Abstract

A rapid emergence in big scholarly data and development in information technology has let the researchers to search preferred articles. This becomes a major challenging issue due to information overloaded problem. Further performing literature survey for scholarly articles has become a challenging task. Especially for junior researchers it becomes a time consuming task due to rapid growth in scientific articles. This requires intelligent algorithms for searching and navigations. Hence an automatic recommendation of cited articles has been provided. In existing approaches citation recommendation are provided based on full article or the citation context is taken as input. Researchers for having survey of papers quickly and to have a broad view of given topic a keyword based approach is used as the search query that provides important cited articles. Graph based article ranking are done based on vertex reinforced random walk (VRRW). Further hierarchical clustering is done to provide recommendation selection for three sets of user (user list).

I.

A recommendation engine, also known as a recommender system, is software that analyzes available data to make suggestions for something that a user might be interested in, such as a book, a video or a job, among other possibilities. An engine, in a software context, is a
special-purpose program that performs a task through a variable algorithm, often as a feature of some larger program. A search engine is one type of recommendation engine, responding to search queries with pages of results that are (at least theoretically) the search engine’s best suggestions for websites that satisfy the user’s query based on the search term plus other data, such as location and trending topics. Recommendation engines are common among e-commerce, social media and content-based websites.

Citation is a reference to a published or unpublished source. At the time of 20th century when researchers have to do their research work, document procurement process became a challenging work, and to find related articles were found very difficult from the stack of books. Later after the 20th century digital document library provided an ease of access in finding articles of their interest. Google scholar, IEEE explore are some examples. Even though this search engines provide relevant articles the appropriateness of the articles to the given query is not that much effective. Therefore, a recommendation system that leads the readers to the key literature in the areas of their interest has to be developed. Citations are important in academic dissemination. It helps in providing better articles to the researchers when compared with the other parameter recommendations like venue, author etc. The most common way of finding reference papers is to, Search documents on search engines such as Google and traces the cited references by starting with a small number of initial papers (seed-papers). Unfortunately, this would be very difficult in finding a comprehensive keyword list to cover all papers, especially for the researchers who are new to the field.

Hence a graph based recommendation system is provided using the citation network. They are simple and robust when compared to other recommendation systems. They can be used to recommend other items like academic collaborators, movies, online items.

Figure 1: Outline of citation recommendation system

II. LITERATURE SURVEY
Citation recommendation system: one can assort different ways in literatures based on several dimensions. The first way of is specification of query, i.e., the input query specified by the user. The other dimension is the way by which citations are used as query. Citations can be either global or local. A global citation recommendation provides a list of articles for
the given user query. Strohman et al. [2] combined text features to measure the relevance between articles. Gori and Pucci [3] used citation relations between articles to built the citation graph and applied random walk algorithm to compute the ranking scores of the targeted articles. on the other side local citation systems aim in providing citations based on the citation context. Lu et al. [5] aimed in recommending citations using translation model which is used in translating text in one language to other. Jevin et al. [1] proposed an recommendation system based on hierarchical clustering of article-level citation network using Eigen factor algorithm- a modified version of page ranking algorithms. Feng et al. [4] incorporated information on common author relations between articles and provided recommendations based on graph based article ranking. Input approaches like providing full articles as input [6], [7], [8] requires latent information about the articles. Random walk with restarts [9] and well-known PageRank [10] are used by PaperRank [11] and ArticleRank [12] approaches to rank the citations using the citation graph. Quan et al. [13] introduced an model using greedy clique expansion algorithm which is done to discover communities. At last page ranking algorithm were used to rank and recommend papers. Diversification of results of random walk has recently attracted recent researches. One of the earlier algorithms that address diversified graph ranking by vertex selection with absorbing random walks is GRASSHOPPER [14]. Finding out more complex and diversified set of citations were given by vertex reinforced random walk with restart which solves the disabilities of GRASSHOPPER.

In this paper we take the user query (title, abstract), extract the keywords, rank the articles formed from citation network and recommend papers for three sets of users. Using of citation context and other metadata of articles might me the time consuming task. Moreover graph based article ranking is a robust and generic systems that aims in recommending other items apart from papers such as music, movies..

III. PROBLEM STATEMENT

Given a heterogeneous bibliographic network, terms, authors, venues of a query, the system aim in building a recommendation model specifically for a given query q and recommend a small subset of papers for the researchers. This approach uses hierarchical clustering, for providing relevant and important papers for three sets of users.

IV. PROPOSED SYSTEM

Research papers are needed in citing relevant and important previous work. Hence there is a need in understanding its background, context and innovation. There is a large, rapidly growing body of scientific literature. Hence automatic recommendation of high quality citations should be provided for the researchers to integrate with their authoring flow. The proposed framework for citation recommendation system is shown in Figure. User when posting queries are processed using natural language processing techniques.

The input keywords are further extended that filter outs unwanted queries and appropriate query is chosen which are mapped in to publication database. Meanwhile citation networks are generated from publication database that results an induced sub graph after mapping the resultant collection of papers to citation networks. Ranking are performed
using a reinforced random walk algorithm. These ranked nodes are clustered and recommendations are provided for different users.

Here we first describe the key ingredients of our proposed citation recommendation system and then we present the algorithms.

A. Data preprocessing

Text mining is the process of seeking or extracting the useful information from the textual data. Our data is preprocessed with the help of NLP (Natural Language Processing). Natural Language Processing (NLP) is an area of research and application that explores how computers can be used to understand and manipulate natural language text.

1. Tokenization

User gives the query, a textual data from which keywords are to be extracted. Hence the given query is tokenized i.e. the process of breaking a stream of text in to words, phrases, symbols or other meaningful elements called tokens. The list of tokens becomes input for further processing such as parsing or text mining.

