Intelligent Information Management, 2016, 8, 27-40
Published Online March 2016 in SciRes. http://www.scirp.org/journal/iim
http://dx.doi.org/10.4236/iim.2016.82004

Incorporating User's Preferences into Scholarly Publications Recommendation

Tobore Igbe, Bolanle Ojokoh

Department of Computer Science, Federal University of Technology, Akure, Nigeria
Email: tobore2ng@gmail.com, itobore@futa.edu.ng, bolanleojokoh@yahoo.com, baojokoh@futa.edu.ng

Received 3 March 2016; accepted 25 March 2016; published 29 March 2016

Abstract

Over the years, there has been increasing growth in academic digital libraries. It has therefore become overwhelming for researchers to determine important research materials. In most existing research works that consider scholarly paper recommendation, the researcher's preference is left out. In this paper, therefore, Frequent Pattern (FP) Growth Algorithm is employed on potential papers generated from the researcher's preferences to create a list of ranked papers based on citation features. The purpose is to provide a recommender system that is user oriented. A walk through algorithm is implemented to generate all possible frequent patterns from the FP-tree after which an output of ordered recommended papers combining subjective and objective factors of the researchers is produced. Experimental results with a scholarly paper recommendation dataset show that the proposed method is very promising, as it outperforms recommendation baselines as measured with nDCG and MRR.

Keywords

Personalization, Digital Library, Information Retrieval, Recommender System, Citation Analysis, User Preferences

1. Introduction

With the increasing growth in scientific publications, access to qualitative documents becomes more difficult. Search engines are faced with the challenge of returning good ranking results with best in quality documents first, as most users rarely go beyond the first couple of pages. In addition, it is also necessary to produce results that are relevant to the specific need of each user early enough in the search process. Recommender systems have emerged as important response to these challenges and they have been deployed successfully in the digital library domain [1]-[3]. Earlier approaches focused on counting the number of publications [4]. After these, more

How to cite this paper: Igbe, T. and Ojokoh, B. (2016) Incorporating User's Preferences into Scholarly Publications Recommendation. *Intelligent Information Management*, **8**, 27-40. http://dx.doi.org/10.4236/iim.2016.82004

complicated computations related to citations came up, leading to a concept referred to as citation analysis. Citation analysis essentially involves inspecting the number of times a scientific paper or scientist is cited in publications, and it works on the assumption that influential scientists and important works will be cited more frequently than others [5]. Researchers and administrators in many academic institutions worldwide also make use of citation data for hiring, promotion, and tenure decisions among others [6]. Citation and citing data could also provide researchers and administrators with a reliable and efficient indicator for assessing the research performance of authors, projects, programs, institutions, and countries and the relative impact and quality of their work [7] [8].

A proper and efficient citation analysis can uncover important research documents to help researchers appreciate previous works and formulate better research ideas. However, in spite of the many benefits of improving the quality of search through citation analysis, it is important to find relevant publications applicable to every researcher's desired area. Several researches [9]-[14] have been conducted to provide improved search results to satisfy researchers' requirements. For instance, an analytic model called Human-Recommender Interaction (HRI) was developed in [10] to create a deeper understanding of users for a descriptive recommender system. Sugiyama and Kan [15] developed a framework for modeling researchers' past works in recommending potential papers, confirming the fact that the interest of a researcher can be linked to his previous works. Nevertheless, most of these studies have provided recommendation to publications in the context of available information present in the publications, while neglecting the need to carefully consider researchers' preference in the choice of recommendation.

In this paper, a combination of objective features obtained from the papers and subjective features obtained from the researcher is proposed. Objective features, according to this work refer to metrics that are factual and quantifiable. These could be measured from the research publications. They include a number of references, a number of authors, and a number of citations among others. All these create significant value for the paper. The subjective features are personal, individual preferences that could be specified by the user in the search process, for example, author name and keywords in abstract and title. These determine personal influence on the choice of paper recommendation.

Our proposal is that not only the features used to find good publications should be considered, but that there should be the inclusion of personal preferences. For instance, an individual may want publications with a specific keyword or phrase in the abstract or title; another consideration may be range of year of publication (for example, 2000 to 2014). Due to differences in individual choices, there is the need to satisfy that while recommending potential papers. In this work, we therefore propose the integration of specific user's preferences to paper recommendation in order to provide personalized search. We apply these preferences to find potential citation papers. Objective feature score is used to scale papers and the use of mean value rate eliminates weaker papers. The use of frequent pattern growth algorithm based on subjective features on potential papers obtained from objective feature analysis, ranks papers based on the preferences of the individual. Our proposed method is entirely unsupervised and can be applied to any collection of publications where recommendation is necessary. We apply our method on a scholarly paper recommendation dataset and the result is remarkable. Recommending papers with both objective and subjective features improve accuracy as evaluated by both mean reciprocal rank (MRR) and normalized discounted cumulative gain (nDCG).

The rest of the paper is organized as follows: Section 2 discusses related works. In Section 3, the proposed method is presented. The experiments and results are discussed in Section 4. Finally, conclusion and direction to extend the work are discussed in Section 5.

2. Related Works

Recommendation system is becoming a useful tool for researchers and as such there are different viewpoints by different authors for paper recommendation. A lot of research works have been carried out in the scholarly paper recommendation domain. In order to make correct recommendation, some works have proposed performance measurement. Erasmus [16] stated some factors and performance metrics that can be used to grade research article based on the degree of citation. These performance indicators include: h-index, an objective measure for the relative scientific value of an author. The index is named after Hirsch who proved mathematically that with a simple procedure, a value for this measure can be calculated. The procedure arranges all articles according to decreasing number of referrals to each article, in such a way that the most cited article is on top.

