Application of Content-Based Approach in Research Paper Recommendation System for a Digital Library

Simon Philip[1]
Department of Computer Science
Federal University Kashere, Gombe, Nigeria

P.B. Shola (PhD) [2]
Department of Computer Science
University of Ilorin, Ilorin, Nigeria

Abari Ovye John[3]
Department of Computer Science
Federal University Lokoja, Kogi, Nigeria

***Abstract*—Recommender systems are software applications that provide or suggest items to intended users. These systems use filtering techniques to provide recommendations. The major ones of these techniques are collaborative-based filtering technique, content-based technique, and hybrid algorithm. The motivation came as a result of the need to integrate recommendation feature in digital libraries in order to reduce information overload. Content-based technique is adopted because of its suitability in domains or situations where items are more than the users. TF-IDF (Term Frequency Inverse Document Frequency) and cosine similarity were used to determine how relevant or similar a research paper is to a user's query or profile of interest. Research papers and user's query were represented as vectors of weights using Keyword-based Vector Space model. The weights indicate the degree of association between a research paper and a user's query. This paper also presents an algorithm to provide or suggest recommendations based on users' query. The algorithm employs both TF-IDF weighing scheme and cosine similarity measure. Based on the result or output of the system, integrating recommendation feature in digital libraries will help library users to find most relevant research papers to their needs.**

Keywords—Recommender Systems; Content-Based Filtering; Digital Library; TF-IDF; Cosine Similarity; Vector Space Model

I. INTRODUCTION

Library users do experience difficulties in getting or finding favorite digital objects (e.g. research papers) from a large collection of digital objects in digital libraries. The increasing volume of information available on the internet has made it even more difficult for internet users to find exact information of interest. So, a recommender system becomes an important requirement in the design of digital libraries. This would assist library users in getting favorite digital objects (e.g. research papers) from the large collection of digital objects in general [1].

Recommender systems are software applications that suggest or recommend items or products (in the case of e-commerce) to users. These systems use users' preferences or interests (supplied as inputs) and an appropriate algorithm in finding the relevant or desired items or products. Recommender systems deal with information overload problems by filtering items that potentially may match the users' preferences or interests. These systems aid users to efficiently overcome the problem by filtering irrelevant information when users search for desired information [2].

Recommender systems use filtering algorithms to provide recommendations to users. These algorithms are classified or categorized majorly into collaborative-based filtering, content-based filtering, and hybrid algorithms [3].

Collaborative Filtering (CF) refers to an algorithm or technique that recommends items or products (in the case of e-commerce) to users based on the past ratings of other users (with similar interest or preferences) on the items or products collectively. It works by collecting users' feedback in the form of ratings for items in a given domain and exploring similarities in rating behavior amongst several users in determining how to recommend an item. This technique is subdivided into neighborhood-based and model-based techniques [4].

Content-based recommenders provide recommendations by comparing representation of contents describing an item or a product to the representation of the content describing the interest of the user (User's profile of interest). They are sometimes referred to as content-based filtering [1].

Hybrid algorithm combines both content-based and collaborative-based techniques to produce separate ranked lists of recommendations and then merge their results to produce a final list of recommendations [5], [6].

The content-based technique is adopted or considered here for the design of the recommender system for digital libraries. Content-based technique is suitable in situations or domains where items are more than users.

II. PROBLEM STATEMENT

Digital libraries offer a wide variety of digital objects (research papers, publications, journals, research projects, newspapers, magazines, and past questions). Some digital libraries even offer millions of digital objects. Therefore, getting or finding favorite digital objects (e.g. research papers) from a large collection of available digital objects in the digital library is one of the major problems library users encounter while using the library. The users need help in finding items (e.g. research papers) that are in accordance with their interests. Recommender systems offer a solution to this problem as library users will get recommendations using a form of smart search (users spend less time searching for digital objects). The problem considered here is then to develop or produce a software or system that users can use to locate quickly items of interest in a digital library containing a large collection of items.

III. Related Work

Reference [7] worked on a restaurant recommender system that was based on case-based recommendation technique. The adopted technique was used to select and rank restaurants. It was implemented to serve as a guide to attendees of the 1996 democratic national convention in Chicago and operated as a web utility.

Reference [8] applied content-based technique in paper recommendation system. The author used Jaccard similarity coefficient or jaccard index to compute similarity between users' query (users' attributes) and the attributes of the papers. The recommendations suggested by the system were sent via emails to the intended users.

Reference [9] designed a group recommender system for Facebook. He used hierarchical clustering and decision techniques to suggest or recommend the most suitable Facebook group (s) to Facebook users. He extracted profile information of the Facebook members at University of North Texas and used it as a test data.