2. Pos tagging

The process of assigning one of the parts of speech to the given word is called Parts Of Speech. It is commonly referred to as POS tagging. Parts of speech include nouns, verbs, adverbs, adjectives, pronouns, conjunction and their sub-categories. We collect the noun and verb and do further processing.

3. Stop word removal

Stop words are words which are filtered out before or after processing of natural language data (text). Though stop words usually refer to the most common words in a language, there is no single universal list of stop words used by all natural language processing tools, and indeed not all tools even use such a list.

4. Stemming

In linguistic morphology and information retrieval, stemming is the processes of reducing inflected (or sometimes derived) words to their word stem, base or root form generally a written word form.
B. Mapping query with keyword communities

This work consists of recommending citations of given search query, which requires mapping query with the keyword cluster formed from the keyword-keyword network.

1. Formation of keyword network

Keywords are either extracted from publication database. Some data usually have the keywords which can be directly used to construct the keyword-keyword network. Keyword network is constructed which is an undirected and weighted graph where each node corresponds the keyword. Two nodes are connected by an edge if there is one article that contains both the keywords.

2. Formation of keyword communities

After keyword network construction the next step is clustering the keywords using Louvain community algorithm, a well known state-of-the-art-algorithm which is used to find the communities from the keyword network. This algorithm uses greedy optimization. This optimization is performed of two types, first the method looks for the smaller communities by optimizing modularity locally and aggregates the nodes which is belonging to same community and builds the network. Then the input query is mapped with keyword communities and constituent keywords from cluster are fetched to the next step of the framework.

C. paper-keyword association

After expanding the input query further, they are mapped with the publication database to get the reference articles. This is achieved by indexing. The purpose of storing an index is to optimize speed and performance in finding relevant documents for a search query. Finally collections of reference articles are collected.

D. Assembling citation networks

Citation network is a social network which contains paper sources and linked by citations relations. This is explained when a document di cites a document dj, this is shown by an arrow from the node representing di to the document representing dj. In this way the document from collection D form a graph, which is called citation graph or citation network. Based on the reference papers obtained from the paper-keyword association they are mapped with citation network which gives an induced graph.

E. Ranking the nodes and hierarchical clustering

The next step is ranking nodes according to reinforced random walk algorithm using div rank. This model automatically balances the prestige and the diversity of the top ranked vertices in a principled way.

1. Ranking nodes

Here we present the overview of random walk with restart for ranking articles.

Let us consider a graph G= (V, E) where v is the vertices and e is finite set of edges. The task of ranking the vertices can be casted as the prestige in a network by finding a suitable prestige function f: V → R+.Researchers have recently focusing on centrality measures based on stationary distribution of random walk in the network. A random walk is random process that describes a path that consists of a succession of random steps on some
mathematical space such as the integers. For example, the path traced by a molecule as it travels in a liquid or a gas

A random walk on a graph is a very special case of a Markov chain. Unlike a general Markov chain, random walk on a graph enjoys a property called *time symmetry* or *reversibility*. It can be either be directed graph or undirected network. Each vertex represents a state and a walk transmit from one state to another based on transition probability.

### 2. Vertex reinforced random walk with restart

In most cases transition probabilities do not change in time. They are time homogeneous. If one vertex is visited frequently by walk then all its neighbors are likely to be visited. This is called as smoothing process [15]. By this way top ranked prestigious articles are viewed. But for diversity in random walk is achieved through Vertex reinforced random walk with restart. It is a time-variant process that takes into account of both prestige and diversity. The probability of jumping from one node to the other is constant over the time. The transition probabilities at each time are influenced by the number of times each state has been visited and by a priori likelihood matrix, which is real, symmetric and nonnegative.

### 2. Hierarchical clustering

Hierarchical clustering involves clustering the nodes that uncover the boundaries between domains, fields, subfields, etc. The core idea is to compress a description of a random walk on the network. If the network has localized regions such that a random walker has a long persistence time in a small group of nodes, a random walk can be concisely encoded when this structure is exploited in the coding scheme. This involves an approach of Bottom-up hierarchical agglomerative clustering, i.e., Building a tree-based hierarchical taxonomy (dendrogram) from a set of documents. Based on these Div scores recommendations are provided to three types of users: researchers, students, serendipity.

### V. EXPERIMENTAL STUDY

For research purpose in citation recommendation most of the system is worked with smaller dataset. Our dataset consist of citation datasets from Microsoft Academic Search (MAS) which houses over 4.1 million publications and 2.7 million authors this dataset has more than 2 million distinct papers altogether comprising of computer science domain having bibliographic information having title of paper, unique index of publication, the venue, and year of publication. If keyword information is missed in the dataset they can be extracted by using keyword extraction methods. High energy physics dataset presents the papers information containing abstract, citation information. Since keyword is missed with the dataset, Keyword extraction is done by NLP techniques using java.

### VI. CONCLUSION

This paper deals with a simple citation-based method for recommending articles that is been considered for different types of users. This method is based on the hierarchical structure of scientific knowledge, allowing for different scales of influence. It is based on content based...
recommendation. Its goal is to help researchers in finding publications related to their research area. Given the constant success of machine learning methods in several other domains, it is very interesting to see that network based methods perform better on scholarly papers. Furthermore, graph-based citation recommendation provides a better ranking of articles to the users. User query is processed using NLP techniques. Queries are extended further and cited articles are ranked using VRRW. In future hybrid recommendations can be provided that would further enhance the recommendation systems.

VII. REFERENCES


TO CITE THIS PAPER