Another indicator used was Impact Factor of a journal for a certain year, which is calculated from the number of citations to that journal in the preceding two years, divided by the number of articles that were published during the same two years. The more citations to a journal, the higher the impact factor, and the more important the journal should be. The third indicator used was the cited half-life of a journal, which is a measure for the time that articles in that journal are cited. If there are more fundamental articles in a journal, the longer those articles will be cited, and the higher is the cited half-life. The cited half-life of that journal is the median of the age of the articles that are cited in the previous year, half the referrals is to articles older than that time moment; the other half of the referrals is to newer articles. The last indicator used was the immediacy index of a journal, which indicates how quick articles in that journal are cited. A journal with many "hot item" articles published in the first half of the year will show a high immediacy index. These indicators were successfully used to ranked journals but the consideration were only features of the journals without taking into account the subjective factors of the researcher within his or her discipline [17].

Page *et al*. [18] presented Page Rank algorithm that simulates a user navigating the Web at random, by choosing between jumping to a random page with a certain probability (referred to as the damping factor d), and following a random hyperlink. While this algorithm has been most famously applied to improve ranking of Web search results, it has also been applied to the digital library field in two ways: 1) in improving the ranking of search results; and 2) in measuring the importance of scholarly papers. There was no consideration for subjective features, which would have yielded a high-quality personalized recommendation for users.

Nascimento *et al*. [10] presented a source independent framework for research paper recommendation, a framework that requires as input only a single research paper and generates several potential queries by using terms in that paper, which are then submitted to existing Web information sources that hold research papers. Once a set of candidate papers for recommendation is generated, the framework applies content-based recommending algorithms to rank the candidates in order to recommend the ones most related to the input paper. This is done by using only publicly available metadata (that is, title and abstract). The success of this recommendation system would have been enriched, if the researcher's preferences were considered.

Jarvelin and Kekalainen [9] presented Context-aware Citation Recommendation system. In the context-aware system, a user can submit either a manuscript (that is, a global context and a set of outlink local contexts) or a few sentences (that is, an outlink local context) as the query to the system. There are two types of citation recommendation tasks, which happen in different application scenarios. *Global Recommendation*: given a query manuscript *d* without a bibliography, a global recommendation is a ranked list of citations in a corpus *D* that are recommended as candidates for the bibliography of *d*. They noted the fact that different citation contexts in *d* may express different information needs. The bibliography candidates provided by a global recommendation should collectively satisfy the citation information needs of all outlink local contexts in the query manuscript *d*. *Local Recommendation*: given an outlink local context (c) with respect to *d*, a local recommendation is a ranked list of citations in a corpus *D* that are recommended as candidates for the placeholder associated with c. For local recommendations, the query manuscript *d* is an optional input and it is not required to already contain a representative bibliography. In this system, the user subjective factors were not considered as part of the criteria for recommendation.

Sugiyama and Kan [15] presented a system for recommendation system to help generate relevant suggestions for researchers based on exploiting potential citation papers in scholarly paper recommendation. Potential citation papers were identified through the use of collaborative filtering. Different logical sections of a paper were given different significance; the system further investigated which sections of papers can represent a paper effectively. Further experiments on a scholarly paper recommendation dataset showed that proper modeling of potential citation papers as well as properly representing papers with both their full text and assigning more weight to the conclusion-improved recommendation accuracy significantly as judged by both mean reciprocal rank (MRR) and normalised discounted cumulative gain (nDCG). The success of this system was impressive; however a high-quality personalized recommendation for researchers could have been achieved if their personalized attributes were considered in the recommendation.

3. Proposed Method

It is imperative to have a system capable of recommending appropriate papers for researchers. This will enhance research outputs and reduce the burden involved in exploring numerous papers. In order to achieve a suitable system, it is necessary to consider objective and subjective features that could influence recommendation.

Figure 1 gives a diagrammatic description of our proposed system. Candidate and relevant papers recommendation for researcher is achieved by considering both objective and subjective features. Potential publications consist of a database of publications from different authors, publication venues and years of publication. Objective feature score is computed for all potential publications. Candidate publications are papers that fulfill the minimum objective rate value and they form collections from which papers are recommended. Researchers can specify a set of personal preferences, which serves as the subjective features criteria for recommending papers from candidate papers. An iterative walk through process is performed to arrive at potential papers for recommendation. The result of the recommendation process shows recommended papers ranked based on best match.

In this work, we adopt the Frequent Pattern (FP) Growth Algorithm, which is an efficient and scalable method for mining a complete set of frequent patterns from data, and creating an extended prefix-tree structure called frequent-pattern tree (FP-tree) [19], for storing and compressing the database on frequent patterns. It has been proved that this method outperforms other popular methods for mining frequent patterns, such as the Apriori Algorithm [14] [20] [21] and the Tree Projection [22]-[24]. In some later works, it was proved that FP-Growth has better performance than other methods [12] [25].

The integration module uses FP-growth algorithm for potential paper recommendation. It consists of two major steps. First is the computation of FP-tree which is a condensed representation of the dataset. The dataset consist of papers that satisfy objective feature score. The patterns used are the subjective features, which are used to group the papers into category based on the occurrence of the pattern. A minimum support count (search frequency) indicated as part of the subjective feature is used to create a frequency table from the dataset. The papers are prioritized in ascending order in a frequency-table based on frequency occurrence of the patterns found in potential papers. The FP-tree is constructed from the frequency-table in relation to the dataset of potential paper; this is done so that the tree can be processed quickly.

Secondly, a walk-through of the created FP-tree to generate an output of ordered recommended papers based on objective and subjective factors for the researcher is generated. The following subsections describe the outlined procedures.

