Publication Recommendation System for Scientific Community in Agriculture

Sowndarya C.A., Shashi Dahiya, Anshu Bharadwaj, Sudeep, Soumen Pal and Rajender Parsad

ICAR-Indian Agricultural Statistics Research Institute, New Delhi

SUMMARY

There exist several publication recommendation systems which recommend publications to researchers and academicians based on their area of interest. A publication recommendation system specifically meant for recommending the agricultural publications will help the agricultural researchers and academicians in getting recommendations in the form of new publications published in their area of interest/work. The developed system uses the keywords and title of the publications to find out the similarity between the newly added publication and all existing publications in the publication repository. The TF-IDF model and the Cosine similarity measure have been used for finding out the similarity using R statistical software. The recommendation of the new publication is sent to the authors of the five most similar publications through mail. The system has been tested using the data from KRISHI Publication and Data Inventory Repository. The agricultural publication recommendation system successfully sent publication alerts to the authors of most similar publications from the data. This system can help the agricultural researchers by recommending the newly published research to the authors of existing publications working in the similar area using an email alert. It will support the researchers to stay up-date with the latest work carried out in their area of interest.

Keywords: Recommendation system, Text mining, Agricultural publications, Similarity, TF-IDF, Cosine similarity.

1. INTRODUCTION

Since the inception of internet, overloading of information and content proliferation is a problem. Technology interventions are needed to ensure that only relevant items are viewed, and other items are filtered out. Today, a tremendous number of recommendation systems are evolving alongside the growth of web-based information systems. A Recommendation System (RS) is a tool that actively finds information from a large amount of information which might be of interest for a user (Liling *et al.* 2019). These are computerized systems for filtering out possibly items which are of little or no interest to users (Low *et al. 2019*). They filter the vital information out of the large amount of dynamically generated information according to user's preferences or observed behavior about an item. From electronic commerce to digital marketing, recommender systems in our day-to-day online journeys are inevitable and deal with the problem of information overload.

Publication recommendation systems supports academic researchers with relevant papers in the area of their interest and help them quickly locate the papers they need. There exist various publication repositories like Google Scholar, Research Gate etc. which provides recommendations to the authors regarding newly published papers. Various approaches for developing successful publication recommendation systems have been adopted in literature. Content-based Recommendation Approach has been used to improve the access to digital research libraries in which a module for key phrases retrieval is used to provide a detailed description of both paper content and consumer preferences (Ferrara *et al.* 2011). A user profile-based personalized recommendation approach has been used to measure the similarity between the given topic and collected papers to recommend the initial paper for each topic in retrieval of information (Hong *et al.* 2012). A Semantic Web based recommendation approach has

Corresponding author: Shashi Dahiya
E-mail address: shashi.dahiya@icar.gov.in

been used in the knowledge base of agriculture and spatial data (Kumar *et al.* 2013) to provide agricultural recommendation. An algorithm employing Tf-Idf weighing technique and cosine similarity measure was used to make recommendations based on user's query (Philip *et al.* 2014). A recommender system for research paper was proposed by enhancing existing search engines with recommendations based on previous searches carried out by other researchers that prevent time absorbing searches (Patil *et al.* 2015). To facilitate the development of literature and to find useful research paper or article in short period, a new concept-based recommendation technique is suggested by (Sharma *et al.* 2017), which defines research articles about its concept is used to recommend semantically relevant papers to researcher. For scholarly paper recommendation, a citation-based recommendation system was proposed which focused on the research paper's latent relationship, with a goal of personalizing academic recommendation irrespective of user expertise and research field focused solely on papers-citations relationships (Haruna *et al.* 2018). A method to recommend most relevant articles to the user's profile by considering user's publications, user's co-authors and co-authors' papers had been proposed by Kaya, 2018. Bai *et al.* 2019.

1.1 Publication Recommendation System for Scientific Community in Agriculture

The Indian Council of Agricultural Research (ICAR), (https://icar.org.in) is an autonomous organization under the Department of Agricultural Research and Education (DARE), Ministry of Agriculture and Farmers Welfare, Government of India. It has played a pioneering role in ushering Green Revolution and subsequent developments in agriculture in India through its research and technology development that has enabled the country to increase the production of food grains, horticultural crops, fish, milk and eggs since 1950-51, thus making a visible impact on the national food and nutritional security. It is engaged in cutting edge areas of science and technology development and its scientists are internationally acknowledged in their fields. ICAR has 113 research institutes, 60 All India Coordinated Research Projects and more than 731 Krishi Vigyan Kendra's/ Agricultural science center's spread across the length and breadth of the big country. The ICAR employees are engaged in research, academics and

extension in the field of agriculture. For bringing the widely spread knowledge resources of ICAR to all stakeholders at one place, ICAR has brought its initiative KRISHI Portal. KRISHI is the Agricultural Knowledge Resources and Information System Hub for Innovations" (https://krishi.icar.gov.in/). The portal serves as a centralized data repository system of ICAR consisting of Technology, Data generated through Experiments, Surveys, Observational studies, Geo-spatial data, Publications, Learning Resources etc. In the KRISHI publication repository, agricultural researchers from all ICAR institutes have the authority to access and upload all types of published documents including reports, bulletins, research papers, web resources, media resources etc.As the number of publications are growing in KRISHI publication repository, a recommendation system can be very helpful in recommending the newly published and uploaded papers to the agricultural research community as per their area of research interest.

