identified as a candidate landmark paper corresponding to that keyphrase.

   The above parameter \( minsup \) ensures that the keyphrases considered are persistent and thereby important. In addition to the above steps, we have the following additional pruning step to refine the results.

- **f)** Remove keyphrases that are contained in the References section of their corresponding candidate landmark papers.

V. EXPERIMENTAL EVALUATION

To evaluate the performance of our algorithm, we prepared a database of well tagged "Data Mining/Databases" research papers from DBLP website. DBLP (Digital Bibliography and Library Project) is a computer science bibliography website hosted at University of Trier in Germany. The web-site maintains information regarding research papers in various fields of computer science. The web-site currently has more than \( \gamma^{20} \) research papers indexed. We experimentally evaluated our approach on a large data set of research papers in the databases and data mining areas. The research papers are classified on their topic, their conference, year of publication and their authors. We have extracted the information related to data mining/databases conferences like sigmod, vldb etc. and store the results in our database to perform the experiments. The information we extracted include the year of publication, authors, conference, paper title, general paper topic and we also got the link from where the user can download the full-text pdf files of research papers. We have converted full-text pdf files into text files by the pdf2text software in Linux. The basis statistics of the data sets are shown in table I. We intentionally did not perform stemming or stop word pruning in order to test the robustness of our algorithm.

Methodology described in section IV is performed on whole dataset. After performing these steps on whole dataset, we showed our results by using occurrence of keyphrases in the references of this document, implies the keyphrase introduced earlier and corresponding document will not be a landmark paper. If it is not present any of the references implies keyphrase introduce first time in that document and recommend that document as a landmark paper.
The results are shown in table II, where first column represents keyphrases, second column represents corresponding document from sorted list of documents and third column represents landmark paper by reference checking and says Yes, if the keyphrase is not in the reference of document else No, if it is present in the reference.

<table>
<thead>
<tr>
<th>Keyphrases</th>
<th>First document from sorted list</th>
<th>Landmark paper by references checking</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Management</td>
<td>05_sigmoid_16_Paper Title</td>
<td>No</td>
</tr>
<tr>
<td>RFID data</td>
<td>05_vldb_97_Paper Title</td>
<td>Yes</td>
</tr>
<tr>
<td>Load shedding</td>
<td>05_vldb_122_Paper Title</td>
<td>No</td>
</tr>
<tr>
<td>Data publishing</td>
<td>05_sigmoid_82_Paper Title</td>
<td>Yes</td>
</tr>
<tr>
<td>Schema mapping</td>
<td>06_vldb_111_Paper Title</td>
<td>No</td>
</tr>
<tr>
<td>BPEL specification</td>
<td>06_vldb_31_Paper Title</td>
<td>No</td>
</tr>
<tr>
<td>Access control</td>
<td>05_sigmoid_27_Paper Title</td>
<td>No</td>
</tr>
<tr>
<td>Twig query</td>
<td>05_vldb_32_Paper Title</td>
<td>No</td>
</tr>
<tr>
<td>Partial answers</td>
<td>06_vldb_20_Paper Title</td>
<td>Yes</td>
</tr>
<tr>
<td>Graph database</td>
<td>05_sigmoid_53_Paper Title</td>
<td>Yes</td>
</tr>
<tr>
<td>Road network</td>
<td>05_vldb_73_Paper Title</td>
<td>No</td>
</tr>
<tr>
<td>Data clustering</td>
<td>05_kdd_73_Paper Title</td>
<td>No</td>
</tr>
<tr>
<td>Quasi-identifier attributes</td>
<td>05_sigmoid_112_Paper Title</td>
<td>Yes</td>
</tr>
</tbody>
</table>

We validated our results by checking manually in the documents. We took a random sample of approximate 50 keyphrases and checked manually in corresponding documents whether these keyphrases introduced first time or not. If keyphrase introduced first time in a document and have some meaningful significance and also not present in references, then we are keeping those documents as a landmark papers. Out of this 50 random sample, we took 13 random sample of keyphrases and corresponding documents as a representative to show our results. In table III, there are 3 columns, where first column represents keyphrases, second column represents corresponding document from sorted list of documents and third column represents landmark papers by manual checking and says Yes, if a keyphrase is present first time in a document also not in the references of and No, if it is not.

