Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 89 (2016) 483 – 491

Procedia Computer Science

Twelfth International Multi-Conference on Information Processing-2016 (IMCIP-2016)

Recommender System for Academic Literature with Incremental Dataset

Mahak Dhanda and Vijay Verma*

National Institute of Technology Kurukshetra, Kurukshetra 136 119, India

Abstract

On account of the colossal expansion in the size of research paper repository, the stature of Recommender System has increased, as it can guide the researchers to find papers akin to them from this vast collection. Furthermore, the recommendation methods like collaborative-filtering or content-based do not allow the user's to provide their personalized requirements explicitly; hence the focus is shifted towards the customized Recommender Systems that can scrutinize user's preferences by contemplating their inputs. But the state-of-art recommendation techniques satisfying user's personalized requirements make a strong assumption of static dataset. So, in this work we are going to present a customized Recommender System that can acknowledge the ever growing nature of research paper repository. To accomplish this, the Efficient Incremental High-Utility Itemset Mining algorithm (EIHI), which has been recently introduced in the literature, is used which is specialized to work with dynamic datasets. Experimental results prove that the proposed system satisfies the researcher's personalized requirements and at the same time handles the incremental nature of the research paper repository efficiently.

Keywords: High-Utility Itemset Mining (HUIM); Recommender Systems (RS); Utility-based Recommendation.

1. Introduction

The massive hike in the extent of information over the web has turned it arduous for the users to search for the information admissible to them. Recommender System (RS) has come out as a revolutionary concept to ride out through this situation[14]. It is a tool (software) that provides the users with the suggestions of information that may be useful to them. These suggestions may turn out helpful to the users in many scenarios where decision making is involved ex. selecting books to read, movies to watch etc. A lot of techniques are available for recommendation which are majorly categorized as collaborative filtering[7] and content-based filtering[5]. Collaborative filtering works on the concept of finding out similar users so as to make recommendations, while content-based techniques work on the basis of similarity in features of the item and the user.

RSs have gained a lot of emphasis in the commercial environment[9] and along with it they have proven to have a crucial role in academic literature domain[11] also. As a result of continuous growth in the size of research paper repository (because of lot of papers coming out of journals and conferences each year), it has turned out to be

*Corresponding author. Tel.: +91-8901370565.
E-mail address: mahak0570@gmail.com

doi:10.1016/j.procs.2016.06.109

troublesome task for the researchers to get papers akin to them. To endure this problem, recommender systems have become necessary in the academic literature domain[11]. Also, as different researchers may have their own personal requirements; customized recommender systems are drawing researcher's attention[3], where users can provide their specific preferences before the recommendation is provided to them. But still the major complication resides due to the dynamic nature of research paper database.

The recommendation technique[3] considering user's preferences, make a stronger assumption about the nature of database to be static. Here, through this work we are going to propose a new customized RS for research paper which can incorporate the dynamic aspect of the dataset. The size of research paper repository may change because of new papers getting added to it time to time and this yearns for a recommender system that can support the dynamic dataset. So, the EIHI algorithm[1] is utilized in our proposed approach, as it is compatible with datasets which are dynamic in nature.

The proposed technique works in two stages 1) first stage is used to filter out the papers pertinent to the topic of interest of the user based on their content; 2) second stage makes sure that the personalized requirements of the user are satisfied (on the basis of usability assigned to the papers with the help of input taken from the user).

The rest of the article is organized as: Section 2 covers Preliminaries for the approach; Section 3 describes the Related work, Section 4 and 5 covers the Proposed approach and Simulation results respectively; and Conclusion is added as Section 6.

2. Preliminaries

EIHI[1] is a High Utility Itemset Mining (HUIM) algorithm, designed to work in case of the incremental datasets. This technique can perform well in case of such dataset because whenever updation is made in the dataset, it can compute the new High Utility Itemsets (HUIs)[2] only by scanning new transactions and previous HUIs (without starting the process again from scratch).

EIHI works by calculating Transactional weighted Utility (TWU)[2] and making utility lists and Estimated Utility Co-occurrence Structure (EUCS)[13] only for the items present in newly added transactions and also by exploring the extensions of only those items[1]. As EIHI is tree-based algorithm it becomes very easy to add new HUIs to the tree as well as to make updations to utilities of existing HUIs[1].