3.1. Objective Features

For paper P_i contained in the dataset of publications W^{paper}, listed below are the objective factors, which form the objective feature set (F_j^{obj}) where $j = 1, 2, \cdots, 8$. The basis for these feature sets is to create an objective feature score for potential papers for recommendation which is considered in the subjective feature processing.

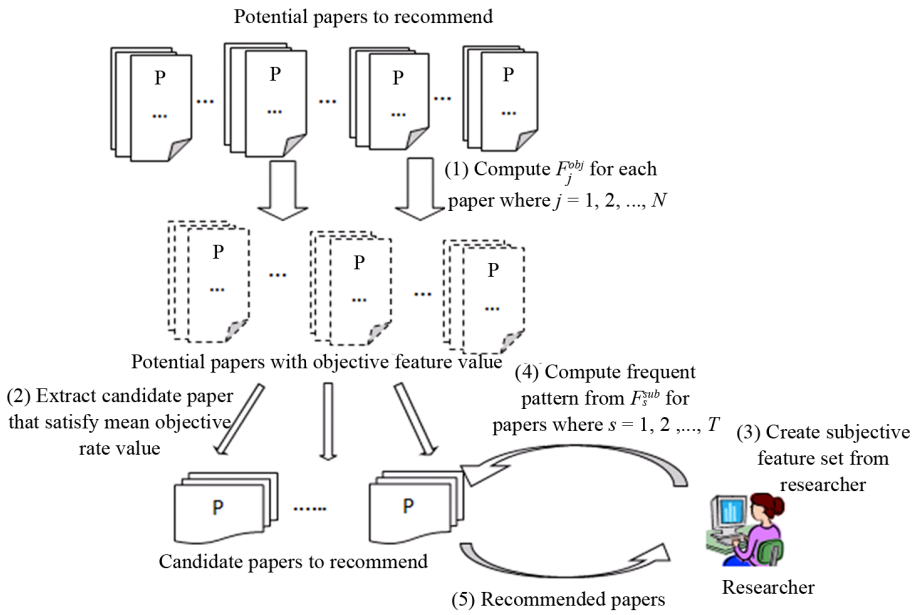

Figure 1. Proposed system description.

1) Number of other papers that cite paper P_i [Ncp]. The count of the number of other papers that acknowledged P_i. High value of Ncp shows that P_i is valuable in content and acknowledged by other researchers.

2) Number of reference papers in P_i [Nrp]. This count value refers to the number of other papers that P_i acknowledged. The number of papers reviewed and considered for research indicates the depth of knowledge paper P_i possesses.

3) Number of authors [Na]. The number of person(s) who wrote the paper. The contributions made by more authors in a research produce a better publication because the knowledge from the authors is harnessed into the research. A research carried out by two or more persons tend to be more resourceful than a research carried out by one person.

4) Summation of the total number of publications for each author(s) in P_i [NPa]. The higher the number of NPa shows the wealth of experience of the authors. It reveals the participation of the author(s) in other publications.

5) Homogenous venue score [Hmp]. It is the rate paper P_i is cited with respect to the average number of citation of papers in the same publication venue of P_i. This reveals how much the paper P_i is acknowledged within the same publication venue.

6) Heterogeneous venue score [Htp]. It is the summation of the rate paper P_i is cited with respect to the average number of citation of papers in other publication venues where P_i was cited. This reveals how much the paper P_i is acknowledged from other publication venues.

7) Year of publication of paper P_i [NPy]. This objective feature assigns a higher weight value to more recently published papers. It is computed by dividing the difference between the year P_i is published and the least year (that is, the oldest) of available paper published by the difference between current year and the least year of available paper published.

8) Mean objective rate value [Rm]. The average value of P_i from the cumulative objective score determines if P_i will be considered for recommendation. The lower the value of R_m, the more the papers will be considered and the higher the value the less the papers to be considered. A value of 0.00, indicate that all potential papers that satisfy the subjective features should be considered.

For the fifth objective feature, we have:

$$Pub_H = \frac{\sum_i^{N_{pub}} H_{cit}\left(p_i\right)}{N_{pub}} \tag{1}$$

$$F_5^{obj} = \left[Hmp\right] = \frac{C_m}{Pub_H} \tag{2}$$

where C_m is the number of times P_i is cited within the same publication venue. N_{pub} is the total number of publications and Pub_H is the average number of citation for papers in the same publication venue as P_i, and for the sixth, we have:

$$Pub_j = \frac{\sum_i^{N_{pub}} H_{cit}\left(p_i\left(j\right)\right)}{N\left(j\right)_{pub}} \tag{3}$$

$$F_6^{obj} = \left[Htp\right] = \sum_j^{V_{pi}} \frac{C_j}{Pub_j} \tag{4}$$

Equation (4) computes the heterogeneous venues core for paper P_i; C_j is the number of times P_i is cited in j publication venue. V_{pi} is the number of other venues where paper P_i is cited apart from the venue where it is published. Pub_j is the average number of citation for papers in venue j and $N\left(j\right)_{pub}$ is the total number of papers in venue j. In order to assign a higher weight value to more recently published papers, we propose the following equation for the seventh feature:

$$F_7^{obj} = \left[NPy\right] = \frac{Y_{pi} - Y_{p\min}}{Y_{curr} - Y_{p\min}} \times 100 \tag{5}$$

Equation (5) computes the objective feature value [NPy], where Y_{pi} is the year Pi was published; $Y_{p\min}$ is

the least year a paper was published, and Y_{curr} is the current year.