<table>
<thead>
<tr>
<th>Keyphrases</th>
<th>First document from sorted list</th>
<th>Landmark paper by Manual checking</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Management</td>
<td>05_sigmoid_16_Paper Title</td>
<td>No</td>
</tr>
<tr>
<td>RFID data</td>
<td>05_vldb_97_Paper Title</td>
<td>No</td>
</tr>
<tr>
<td>Load shedding</td>
<td>05_vldb_122_Paper Title</td>
<td>Yes</td>
</tr>
<tr>
<td>Data publishing</td>
<td>05_sigmoid_82_Paper Title</td>
<td>Yes</td>
</tr>
<tr>
<td>Schema mapping</td>
<td>06_vldb_111_Paper Title</td>
<td>No</td>
</tr>
<tr>
<td>BPEL specification</td>
<td>06_vldb_31_Paper Title</td>
<td>Yes</td>
</tr>
<tr>
<td>Access control</td>
<td>05_sigmoid_27_Paper Title</td>
<td>No</td>
</tr>
<tr>
<td>Twig query</td>
<td>05_vldb_32_Paper Title</td>
<td>Yes</td>
</tr>
<tr>
<td>Partial answers</td>
<td>06_vldb_20_Paper Title</td>
<td>Yes</td>
</tr>
<tr>
<td>Graph database</td>
<td>05_sigmoid_53_Paper Title</td>
<td>Yes</td>
</tr>
<tr>
<td>Road network</td>
<td>05_vldb_73_Paper Title</td>
<td>No</td>
</tr>
<tr>
<td>Data clustering</td>
<td>05_kdd_73_Paper Title</td>
<td>No</td>
</tr>
<tr>
<td>Quasi-identifier attributes</td>
<td>05_sigmoid_112_Paper Title</td>
<td>Yes</td>
</tr>
</tbody>
</table>

In table V, we have combined the results from both table II and III. In this table, first and second columns indicate keyphrases and corresponding first document from sorted list of the documents, and, third and fourth column represents results from references checking and manual checking and finally last column represents landmark papers and show Yes, if a document is satisfying above criteria, else leave “-“ in that column. Also, corresponding keyphrase is important and new to the user.

The physical significance of first document for eg: "05_vldb_97_Paper Title" from sorted list, implies that this is the first document where keyphrase "RFID data" introduced first time. The output of our algorithm is not limited and it will be more accurate if we will run our algorithm on large datasets. Also, for the keyphrase "twig query" we have seen that it occurs only on vldb 2005 and 2006 year research papers not in the other conferences and in a given time stamp in our database. This implies that the keyphrase is not used by after this conference and years. So, it will be helpful for new researchers or user, to choose this keyphrase as an important keyphrase to continue his work in this area. Also, some of the keyphrases occur in recent years for eg: 2007, not before in our database, give intuition to user or researcher that this is the important or new keyphrase which introduced very recently in the
coming years. For keyphrase "graph database" we analyze that this keyphrase occurred in our whole database very frequently implies the keyphrase has equal influence throughout the time stamp window and it may occurs in coming years implies that the keyphrase is neither new nor important for a new user. In addition to this, for the keyphrases which are new and important for a user, we are also providing first document to the user, from where it originates. This will provide help to the new researcher or user to read the documents to get the overview of the keyphrases from where it introduced.

To check the accuracy of our results, we use common measurements in information retrieval that are recall and precision. Precision (P) is the percentage of retrieved documents that are in fact relevant to the query (i.e., “correct” responses), and Recall (R), is the percentage of documents that are relevant to the query and were, in fact, retrieved.

\[
\text{precision} = \frac{tp}{tp + fp}
\]  
(1)

And

\[
\text{recall} = \frac{tp}{tp + fn}
\]  
(2)

Where, tp represents true positive, tn is true negative, fp is false positive and fn is false negative respectively.