As the academic literature repository is incremental (as new papers may get added to the repository from time to time) and it's not possible to scan the complete repository again and again. Hence we take the advantage of concepts used in EIHI to handle the dynamic nature of this repository. Our proposed approach uses EIHI to mine the High Utility Reference-sets (HURs)[3] and recommend them to the users.

3. Related Work

Recommender Systems have become an important part of academic literature domain because of the ever growing size of research paper repository[16]. Research paper recommender systems are mainly influenced either by the frequency of citations of a paper or by the pertinence of its contents to the user, while making recommendations[11].

The approaches based on frequency of citation may face *cold start* problem and hence recommender system based on Belief-Propagation has been proposed[10]. In order to further improve the quality of recommendation some approaches using semantic data for recommendation have been created[15]. Research paper recommender system supporting diversity has also been developed[6] which uses the concepts of co-authors and dissimilar users to make recommendations.

Also as different researchers may have their own personalized requirements in terms of publishing dates etc., recommender systems based on user's recent interests are gaining importance. To find user's interests, recommender systems may either use the author's published work[4] or can be customized to take input from the user before recommending papers[3]. But the recommendation technique proposed in[3] suffers from dearth of efficiency because it relies on Two-phase algorithm[2] which does not contemplate the incremental nature of research paper repository. To overcome this limitation we are proposing EIHI based recommendation approach having support for dynamic datasets.

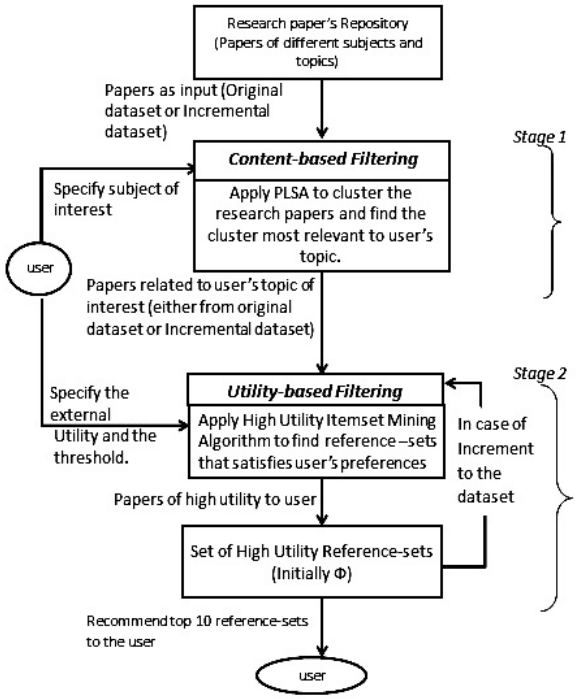

Fig. 1. Architecture of the System.

4. Proposed Approach

Our developed approach is customized research paper recommendation approach having compatibility for incremental research paper repository; that works using incremental high-utility itemset mining technique EIHI.

The complete process of recommendation is performed by the approach in two stages 1) choosing the papers pertinent to researcher's topic of interest; 2) recommending the researchers with the papers having higher usefulness to them.

To make it work efficiently even when the number of research papers in the repository can increase (as new papers may get added to the repository from time to time), we have used EIHI to mine HURs (as EIHI can handle dynamic datasets).

4.1 Architecture of the system

The basic architecture of the system is as shown in Fig. 1.

4.2 Working

The proposed approach works as a two-stage method:

Stage I. Finding papers pertinent to user on the basis of their contents:

In this step, the research papers present in the dataset (repository) are segregated into 'k' clusters using the PLSA algorithm. Different topic is accredited to each of the cluster implicitly and then according to the word distributions of papers their probability of getting fit into a particular cluster is found out. A paper is allocated to the cluster with which its clubbing probability is greatest.

After the clusters are made, proximity of each cluster to topic of interest of the user is estimated with the help of the similarity measures. The cluster which has the highest affinity towards user's topic of interest is selected and is worked upon further.

Whenever the increment is made to the repository by adding some new papers, the PLSA algorithm is run again in order to assign the newly added papers with their respective clusters (with which their association probability is highest). And only the cluster whose relevance to user's topic is highest is fed to the next step.