Facebook recommendation system provides friends recommendations or suggests friends as "people you may know". These suggestions or recommendations are based on mutual friends, work and educational information, groups you are part of, contacts you have imported using friends finder and many other factors. This recommendation system uses facebook users' profile [10].

Amazon's customers who bought, CDNOW.com's Album Advisor, MovieFinder.com's Match Maker, and Reel.com's Match Maker use item to item correlation as recommendation technology to provide recommendations to their customers. Amazon's customers who bought feature recommend products to its customers. CDNOW.com's Album Advisor suggests music to its customers. MovieFinder.com's Match Maker, and Reel.com's Match Maker recommend Videos to their customers [11], [12], [13], [14].

IV. Methodology

The use of collaborative-filtering technique in recommending research papers has been criticized by some authors. Authors like [15] suggest that collaborative-filtering technique is ineffective in domains where items (e.g. research papers) are more than users. [16] Said; "Users are not willing to spend time to rate items explicitly". Hence, content-based approach is adopted for the design and implementation of research paper recommender system. This approach does not depend on the ratings of other users but uses the contents describing the items and the users' taste or needs. The researchers used the following data collection procedure and methods in representing the research papers, users' profile of interest, and also in providing recommendations to the users.

- Dataset for the system: Sources of the research papers are the research papers published by the academic staff of federal university kashere, and also the ones from open sources obtained on the internet. Information about users' profile of interest is collected from the users during their transactional behaviors or the usage of the system. For instance, information about the users' profile can be collected when a user downloads, opens or likes a research paper.
- Keyword-Based Vector-Space Model: The researchers used this model with basic TF-IDF weighing technique to represent a research paper as a vector of weights, where each weight indicates the degree of association between a research paper and a term or keyword.
- Item Representation: The items (research papers) are represented by a set of features (also called attributes or properties). These attributes are: title of the paper, abstract, keywords, research area, ID of the paper, and the authors. The abstract represents the research paper when the frequency of a term in the research paper is being determined.
- TF-IDF and Cosine Similarity: The researchers used TF-IDF and cosine similarity to determine how relevant or important a research paper is to a user's query. The importance increases proportionally to the number of times a term (in the user's query) appears in the research paper. TF-IDF is given by:

$$tf - idf(t, d, D) = tf(t, d) \times idf(t, D) \quad (1)$$

Where t=term in the user's query

d= a document in the collection, D= a collection of documents

TF=Term Frequency given by:

$$tf(t, d) = \frac{N_{t,d}}{N_d} \quad (2)$$

IDF= Inverse Document Frequency which is given by:

$$idf(t, D) = log \frac{N}{|d \in D : t \in d|} \quad (3)$$

Where

N = number of documents in the collection

N_d= Number of terms in the document *d*

$N_{t,d}$ = Number of times term *t* appears in document *d*

The Cosine similarity is a measure of similarity between two vectors of an inner product space that measures the cosine of the angle between them. The researchers used this method to determine how similar a research paper is to a user's query or paper that a user has liked in the past. The research papers are represented as vectors of weights, where each weight indicates the degree of association between the research papers and the tern.

Given two research papers or documents d_j, d_k represented as vectors of weights, their similarity is measured by:

$$\text{Sim}(d_j, d_k) = \frac{\overline{d_j}.\overline{d_k}}{|\overline{d_j}|.|\overline{d_k}|} = \frac{\sum_{i=1}^{n} w_{i,j}.w_{i,k}}{\sqrt{\sum_{i=1}^{n} w_{i,j}^2}\sqrt{\sum_{i=1}^{n} w_{i,k}^2}} \quad (4)$$

Where $W_{i,j}$ = Weight of term *i* in document *j*

$W_{i,k}$ = Weight of term *i* in document *k*

Note: Stemming was not applied because of the following reasons: Losing context of search, may reduce precision, and cannot be applied to proper nouns [17].

V. Recommendation Algorithm Based on Users' Query

The proposed algorithm for generating recommendations for an active user based on the user's query is shown below:

1) START

2) Enter a search query Q

3) Process the user's query Q

*a) Extract tokens from the user's query e.g if the query is "**What are the data mining techniques**", the system will explode the query into the following tokens: **What, are, the, data, Mining, Techniques***

*b) Remove **stop words** e.g. using the query above, the system will remove the following tokens (**stop words**): **what, are, the***

c) Store the remaining tokens in an array after stop words removal

4) Retrieve relevant or similar research papers to the user's query Q which forms a collection "C".