The cumulative sum of all the objective features for each paper in W^{paper} is obtained in Equation (6). $P[t]$ is an associative array that maps each paper in W^{paper} to its objective feature score.

$$P[t] = W^{\text{paper}} \xrightarrow{\text{calculate}} \sum_{j=1}^{6} F_j^{obj} \tag{6}$$

$$F_8^{obj} = D^{\text{paper}} = \begin{cases} \dfrac{P[t]_i}{\sum P[t]} \geq R_m P_i \\ \text{elsenull} \end{cases} \tag{7}$$

D^{paper} are papers that meet the minimum average objective score requirement, which are considered for recommendation and ranked based on subjective feature.

3.2. Subjective Features

For user u, the following are the subjective factors, which form the subjective feature set (F_s^{sub}) where s = 1, 2, 3, 4, 5.

1) Keyword in title [Kt]. Researcher specified keyword is examined if it is present in the title of the paper. If found, then, it is a potential paper for recommendation.

2) Keyword in abstract [Ka]. Inspection of Researcher's keyword (or phrase) in the abstract section of the publication. If present, it is considered as a potential paper.

3) Author name specification [SNa]. The choice of a particular author can be specified as part of the search criteria, to indicate the preference of the researcher.

4) Publication years to be considered [Py]. The range of years for publications can be specified to indicate the preference of publications that will be displayed.

5) Publication venue to be considered [Pv]. The researcher may be interested in publications from one or more publication venues (such as IEEE, ECCC) based on experience, service and other personal preferences.

6) Minimum support [Ms]. This value is used to compute feature occurrence for each paper that satisfies each subjective feature 1, 2, 3, 4 and 5 above. This value is used to eliminate weaker paper considered under subjective features. The higher the minimum support the lower the number of publications that will be considered for processing and recommendation.

The subjective function returns a list of papers that matches each subjective feature set from the potential papers.

The result produced is described in **Figure 2**, where each row shows the paper(s) where the corresponding subjective feature was found. It shows that a paper, P_i can appear in multiple rows. The sixth feature is the minimum frequency occurrence expected from papers in **Figure 2**. Its value is rounded up to the nearest whole number and it is used as the frequency criteria to determine papers that will be listed. The frequent pattern is created from D^{paper} for each subjective feature set.

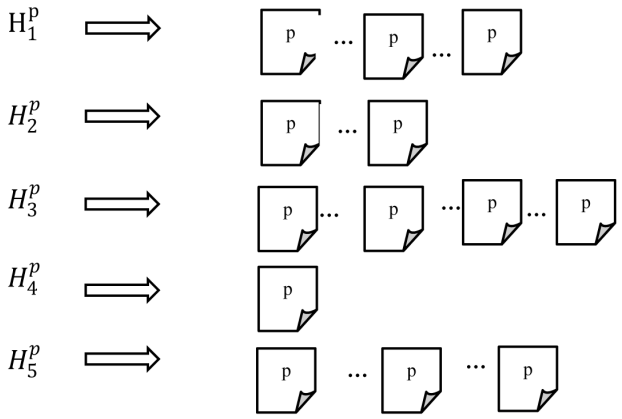

Figure 2. Frequent pattern of candidate papers based on occurrence of subjective features.

3.3. Leveraging Objective and Subjective Features

In recommending personalised potential papers, the objective and subjective feature set form the criteria used for recommendation. The objective features select the possible papers from the dataset based on mean objective value. An initial table is constructed holding papers that meet each subjective feature value as a row in the table, as shown in **Figure 2**.

A distinct list of papers is constructed from the aggregated papers that are found in the subjective feature set H_w^p. Each paper in the feature set is counted to determine the number of occurrence, if the frequency satisfies minimum support value (Ms), it is added to the list. Potential papers that did not meet the Ms value or already exist in the list are not considered. A connection tree is constructed between the papers in the list based on similar Ms value. A list of recommended papers is generated in ascending order with respect to subjective feature set of the researcher from the tree by recursively traversing the constructed tree of papers.

4. Experiments

Dataset is collected from DBLP version of arnetminer containing 2,084,055 papers with 2,244,018 number of citation relationships, published until 2013. The papers in the dataset are obtained from different publication venues. The dataset is organized into groups of published papers in different venue; in each group, each line starts with a specific prefix indicating an attribute of the paper. The DBLP version is organized into 1,511,035 groups. The size of the dataset is 1.9 GB. The data set can be used for clustering, publication reference information, studying influence in the citation network, finding the most influential papers, topic modeling analysis among others [26]-[29]. Processing the dataset, 19 distinct files of 50 MB were created from the initial dataset file. The dataset contains papers of different topics in Computer Science such as: database management system, programming language, security, artificial intelligence, data processing, computer graphics, and software engineering.

For our implementation, we computed the objective feature values for all papers in the dataset and the summation of the objective feature set is computed. The papers that match [Rm] objective value are extracted for subjective feature processing.

A set of subjective feature attributes were constructed to form the subjective feature set. These attributes were used to obtain the possible papers for recommendation. For each of the distinct file containing the dataset, a thread was constructed to search through the content for records that match the feature set. The purpose of using threads was to speed up the time required to perform the search.

We implemented the algorithm in Java, and the output (process statistics, citation analysis, performance analysis, and ranked publications) are stored in MySQL. All the experiments were conducted on windows operating system.

4.1. Experimental Results

Table 1 shows a sample subjective feature with corresponding values that forms preferences for recommendation of publications. From the table, authors considered (SNa), range of publication year (Py) and publication venue (Pv) such as Institute of Electrical Electronics Engineering (IEEE) and Electronic Colloquium on Computational Complexity (ECCC) conference are indicated. The keywords to search for in the title (Kt) and abstract (Ka) are also shown in the table. **Table 2** shows some possible potential papers and their corresponding objective feature values. The potential papers are the ones that meet the minimum mean objective rate value. The experiment was carried out with values of [Rm] = 0.2, 0.4, 0.6, 0.8 and 1.0. The minimum support is varied between 10 and 100 for each value of [Rm].