An information retrieval system often needs to trade off recall for precision or vice-versa. One commonly used trade-off is the F-score, which is defined as the harmonic mean of recall and precision:

\[
F - \text{score} = \frac{2 * R * P}{P + R}
\]  
(3)

We also show the confusion matrix, accuracy rate and error rate for our experimental results. The confusion matrix is a useful tool for analyzing how well our predicted results can match to actual results. A confusion matrix for predicted class and actual class is shown in table V.

**TABLE IV. A CONFUSION MATRIX FOR POSITIVE & NEGATIVE TUPLES**

<table>
<thead>
<tr>
<th>Actual Results</th>
<th>Predicted Results</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5 (tp)</td>
</tr>
</tbody>
</table>

From table IV, calculated precision is 0.714, recall is 0.625 and F-score is 0.667. We also estimated accuracy and error rate for our experiments. The accuracy of any system is the percentage of test set tuples that are correctly classified and error rate identify how many are misclassified. The accuracy and error rate is given by:

\[
\text{Accuracy} = \frac{tp + tn}{tp + tn + fp + fn}
\]  
(4)

And,

\[
\text{error-rate} = 1 - \text{Accuracy}
\]  
(5)

In our experiments, calculated accuracy is 0.62 and error rate is 0.38. This implies that our algorithm is reasonably accurate. On the other hand it is easy, direct and simple to understand.

VI. CONCLUSION AND FUTURE WORK

In this paper we have identified those documents within a large set of data that discuss a keyphrase that has not already been reported in earlier documents. In this approach, algorithms look for keyphrases in a corpus of research papers and compare the keyphrase with earlier papers. For a given keyphrase, we have identified a document from the text documents collected over time with the condition is, that keyphrase should occur first time at this paper and also have some meaningful definition in this paper. In addition of this, we also identified the presence of these keyphrases in the references (e.g. Title, Conference name etc.) of the document, if it is not present any of the references of a particular document means it is not introduced before this document and before this time stamp. So, that keyphrase is new to that document and introduced first time at this document with some meaningful definition, will known as a landmark paper. Our approach combines simple ideas from text mining, information extraction and retrieval. In future, we are trying to run our algorithm on large database of conferences and time stamp. In addition to this, we can extend this work to find the landmark keyphrases by using evolutionary graph of important keyphrases and also we can extend this work to find the impact of the document based on frequent keyphrases.

REFERENCES


TABLE V. LIST OF PAPERS BY CHECKING REFERENCES

<table>
<thead>
<tr>
<th>Keyphrases</th>
<th>First document from sorted list</th>
<th>Landmark paper by references checking</th>
<th>Landmark paper by manual checking</th>
<th>Landmark paper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Management</td>
<td>05_sigmod_16_Paper Title</td>
<td>No</td>
<td>No</td>
<td>-</td>
</tr>
<tr>
<td>RFID data</td>
<td>05_vldb_97_Paper Title</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Load shedding</td>
<td>05_vldb_122_Paper Title</td>
<td>No</td>
<td>Yes</td>
<td>-</td>
</tr>
<tr>
<td>Data publishing</td>
<td>05_sigmod_82_Paper Title</td>
<td>Yes</td>
<td>Yes</td>
<td>-</td>
</tr>
<tr>
<td>Schema mapping</td>
<td>06_vldb_111_Paper Title</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>BPEL specification</td>
<td>06_vldb_31_Paper Title</td>
<td>No</td>
<td>Yes</td>
<td>-</td>
</tr>
<tr>
<td>Access control</td>
<td>05_sigmod_27_Paper Title</td>
<td>No</td>
<td>No</td>
<td>-</td>
</tr>
<tr>
<td>Twig query</td>
<td>05_vldb_32_Paper Title</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Partial answers</td>
<td>06_vldb_20_Paper Title</td>
<td>Yes</td>
<td>No</td>
<td>-</td>
</tr>
<tr>
<td>Graph database</td>
<td>05_sigmod_53_Paper Title</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Road network</td>
<td>05_vldb_73_Paper Title</td>
<td>Yes</td>
<td>No</td>
<td>-</td>
</tr>
<tr>
<td>Data clustering</td>
<td>05_kdd_73_Paper Title</td>
<td>No</td>
<td>No</td>
<td>-</td>
</tr>
<tr>
<td>Quasi-identifier attributes</td>
<td>05_sigmod_112_Paper Title</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>