Stage II. Finding Reference-sets of high utility to the user on the basis of user's personalized requirements:

To contemplate user's personalized preferences (ex. authority, publishing date etc.) the proposed system utilize the HUIM technique EIHI[1] which generates as output the reference-sets that best fits user's requirements. Originally, the HUIM algorithms are designed to find out those itemsets whose utility satisfies the minimum threshold. But in academic domain, utility signifies degree with which a paper is preferred by the researcher. We may consider date of publishing or publishing authority etc. as basis for finding utility of a reference-set.

EIHI when applied in this domain has to work papers rather than the transactions. Also items are supplanted by references and hence in our approach, EIHI is applied to reference-sets instead of itemsets. Here, the *internal utility(i)* of a reference can take value that are either 1 or 0 denoting if a reference is being cited by a paper or not. The *external utility(e)* is provided by researcher based on his preferences (which in the presented approach is publishing date).

The *utility of a reference (r) in any paper (P)* is $u(r, P) = i * e$ where i and e are the internal and external utilities respectively and the *utility of a reference-set (R)* is the sum total of utilities of all the references present in that set. Θ defines the utility threshold specified by the user and a reference-set's utility should be either greater than or equal to Θ in order to get a place in the recommendation list. Among all the reference-sets whose utility satisfies a minimum utility condition, we choose the top 10 reference-sets and then recommend those to the researcher (user).

In case more papers are added to the repository, EIHI is implemented again on the newly added papers to get the new set of HURs (as some new HURs may appear and the utilities of previous HURs may get updated). EIFI is capable to deal with dynamic datasets as it can discover the new HURs only by considering the newly added papers along with previously found HURs (as compared to the static dataset approaches that have to find HURs from scratch in case of any increment to dataset).

4.3 Example

Take the dataset in Table 1 as toy example. Here, the columns contain the references as r1, r2, r3 and so on, used in research papers; whereas, rows contain the research papers, themselves. The entries in the cells denote the number of

Table 1. Research paper dataset (original dataset D contains papers P1–P10 and the updated dataset D' contains P1–P12.

ID of Paper	r1	r2	r3	r4
P1	1	1	0	0
P2	0	1	0	0
P3	1	0	1	0
P4	0	1	1	1
P5	1	1	0	0
P6	0	0	0	0
P7	1	1	1	0
P8	0	1	1	1
P9	1	0	0	0
P10	0	1	0	0
	1	1	1	0
	0	0	1	1

Table 2. Weightage table (shows weight assigned to each reference).

Reference	Weight Assigned
r1	1
r2	4
r3	3
r4	2

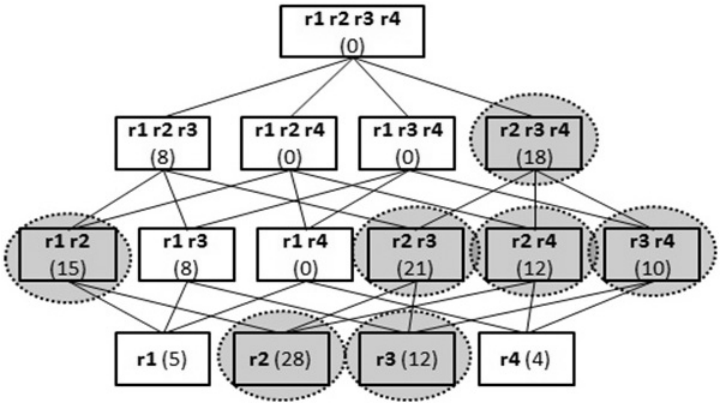

Fig. 2. All Possible Reference-Sets along with Utility Values (for original dataset (D)).

times citation is given to a reference (corresponding to column) in a given paper (corresponding to row). Here we have considered ten research papers using any number of references from the set of given four references. Table 2, is used to represent the weight/utility value accredited to each of the reference by the user.

Here utilities of different reference-sets in original dataset D are as follows:

u({r2}, P2) is 4, u({r2}) is 28, and u({r1, r3}) = u({r1, r3}, P3) + u({r1, r3}, P7) = 4 + 4 = 08. If the utility threshold Θ is taken as 10, then the reference-set {r1, r3} will not be in the recommendation list. But if the threshold is taken as 8, {r1, r3} will be in the recommendation list as it is high utility reference-set (HUR). Now consider the system that recommends reference-sets on the basis of frequency of citation and take minimum threshold as 2 then r4 will be there in the recommendation list but if we consider the utility based recommendations r4 is not HUR as u({r4}) = 4 which is less than the utility threshold i.e. 10. Hence, r4 is not present in the recommendation list which proves that r4 does not satisfy personalized requirements of the user.