Recommended papers are shown in **Table 3**. The table shows sample of ranked recommended papers for a sample objective and subjective feature set from **Table 1**, with the following description: year of publication, title, authors, venue of publication and year of publication. The number of papers returned is determined by the minimum support specified by the user. **Figures 3-7** show the ranked papers for different values of minimum support. It is observed that the number of papers recommended is indirectly proportional to the value of support specified. More highly ranked papers are recommended for [Ms] values less than 60. The number of papers recommended reduces as [Rm] value increases.

4.2 Evaluation Measures

How well a recommendation system is capable of effectively positioning relevant citations defines the efficiency

Table 1. Subjective feature set and corresponding values.

SNa	Py	Pv	Kt	Ka
Zhang, Paolo, Chu, Krebs, Chan, Wang, Aranda, Li	2000 to 2010	IEEE, ECCC	mining, rules, fuzzy rule	mining, rules, fuzzy rule
Stonebraker, Hull, Hopcroft, Nierstrasz, Tarlton	1985 to 2000	INGRES Papers, OOC Databases and Application	relational database, system, database, interface	relational database, system, database, interface
Füzesi, Nakamura, Dholakia, Robichaud	1995 to 2005	ICC, IEEE	digital, network, digital subscriber lines, packet, data packet	digital, network, digital subscriber lines, packet, data packet
Leichter, Highnam, Ikedo, Shi	1998 to 2005	IWDM	Digital Mammography, Mammography, Breast x-ray	Digital Mammography, Mammography, Breast x-ray

Table 2. Potential papers with objective feature set and values.

Potential Papers	Nrp	Ncp	Na	NPa	Hmp	Htp	Npy
Paper 1	18	7	4	82	1.11	0.00	71.42
Paper 2	22	5	3	61	4.22	0.00	52.12
Paper 3	28	6	4	58	2.50	0.21	82.04
Paper 4	16	9	2	72	2.12	1.64	78.16
Paper 5	15	5	2	43	3.35	1.43	68.11
Paper 6	18	4	6	88	2.11	1.26	53.62
Paper 7	18	2	7	137	1.11	0.12	60.04
Paper 8	16	5	2	66	5.52	1.37	70.48
Paper 9	19	4	4	82	4.10	2.40	44.19
Paper 10	19	5	5	79	1.26	1.21	52.43
Paper 11	22	6	5	81	1.24	2.15	55.45
Paper 12	27	11	2	84	0.34	0.000	41.35
Paper 13	24	12	6	89	1.03	2.22	70.48
Paper 14	20	5	6	90	1.21	2.52	55.43

Table 3. Sample of ranked recommended papers.

TITLE	Authors	Venue	Year
Extraction of If-Then Rules from Trained Neural Network and Its Application to Earthquake Prediction	Yue Liu, Bofeng Zhang, Gengfeng Wu	IEEE	2004
Release Planning under Fuzzy Effort Constraints	AnNgo-The, GüntherRuhe, WeiShen	IEEE	2004
Mining Fuzzy Rules in A Donor Database for Direct Marketing by a Charitable Organization	Keith C. C. Chan, Wai-Ho Au, Berry Choi	IEEE	2002
Agent Paradigm in Clinical Large-Scale Data Mining Environment	AbdelazizOuali, Z. Ramdane-Cherif, Amar Ramdane-Cherif, NicoleLévy, Marie-Odile Krebs	IEEE	2003
A novel fuzzy neural network: the vague neural network	RuiFang, YibiaoZhao, Wei-Sheng Li	IEEE	2005
Quasi-Morphism and Comprehensibility of Rules in Inductive Learning	Wiphada Wettayaprasit, Chidchanok Lursinsap, Chee-Hung Henry Chu	IEEE	2002
An Evolutive Algorithm for the Data Mining in Population Data Warehouse	Wenchuan Yang, PengWang, Chunyang Gao, Yanyang Fan, Huahua Luan	IEEE	2006
Studies on Fuzzy Information Measures	Shifei Ding, Zhong zhi Shi, Fengxiang Jin	IEEE	2006
Cognitive-Based Rules as a Means to Select Suitable Groupware Tools	Gabriela N. Aranda, Aurora Vizcaíno, Alejandra Cechich, Mario Piattini, Jose Jesus Castro-Schez	IEEE	2006
Decomposition and Hierarchical Process for Fuzzy Cognitive Maps of Complex Systems	Zhang Guiyun, Yang Bingru, Zhang Weijuan	IEEE	2006
A Logical Framework for Fuzzy Collaborative Filtering	Stefano Aguzzoli, Paolo Avesani, Brunella Gerla	IEEE	2001
Logical Foundations for Constraints on Fuzzy Sets in Soft Computing: MV-Partitions and Refinement	Paolo Amato, Corrado Manara, Domenico Porto	IEEE	2001

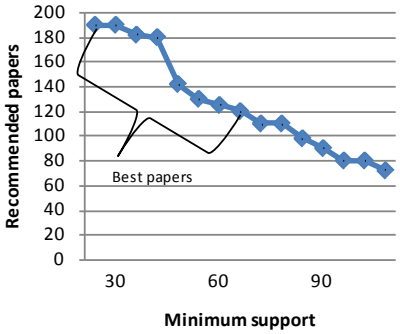

Figure 3. Recommendation papers for different [Ms] values and [Rm] of 0.2.