5) Determine the weight (Using TF-IDF) of each token in the user's query Q and store the weights in an associative array say Q_weights

Note: Query Q is now represented as a vector of *tf-idf* weights

6) FOR k=1 to N

$$\text{Sim}(d_k, Q) = \frac{\sum_{i=1}^{n} w_{i,k}.w_{i,Q}}{\sqrt{\sum_{i=1}^{n} w_{i,k}^2}\sqrt{\sum_{i=1}^{n} w_{i,Q}^2}}$$

Sim_Values[d_k]= $\text{Sim}(d_k, Q)$

NEXT

7) Sort the associative array Sim_Values in descending order with respect to similarity value

8) FOR d=1 TO N

IF Similarity_value of "d" >=0.3 THEN

Retrieve the details of Document 'd', and display it

END

NEXT

Note: *Sim_Values* is an associative array containing the similarity values of all relevant documents to Query Q.

Q = query supplied by an active user.

N= number of documents or research papers in the collection "C".

$W_{i,k}$ = Weight of term i in document k

$W_{i,Q}$ = Weight of term i in Query Q Sim_Values= An associative array containing the similar papers in order of their similarity values to query Q.

VI. Results and Discusion

The results obtained from the developed system were compared with the results of a digital library without recommendation feature and found to be correct and with even additional features that are not available in the digital library. The results therefore, are in conformity with most of the literatures reviewed. Thus, the research paper recommendation system integrated in the digital library has numerous advantages over the ones without recommendation feature.

A. The Library Users' Search Page

Figure 1 allows library users to search for research papers in the digital library. The papers displayed were based on the user's supplied query. The display is done in the order of their importance or relevance (*computed using TF-IDF technique and cosine similarity*) to the user's query.

Fig. 1. Library users' search page

B. The Library Users Recommendation Page Based On the Users' Taste Supplied As a Query

Figure 2 shows the research papers recommended based on the users' taste supplied as a query.

Fig. 2. Library users' recommendation page based on the user's taste supplied as a query.

VII. CONCLUSION

Research paper recommender systems help library users in finding or getting most relevant research papers over a large volume of research papers in a digital library. This paper adopted content-based filtering technique to provide recommendations to the intended users.

Based on the results of the system, integrating recommendation features in digital libraries would be useful to library users. The solution to this problem came as a result of the availability of the contents describing the items and users' profile of interest. Content-based techniques are independent of the users ratings but depend on these contents.

This paper also presents an algorithm to provide or suggest recommendations based on the users' query. The algorithm employs both TF-IDF weighing and cosine similarity measure.

VIII. FUTURE WORK

The next step of our future work is to adopt hybrid algorithm to see how the combination of collaborative and content-based filtering techniques can gives us a better recommendation compared to the adopted technique in this paper.

REFERENCES

[1] J. Raymond. Mooney and R. Loriene.: Content- Based Book Recommendation Using Learning for Text Categorization. In proceedings of the fifth ACM conference on digital libraries, pages 195- 204, San Antonio, TX, June 2000.

[2] P. Resnick, H. Varian: Recommender Systems. Communications of the ACM, pages 56-58 (1997)

[3] Y. Koren., R.M. Bell, C. Volinsky: Matrix Factorization Techniques For Recommender Systems IEEE Computer pages 30-37 (2009)

[4] S. Xiaoyuan. and M.K. Taghi: A Survey of Collaborative Filtering Techniques in Artificila Intelligence, pages 1-20, 2009

[5] R. Burke,: Hybrid Recommender Systems: Survey and Experiment. User Modeling and User-Adaptive Interaction pages 331-370, November, 2002

[6] P. Cotter and B. Smyth: Intelligent Personalized TV Guides. In twelfth conference on innovative applications of artificial intelligence", pages 957- 964, 2000.

[7] J. L. Kolodner: A Restaurant Recommender System, 1993

[8] B.S. Oladapo: A Research Paper Recommender System, 2013 unpublished.

[9] E.Baatarjav, J.Chartree, and T. Meesumrarni: Group Recommendation System for Facebook, 2010

[10] http://www.facebook.com

[11] http://www.amazon.com

[12] http://www.CDNOW.com

[13] http://www.movieFinder.com

[14] http://www.reel.com

[15] N. Agarwal, E. Haque, H. Liu, and L. Parsons: Research Paper Recommender Systems: A Subspace Clustering Approach, Advances in Web-Age Information Management,"*Springer: Heidelberg*., 2005.

[16] R. Torres, S.M. McNee, M. Abel, J. Konstan, J. Riedl: Enhancing Digital Libraries With Techlens, in *JCDL 2004*, 2004, pp. 228–236.

[17] A.B. Manwar, S.M. Hemant, K.D. Chinchkhede, C. Vinay: A Vector Space Model for Information Retrieval:A MATLAB approach, 2012.