The search space (the possible reference-sets) for the toy example is shown in Fig. 2. The utility value of reference-sets is stated as numeric content. If Θ (minimum utility threshold) is presumed to be 10 then only the reference-sets with utility value $\geq \Theta$ will be HURs and hence will be recommended to the user.

Here {r2}, {r3}, {r1 r2}, {r2 r3}, {r2 r4}, {r3 r4}, {r2 r3 r4} are the reference-sets with utility $\geq \Theta$ and hence are defined as HURs (in dark solid circles in Fig. 2.). So these appear in the recommendation list of the user. The High Utility Reference-sets (HURs) are found out by using EIHI algorithm.

Now consider the updation in dataset (as the dataset is incremental and more papers can be added). Let the dataset D becomes D' by inserting Paper P11 and P12 into it (shown in Table 1 as solid dark rows). Along with the dataset, the set of HURs has to be updated (some new HURs may be added and the utility values of older HURs may get changed). To find out the new Reference-sets to be recommended again EIFI is to be applied.

EIFI works on the newly added papers along with the previously found HURs to create new set of references having high utility, as described in Fig. 3. (HURs being shown as solid dark circles). In this case some new HURs are discovered ex. {r1 r2 r3}, {r1 r3} and also the utilities of previously found HURs gets increased ex. for {r2}, {r3}, {r1 r2}, {r2 r3}, {r3 r4} utilities have been updated as 32, 18, 20, 28 and 15 respectively.

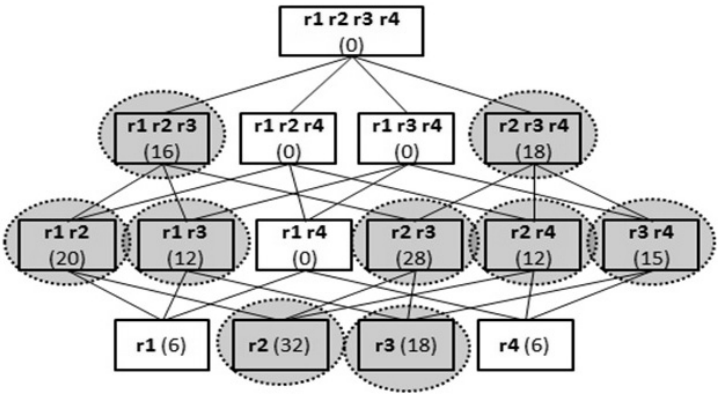

Fig. 3. All Possible Reference-Sets along with Utility Values (for updated dataset (D')).

5. Experimental Evaluation

5.1 Data set

In order to perform evaluation of our **Recommender System for Academic Literature with Incremental Dataset,** we have used ACL Anthology Network[4] (a real-world dataset). Research papers are accumulated by ACL from various venues and are represented in the form of citation network. Because of the consistency of the topic of dataset (on computational linguistics), we will directly implement the second step to mine HURs from the research paper repository. There is no requirement of clustering to find the research papers akin to the user. We are taking into consideration the collection of research papers with publishing date between 1965 and 2007 originally and for making increments into the data we are considering papers from 2008–2013.

The "publishing date" is considered as the factor to denote user's preferences. The weightage will be provided to the references on the basis of their publishing dates explicitly by the user. By using this weight the research papers of higher usability to the users are selected and hence recommended to the user.

Pre-processing of the raw data is carried out. Originally, ACL repository contains various research papers along with the citation network of those papers. To implement HUIM techniques we are required to have a network of "transaction-itemset" type, hence each paper Pi in the citation network is treated as a transaction with its references denoting the items in transaction. To accomplish this IDs (unique) have to be accredited to the papers and result represents the papers with these IDs only.

6. Results

Various experiments have been performed to check the capability of our system. The elicitation of each experiment is provided as:

1) Recommendation by considering citation frequency as basis, where the weight assigned to the papers on the basis of their publishing date is same.
2) Recommendation by considering utility of reference-set as basis, using EIHI algorithm. Here the weights of references with publishing date between 2005–2007, 2000–2004, 1990–1999 and 1965–1989 are 20, 10, 5 and 1 respectively.