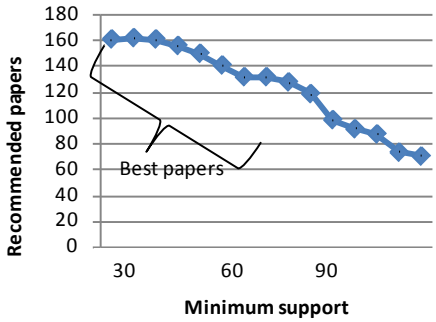

Figure 4. Recommendation papers for different [Ms] values and [Rm] of 0.4.

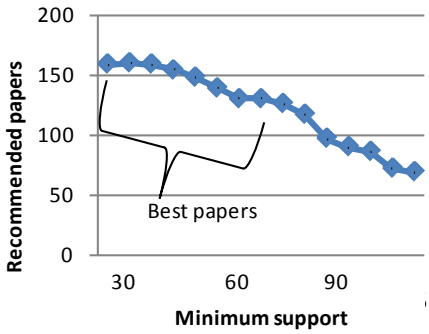

Figure 5. Recommendation papers for different [Ms] values and [Rm] of 0.6.

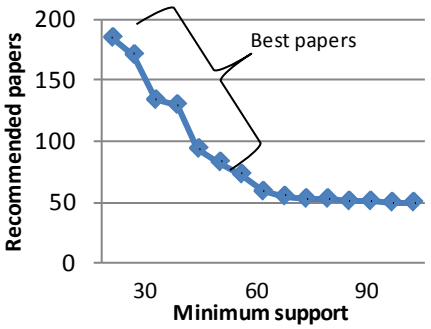

Figure 6. Recommendation papers for different [Ms] values and [Rm] of 0.8.

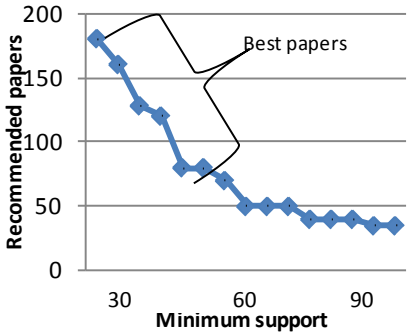

Figure 7. Recommendation papers for different [Ms] values and [Rm] of 1.0.

of the system. As such, we adopt ranked normalized discounted cumulative gain (nDCG) [3] [10] and mean reciprocal rank (MRR) [30]. These measures are standard information retrieval evaluation measures.

1) Normalized Discounted Cumulative Gain (nDCG) is used to evaluate ranked result of items, this makes it suitable for evaluation of our recommendation system. It is desirable that relevant citation paper appear at the top of the list.

$$\mathrm{nDCG}@i = Z_i \sum_{j=1}^{R} \frac{2^{r(j)}-1}{\log(1+j)}$$

where $r(j)$ is the ranking of the j-document in the ranking list, and the normalization constant Z_i is chosen so that the perfect list gets a NDCG score of 1. We use nDCG@R, where R = 3, 5, 7, 10. R is the number of top-R papers recommended by our proposed method.

2) Mean Reciprocal Rank (MRR) is used to determine the ability of the system to return a relevant paper at the top of the ranking. This measure is performed by computing the reciprocal rank, averaged over all target subjective feature set.

$$\mathrm{MRR} = \frac{1}{N_{tr}} \sum_{i=1}^{N_{tr}} \frac{1}{r_i}$$

where r_i is the rank of the highest ranking relevant paper for a target researcher i. N_{tr} is the total number of subjective feature set.

4.3. Discussion

Table 4 shows the recommendation accuracy for different minimum support for nDCG@3, nDCG@5, nDCG@7, nDCG@10 and MRR. The table also shows the comparison with Sugiyama and Kan [11] [15]. The result from the table shows that a minimum support of 60% achieved the best performance for potential paper recommendation. **Figures 8-12** show the graphical representation of the results obtained from **Table 4**. The results from the graph show that our system outperforms the baseline system with a minimum support [MS] of 60.

The recommendation baseline [11] is based on modeling previous works of a researcher in recommending scholarly papers to the researcher. The parameter that was considered are: Weight "SIM", *Th* = 0:4, = 0:23, *d* = 3. Its optimal performance is obtained for nDCG@5, NDCG@10.

The other recommender system [15] outperforms the first at the same nDCG value. It uses collaborative filtering and the investigation of sections available in papers to determine the relevance of the paper for recommendation. The system optimal recommender evaluation value is obtained with *n* = 4 and *Npc* = 5. These baseline values for [15] and [11] are obtained from [18].

5. Conclusion

In this paper, we have presented a significant approach for scholarly paper recommendation. Precisely, we considered the use of both subjective and objective features to construct a personalized paper recommendation system.

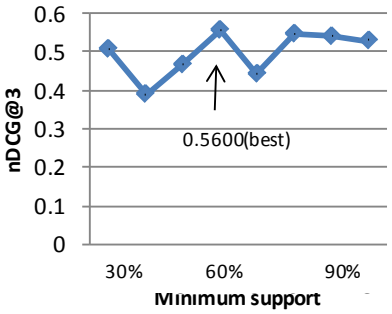

Figure 8. Graph of nDCG@3.

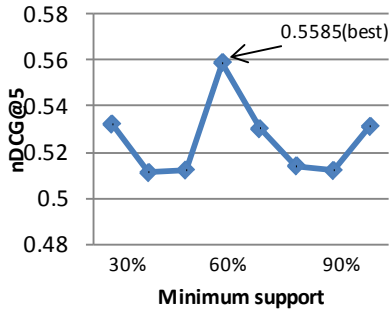

Figure 9. Graph of nDCG@5.

