International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 03 | Mar 2021 www.irjet.net p-ISSN: 2395-0072

Sequential Pattern Miner to Identify the Research Area Shift of Researchers

Baby Manjusha P[1], Shibili T[2], Sreeja M K[3]

[1]*Head of the Department, Dept. of Computer Hardware and Maintenance,, Govt. Polytechnic College, Nedumkandam, Kerala, India*
[2]*Assistant professor, Dept. of Computer science, College of Engineering Vadakara, Kerala, India*
[3]*Lecturer, Sree Rama Govt Polytechnic College, Tripayar, Kerala, India*

Abstract - *For a new researcher, the knowledge of the current research trends is very important. Bibliographic databases are organized as the digital collection of academic publications. It provides a rich source of information about the publications such as title, author, book title, conference name/journal name, pages, year, abstract etc. But it is still hard for new researchers to discover hot or emergent topics and to understand the progression of the research disciplines. Bibliographic databases act as an aid for discovering the academic communities for acquiring a perception of the research community. Moreover, this database holds a vast amount of sequence data that shows emerging patterns. Discovering the vital predictabilities in these data are still challenging. In this paper, a new approach is proposed that explores the pertinence of the sequential pattern mining approach for identifying the research area shift of the researchers.*

***Key Words*: Sequential pattern, parser, research area shift**

1. INTRODUCTION

Bibliographic databases provide rich information about the publications such as title, author, book title, conference name/journal name, pages, year, abstract etc. As the size of bibliographic databases is growing, extracting information from them is a significant issue because much of the unseen information is implied within associations among entities in the data. The most useful fields in a bibliographic database entry are ee, or a reference to the electronic edition of the publication, title and author. A title is a semantic component that serves as the basic building block for identifying a research area in the corresponding discipline. The research area is a vital component for differentiating research communities. The metadata of the publications is used to take out keywords and terms, which can be used for building a taxonomy of topics. The ACM Computing Classification System, which is readily available can be used for building the taxonomy of topics in Computer Science. DBLP database also contains the table of contents of some of the conference proceedings, which includes session titles that could also be taken as topics.

Sequential pattern mining algorithms determine frequent subsequence as patterns in a sequence database. This is one of the important data mining problem with wide-ranging applications like analysing web access patterns, customer purchase pattern, time-related or sequencing processes such as natural calamities, disease occurrences etc. The sequential pattern mining concept was first introduced by Srikant and Agarwal, grounded on the study of customer's buying pattern. Basically, there are two approaches for discovering sequential patterns: Candidate generation-and-test method such as SPADE which is a vertical format based approach; GSP, a horizontal format-based approach; and Sequential pattern growth approach such as FreeSpan, PrefixSpan etc and its extensions.

A sequence database is an ordered collection of sequences of events. A sequence database SD is a set of tuples <sid, S>; sid is the sequence identifier and S is a sequence. The tuple <sid, S> is said to contain a sequence 's', if 's' is a subsequence of S. A minimum support threshold for the sequence should be defined, which is the number of tuples in the database containing the sequence 's'. i.e., supportSD(s)={<sid,S>/ (<sid,S>∈ SD) ∧ (s ⊆ S)}. A sequence 'S' is said to be a sequence pattern in a sequential database if the support value of sequence S is greater than or equal to the given minimum support threshold i.e., supportSD(S) ≥ minimum support.

2. RELATED WORK

Zaiane et al. propose a model DBconnect that utilizes the relations within the DBLP database for discovering research communities and frequent associations [1]. Ichise et al. propose a technique to find research communities by featuring a network model of publications and a word assignment technique for identifying the communities [2]. Chetna et al. conducted a systematic survey on sequential pattern mining and categorize the algorithms into two classes. The first one is based on the algorithms for increasing the efficiency of mining and the second based on the different versions of sequential pattern mining designed for particular applications [5].

3. SYSTEM ARCHITECTURE

Figure 1depicts the proposed architecture of the system.DBLP database is used as the input to the system. The method proposed is described in four phases, viz., Parsing DBLP- XML Records, Keyword Generation, Identification of Research Topic and Research Area Shift Discovery. The first three phases perform pre-processing of data. In the next phase, a model of the sequential pattern miner is constructed to identify the research area shift of researchers. Before describing the proposed method, a snippet of the Bibliographic database is also shown.

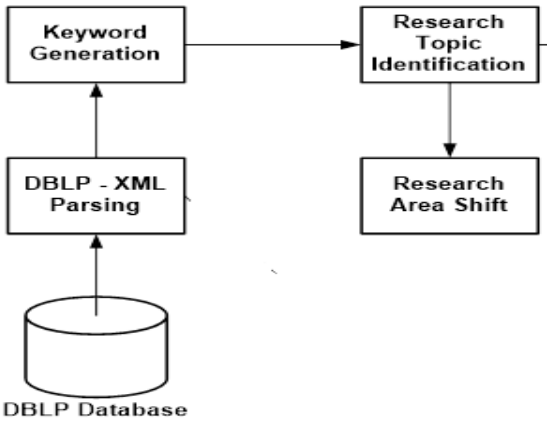

Figure 1- System Architecture

3.1 DBLP Database:

DBLP is a major Bibliographic data source available on the Internet for the Computer Science discipline. DBLP provides information on major research publications in the conference proceedings and computing journals. The DBLP data set can be downloaded from the location http://dblp.uni-trier.de/xml/.The whole dataset of DBLP is available as a large XML file. The file dblp.xml comprises all bibliographic records. Along with the xml file, the document type definition dblp.dtd is also there to validate the XML file. The main advantage of the DBLP dataset is its free accessibility and the inclusion of conference proceedings. But it has the disadvantage of lacking the citation information and the diverse coverage for different subfields in computer science.