   We have taken into consideration the original dataset (i.e. papers from 1965–2007) for experiment 1and 2.
3) Comparing the EIHI-based approach with Two-Phase based approach when several increments are there in the dataset (here we consider papers from 2008–2013 to make increments to original dataset). Here we make 5 increments in the original dataset. In each increment several new papers are added to the dataset. Here the weights

Table 3. Top 10 Recommended Reference-Sets based on the Citation Frequency.

Rank of Reference-set	Reference-sets by using citation frequency as base for recommendation
1.	1953 1902
2.	1950 1953
3.	**1869 13**
4.	1950 1902
5.	1950 1953 1902
6.	1869 1950
7.	1869 1902
8.	**1950 13**
9.	1869 1953
10.	**1902 13**

Table 4. Top 10 Recommended Reference-sets on the basis of Utility of Reference-Sets using EIHI.

Rank of Reference-set	Reference-sets by using citation frequency as base for recommendation (using EIHI).
1.	1953 1902
2.	1950 1953
3.	**2480 1953**
4.	**2598 1953**
5.	**2480 1869**
6.	1950 1902
7.	**2598 1950**
8.	1950 1953 1902
9.	1869 1950
10.	1869 1902

are provided as 20, 10, 5, 3 and 1 to papers published in 2009–2013, 2005–2008, 2000–2004, 1990–1999 and 1965–1989 respectively.

In experiment 2 we have used Θ as 2000. And for experiment 3 Θ is 1200. User can choose any random value for this threshold depending on his requirements.

Table 3 represents the reference-sets being recommended by experiment 1 i.e. by considering the frequency of citation as the basis for making recommendations.

Table 4 represents the reference-sets obtained after performing experiment 2. Here filtering of the reference-sets on the basis of their utility (using the EIHI algorithm). The variations between the results of experiment 1 and 2 (i.e. the recommendations on the basis of citation frequency and recommendation on the basis of utility) are shown as bold letters.

Here the reference-sets {2480, 1953}, {2598, 1953}, {2480, 1869}, {2598, 1950} are discovered by the EIHI based filtering but not by the filtering on the basis of citation frequency. Whereas the reference-sets {1869, 13}, {1950, 13}, {1902, 13} are discovered by the citation frequency based filtering but not by EIHI based filtering. The references 2480 and 2598 are published in year 2007 and 2005 respectively which have been provided with higher weightage by the user (because of their recentness). Whereas the reference 13 is published in year 1993 so less weightage is given to it by the user (because it is not recent). Hence the reference-sets including 13 are not high utility reference-sets. This verifies that our proposed approach brings the HURs into the recommendation list of the user and at the same time eliminates the low utility reference-sets from recommendation list.

The EIHI based recommendation approach performs better than already existing Two-phase based approach[3] in case of incremental datasets because it is able to generate the new HURs only by considering new papers being added and the previous HURs, it does not require scanning the complete dataset again. Whereas in case of the Two-Phase

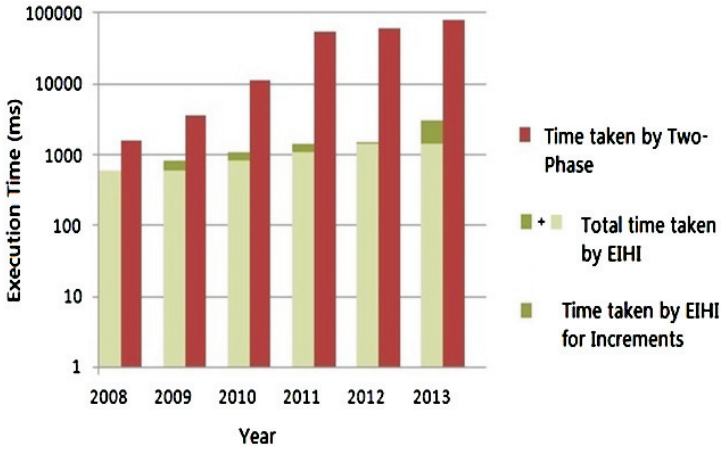

Fig. 4. Execution Results.

algorithm if any increment is made in the dataset the complete procedure for generating HURs has to be started from scratch.