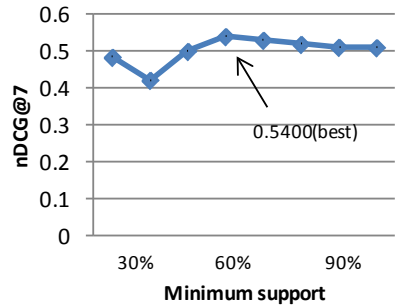

Figure 10. Graph of nDCG@7.

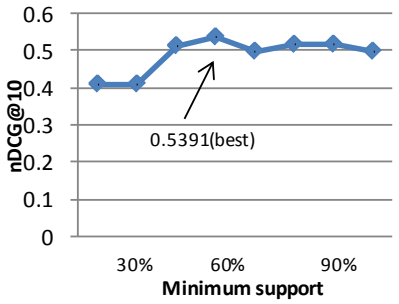

Figure 11. Graph of nDCG@10.

Our system is capable of providing a number of potential citation papers using FP growth algorithm with a minimum support value specified by the researcher.

The result from our implementation shows that potential paper recommendation system can be used by different researchers to meet their specific preferences. In addition, the evaluation performed using nDCG and

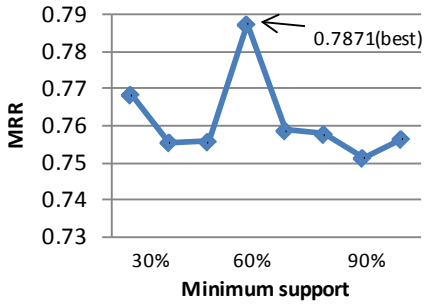

Figure 12. Graph of MRR.

Table 4. Recommendation accuracy for different minimum support.

Minimum Support (%)	nDCG@3	nDCG@5	nDCG@7	nDCG@10	MRR
30	0.5100	0.5320	0.4841	0.4110	0.7681
40	0.3912	0.5111	0.4220	0.4110	0.7552
50	0.4688	0.5122	0.5000	0.5150	0.7554
60	**0.5600**	**0.5585**	**0.5400**	**0.5391**	**0.7871**
70	0.4444	0.5300	0.5300	0.5000	0.7584
80	0.5500	0.5140	0.5190	0.5188	0.7574
90	0.5433	0.5120	0.5112	0.5192	0.7510
100	0.5321	0.5310	0.5100	0.4999	0.7561
Baseline [18]		0.519		0.4869	0.7595
Baseline [11]		0.547		0.5149	0.768

MRR show that it is a more efficient recommendation system in comparison to some baseline systems. We are confident that our method can be used in any field and by any researcher to achieve a personalized recommendation for unearthing potential citation papers. To improve the recommendation system, in future work, we can develop a hybrid approach which will consider, in addition to subjective and objective feature, a network-aware feature which is based on collaborative filtering (CF). This will create a balanced weighing system among all the features that will be used for recommendation.

References

[1] Nascimento, C., Laender, A.H.F., da Silva, A.S. and Gonçalves, M.A. (2011) A Source Independent Framework for Research Paper Recommendation. *Proceedings of the* 11*th ACM/IEEE Joint Conference on Digital Libraries* (*JCDL* 2011), Ottawa, 13-17 June 2011, 297-306.

[2] Wang, C. and Blei, D.M. (2011) Collaborative Topic Modeling for Recommending Scientific Articles. *Proceedings of the* 17*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (*KDD*'11), San Diego, 21-24 August 2011, 448-456. http://dx.doi.org/10.1145/2020408.2020480

[3] Zhang, M., Feng, S., Tang, J. Ojoko, B.A. and Liu, G.J. (2011) Co-Ranking Multiple Entities in a Heterogeneous Network Integrating Temporal Factor and Users' Bookmarks. *Proceedings of the* 13*th Asian-Pacific International Conference of Digital Libraries* (*ICADL*), **7008**, 202-211.

[4] Garfield, E. (1972) Citation Analysis as a Tool in Journal Evaluation. *Science*, **178**, 471-479. http://dx.doi.org/10.1126/science.178.4060.471

[5] He, Q., Pei, J., Kifer, D., Mitra, P. and Giles, C.L. (2010) Context-Aware Citation Recommendation. Copyright Is Held by the International World Wide Web Conference Committee (IW3C2), Distribution of These Papers Is Limited to Classroom Use, and Personal Use by Others, WWW 2010, 26-30 April 2010, Raleigh, North Carolina, USA, ACM 978-1-60558-799-8/10/04.

[6] Andersen, R., Borgs, C., Chayes, J., Feige, U., Flaxman, A., Kalai, A., Mirrokni, V. and Tennenholtz, M. (2008) Trust-

Based Recommendation Systems: An Axiomatic Approach. *Proceedings of World Wide Web*, Beijing, 21-25 April 2008, 199-208.

[7] Borgelt, C. (2005) Keeping Things Simple: Finding Frequent Item Sets by Recursive Elimination. *Workshop Open Source Data Mining Software* (*OSDM*'05, *Chicago*), ACM Press, New York, 66-70.

[8] Meho, L.I. (2006) The Rise and Rise of Citation Analysis. School of Library and Information Science, Indiana University, Bloomington.

[9] Jarvelin, K. and Kekalainen, J. (2000) IR Evaluation Methods for Retrievinghighly Relevant Documents. In: Belkin, N.J., Ingwersen, P. and Leong, M.-K., Eds., *Proceedings of the* 23*rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM Press, New York, 41-48.