The majority of the entries in the DBLP database are either conference or journal publications. There are eight types of entries in the DBLP dataset. Every entry comprises one or more of the following metadata fields: title, author, pages, editor, book_title, year, address, journal, number, volume, month, ee, URL, cdrom, publisher, cite, note, ISBN, cross_ref, school, series, chapter_no.

An extract of the DBLP file is given below.

```
<inproceedings mdate="2003-06-03"
key="conf/ai/RazekFK03">
<author>Mohammed Abdel Razek</author>
<author>Claude Frasson</author>
<author>Marc Kaltenbach</author>
<title>Re-using Web Information for Building Flexible
Domain Knowledge.</title>
<pages>563-567</pages>
<year>2003</year>
<crossref>conf/ai/2003</crossref>
<booktitle>Canadian Conference on AI</booktitle>
<ee>http://link.springer.de/link/service/series/0558/bibs/
2671/26710563.htm</ee>
<url>db/conf/ai/ai2003.html#RazekFK03</url>
</inproceedings>
```

The description of each field is given below.

"inproceedings": Conference proceedings
"mdate": Date of last modification of the record, which comply with ISO 8601 format.
"key: Unique key identifying the record.
"author": Name of the author
"title": Title of the paper
"pages" Range of page numbers
"crossref": Key of the proceedings record
"booktitle": Name of conferences or workshops
"ee", "url" : Location of the electronic edition of the publication if it is available. If the electronic copy is not available, this field holds the position of the Table of Contents of DBLP.

The subtree conf/* in the "key" namespace indicates that the paper is a conference paper or a workshop paper. A record can hold up to two URLs in the "ee" and "url" field. The DBLP dataset also comprehends the Table of Contents of the conference proceedings which can be utilized for finding research topics.

3.2 Parsing DBLP XML Records

Use an XML parser to parse all the records of the DBLP XML files to take out all the relevant information from them. The information such as title, authors, conference/journal name, location of the electronic edition, if available, and year of publication are extracted. A transaction database is maintained to store the extracted information. Every tuple in the database table is viewed as a transaction and the identified information as the items of transaction.

3.3 Keyword Generation

This phase identifies the keywords corresponding to each publication. For that, titles are tokenized and remove frequently used stop words such as for, towards, in, understanding, approaches etc. The next step is to transform the remaining tokens to their root form by the stemming process. These words are treated like keywords. Keywords can also be extracted using "ee" and url fields in the transaction database. Also, additional terms related to the published literature can be identified by crawling the URL of

the publication to retrieve keywords and abstracts from it. Additional keywords can be taken out from the extract utilizing any term extractor readily available. Figure 2 outlines the steps of extracting the keywords.

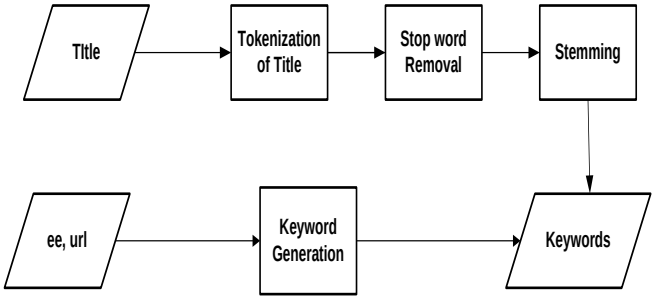

Figure 2- Flow diagram of keyword extraction

3.4 Identification of Research Topic

This phase identifies the research topic of each author in the transaction database. This can be achieved using the keywords of the publications of the author. A mapper function has to be defined for mapping the research areas of the Computer Science discipline with the extracted keywords of publications. The next step is to find the similarity score between the keywords identified from the transaction and the keywords defined in the mapper function. If the similarity score is above the defined threshold, the transaction is mapped to the corresponding research area and obtain the relationship between authors and the research area.

3.5 Research Area Shift Identification

This phase devises a method to identify authors Research area shift. The first step is to sort the transaction database year-wise. Using the Author, Research area information, construct the sequence database. The sequence database is given as input to the sequential pattern miner and derives the patterns containing the author and research area. The sequential patterns with support value greater than or equal to the given minimum support threshold represent the research area of interest shift of the author.

4. CONCLUSION

In this paper, a new method is proposed that explores the pertinency of the sequential pattern mining algorithms for identifying the research area shift of the researchers. The system can be implemented in four phases viz., Parsing DBLP- XML Records, Keyword Generation, Research Topic Identification and Research Area Shift Discovery. DBLP database is used as the input dataset to the proposed system. The output of the system will be sequential patterns that represent the research area shift of the potential researchers.

REFERENCES

[1] Osmar R. Zaiane Jiyang Chen Randy Goebel "DBconnect:mining research community on DBLP data "Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis Pages: 74-81 2007

[2] R. Ichise, H. Takeda, T. Muraki "Research Community Mining with Topic Identification" Proceedings of the Information Visualization (IV'06) 0-7695-2602-0/06 $20.00 © 2006 IEEE

[3] R. Agrawal, and R. Srikant, "Mining sequential patterns," The International Conference on Data Engineering, pp. 3–14, 1995

[4] B.Manjusha, Sumam M. "Chapter 9 Apriori-based Research Community Discovery in Bibliographic Database" , Springer Nature, 2011

[5] Chetna C., Amit T., Amit G., Sequential Pattern Mining: Survey and Current Research Challenges, International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-1, March 2012

[6] Jian Pei, Jiawei Han, Behzad Mortazavi-Asl, Jianyong Wang, Helen Pinto, Qiming Chen,Umeshwar Dayal,"Mining Sequential Patterns by Pattern-Growth:The PrefixSpan Approach"IEEE Transactions on Knowledge and Data Engineering, vol. 16, No. 10, october 2004