Figure 4 shows the graph denoting execution times of the EIHI-based and the Two-phase based recommendation approaches, when the increments are being made to the system. Here, firstly the research papers from 1965–2007 are given as input to the EIHI and Two-Phase based system as year 2008 version; then as the database is incremental some new papers are being added each year and hence new versions are created (2009, 2010 etc.).

In case of EIHI-based system (our proposed approach) whenever new papers are added to the dataset only the increments (i.e. the new papers added) have to be provided as input to the system not the complete dataset and hence for 2009 release only the papers present in 2009 version but not in 2008 one, are taken as input for the system and so on for 2010, 2011, 2012 and 2013 version. Whereas, in case of Two-phase based approach, when a new version is created, the complete version is to be fed as input to the system. Hence each year when new papers are added to the dataset, the complete dataset is added as input to the system.

The red and green bars in the graph shows the execution times of Two-Phase based approach and EIHI-based approach respectively. The darkened portion in the green bar represents time taken by EIHI for the increments. While, the complete green bar (lighter portion + darkened portion) represents the total time taken by EIHI-based approach. The graph verifies that EIHI based approach is always faster than the Two-phase based approach in case the updations (increments) are made to the dataset. Hence our proposed approach can work well even with dynamic dataset.

References

[1] Fournier-Viger and Philippe, Efficient Incremental High Utility Itemset Mining, In *Proceedings of the ASE BigData & SocialInformatics*, pp. 53, (2015).
[2] Y. Liu, W. Liao and A. Choudhary, A Fast High Utility Itemsets Mining Algorithm, In *Proceedings of the 1st International Workshop on Utility-based Data Mining*, pp. 90–99, (2005).
[3] Shenshen Liang, Ying Liu, Liheng Jian and Yang Gao, A Utility-based Recommendation Approach for Academic Literatures, In *ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, pp. 229–232, (2011).
[4] Sugiyama, Kazunari and Min-Yen Kan, Scholarly Paper Recommendation Via user's Recent Research Interests, In *Proceedings of the 10th Annual Joint Conference on Digital Libraries*, pp. 29–38, (2010).
[5] de Gemmis and Marco, Semantics-Aware Content-Based Recommender Systems, In *Recommender Systems Handbook*, pp. 119–159, (2015).
[6] Sugiyama, Kazunari and Min-Yen Kan, Serendipitous Recommendation for Scholarly Papers Considering Relations among Researchers, In *Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries*, pp. 307–310, (2011).
[7] Koren, Yehuda and Robert Bell, Advances in Collaborative Filtering, In *Recommender Systems Handbook*, pp. 77–118, (2015).
[8] Paraschiv and Ionut Cristian, A Paper Recommendation System with ReaderBench: The Graphical Visualization of Semantically Related Papers and Concepts, In *State-of-the-Art and Future Directions of Smart Learning*, pp. 445–451, (2016).
[9] Nguyen and Loc, Introduction to a Framework of E-commercial Recommendation Algorithms, In *American Journal of Computer Science and Information Engineering 2.4*, pp. 33–44, (2015).

[10] Ha and Jiwoon, Recommendation of Newly Published Research Papers using Belief Propagation, In *Proceedings of the 2014 Conference on Research in Adaptive and Convergent Systems*, pp. 77–81, (2014).
[11] Beel and Joeran, Introducing Docear's Research Paper Recommender System, In *Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital libraries*, pp. 459–460, (2013).
[12] http://clair.si.umich.edu/clair/anthology/.
[13] Fournier-Viger and Philippe, FHM: Faster High-utility Itemset Mining using Estimated Utility Co-occurrence Pruning, In *Foundations of Intelligent Systems*, pp. 83–92, (2014).
[14] Ricci, Francesco, Lior Rokach and Bracha Shapira, Recommender Systems: Introduction and Challenges, In *Introduction to Recommender Systems Handbook*, Springer US, pp. 1–35, (2011).
[15] Paraschiv and Ionut Cristian, A Paper Recommendation System with ReaderBench: The Graphical Visualization of Semantically Related Papers and Concepts, In *State-of-the-Art and Future Directions of Smart Learning*, pp. 445–451, (2016).
[16] Beel and Joeran, Research-paper Recommender Systems: A Literature Survey, In *International Journal on Digital Libraries*, pp. 1–34, (2015).