[10] Nascimento, C., Alberto, H.F., Laender, A.S., Silva, M. and André, G. (2011) A Source Independent Framework for Research Paper Recommendation. *Proceedings of the* 11*th Annual International ACM/IEEE Joint Conference on Digital Libraries* (*JCDL*), Ottawa, 13-17 June 2011, 297-306. http://dx.doi.org/10.1145/1998076.1998132

[11] Sugiyama, K. and Kan, M.-Y. (2010) Scholarly Paper Recommendation via User's Recent Research Interests. *Proceedings of the* 10*th ACM/IEEE Joint Conference on Digital Libraries* (*JCDL*'10), Queensland, 14-17 August 2010, 29-38. http://dx.doi.org/10.1145/1816123.1816129

[12] Zaki, M., Parthasarathy, S., Ogihara, M. and Li, W. (1997) New Algorithms for Fast Discovery of Association Rules. *Proceeding of the* 3*rd International Conference on Knowledge Discovery and Data Mining* (*KDD*'97), Newport Beach, 14-17 August 1997, 283-296.

[13] Yang, D., Wei, B., Wu, J., Zhang, Y. and Zhang, L. (2009) A Ranking-Oriented CADAL Recommender System. *Proceedings of the* 9*th ACM/IEEE-CS Joint Conference on Digital Libraries* (*JCDL*'09), Austin, 15-19 June 2009, 203-211.

[14] McNee, S.M., Albert, I., Cosley, D., Gopalkrishnan, S.L.P., Rashid, A.M., Konstan, J.S. and Riedl, J. (2002) Predicting User Interests from Contextual Information. *Proceedings of the* 2002 *ACM Conference on Computer Supported Cooperative Work* (*CSCW*'02), New Orleans, 16-20 November 2002, 116-125.

[15] Sugiyama, K. and Kan, M.-Y. (2013) Exploiting Potential Citation Papers in Scholarly Paper Recommendation. *Proceedings of the* 13*th ACM/IEEE-CS Joint Conference on Digital Libraries*, Indianapolis, 22-26 July 2013, 153-162.

[16] Erasmus, M. (2011) Citation-Analysis, Impact Factors, H Index: Evaluating Journals, Authors, and Institutions. Medical Library Erasmus MC. http://www-fgg.eur.nl/medbib/Manuals/M208_Citatie-analyse.pdf

[17] Krell, F.-T. (2012) The Journal Impact Factor as a Performance Indicator. *European Science Editing*, **38**, 3-6.

[18] Page, L., Brin, S., Motwani, R. and Winograd, T. (1998) The PageRank Citation Ranking: Bringing Order to the Web. Technical Report SIDL-WP-1999-0120, Stanford Digital Library Technologies Project.

[19] Han, J., Pei, J. and Yin, Y. (2000) Mining Frequent Patterns without Candidate Generation. *Proceedings of the* 2000 *ACM SIGMOD International Conference on Management of Data*, *SIGMOD*'00, Dallas, 15-18 May 2000, 1-12. http://dx.doi.org/10.1145/342009.335372

[20] Agrawal, R. and Srikant, R. (1994) Fast Algorithms for Mining Association Rules. *Proceedings of the* 20*th International Conference on Very Large Data Bases* (*VLDB*'94), Santiago, 12-15 September 1994, 487-499.

[21] Fu, X., Budzik, J. and Hammond, K.J. (1998) Mining Navigation History for Recommendation. Infolab, Northwestern University, Evanston.

[22] Agarwal, R., Aggarwal, C. and Prasad, V.V.V. (2001) A Tree Projection Algorithm for Generation of Frequent Item Sets. *Journal of Parallel and Distributed Computing*, **61**, 350-371. http://dx.doi.org/10.1006/jpdc.2000.1693

[23] Tang, J., Zhang, J., Jin, R., Yang, Z., Cai, K., Zhang, L. and Su, Z. (2011) Topic Level Expertise Search over Heterogeneous Networks. *Machine Learning*, **82**, 211-237. http://dx.doi.org/10.1007/s10994-010-5212-9

[24] Price, D.J.D. (1963) Little Science, Big Science. Columbia University Press, New York.

[25] Van Raan, A.F.J. (2005) Fatal Attraction: Conceptual and Methodological Problems in the Ranking of Universities by Bibliometric Methods. *Scientometrics*, **62**, 133-143. http://dx.doi.org/10.1007/s11192-005-0008-6

[26] Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L. and Su, Z. (2008) ArnetMiner: Extraction and Mining of Academic Social Networks. *Proceedings of the* 14*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *SIGKDD*'2008, Las Vegas, 24-27 August 2008, 990-998. http://dx.doi.org/10.1145/1401890.1402008

[27] Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L. and Su, Z. (2010) A Combination Approach to Web User Profiling. *ACM Transactions on Knowledge Discovery from Data* (*TKDD*), **5**, Article No. 2. http://dx.doi.org/10.1145/1870096.1870098

[28] Tang, J., Fong, A.C.M., Wang, B. and Zhang, J. (2012) A Unified Probabilistic Framework for Name Disambiguation in Digital Library. *IEEE Transaction on Knowledge and Data Engineering*, **24**, 975-987.

http://dx.doi.org/10.1109/TKDE.2011.13

[29] Tang, J., Zhang, D. and Yao, L. (2007) Social Network Extraction of Academic Researchers. *Proceedings of* 2007 *IEEE International Conference on Data Mining* (*ICDM*), Omaha, 28-31 October 2007, 292-301. http://dx.doi.org/10.1109/icdm.2007.30

[30] Voorhees, E.M. and Tice, D.M. (1999) The TREC-8 Question Answering Track Evaluation. *Proceedings of the* 8*th Text Retrieval Conference*, *TREC*-8, Gaithersburg, 17-19 November 1999, 77-82.