Authors have developed a methodology in this paper to develop a recommendation system for scientific community in agriculture using web development techniques, R software and text similarity checking techniques of TF-IDF and Cosine Similarity function for identification of similar documents. The objective of this work is to provide an online recommendation platform from where agricultural researchers can get recommendations and details of the latest work (abstract) being done by other research community in similar areas using an email alert, which will support the researchers to stay up-date in their area of interest.

2. METHODOLOGY

2.1 System Architecture

The AgPubRS application has been developed as an online application using the Client-Server architecture. The Database of the RS is deployed using PostgreSQL database management system (Oba and Hsu, 2017). It has all the information related to the publications, including publication Id, title, keywords, authors and co-authors along with their email ids, and Universal Resource Identifier (URI) for the publication. Hyper Text Markup Language (HTML), Cascading Style Sheet (CSS), and JavaScript has been used for designing the client side pages. Server-Side Application Layer is built using Java Server pages (JSP), which provides a safe and fast platform for creating interactive content

on the servers. A connection of the web application with R-studio has been established for checking the similarity of the publications (Silge and Robinson, 2017). The user interface is separated from the logical database. It interacts with the AgPubRS application on a remote server that in turn interacts with software concerned for long term data storage on a database server. For developing and testing the AgPubRS application, publications submitted in KRISHI publication repository were used.

2.2 Identification of Similar Documents

The TF-IDF and the Cosine Similarity algorithms are used for finding the similarity between Title and Keywords of new publication with all existing publications in the database.

The similarity between publications is computed using the following two R functions:

- weightTfIdf (m, normalize = TRUE)

- cosine (p, q = NULL)

The Term Frequency-Inverse Document Frequency (TF-IDF) value is a numerical index for assessing the significance of the term to a document in a cluster. It is often used in information extraction.
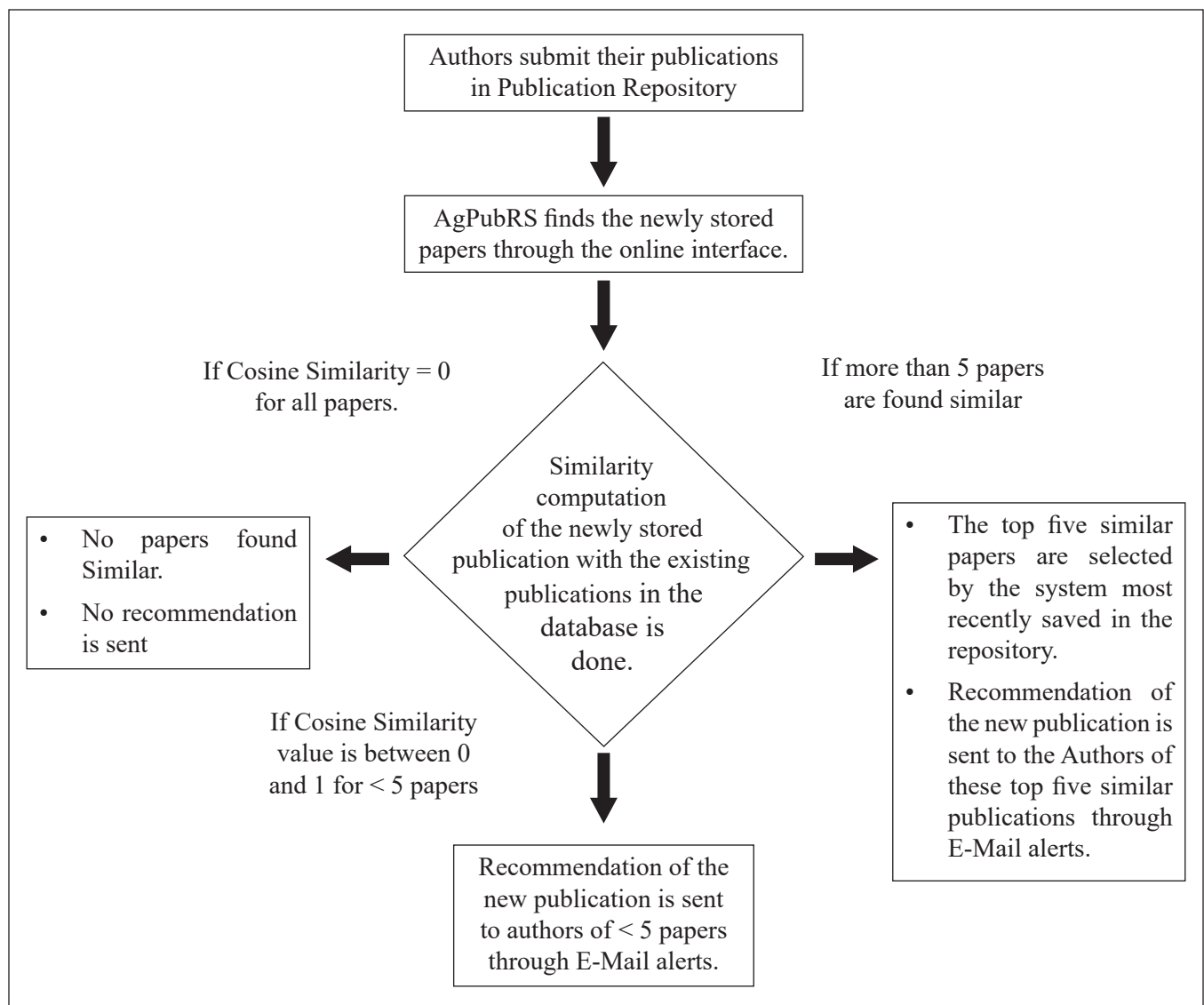
$$f_{ij}*\log [n/(d_j+1)]$$

Fig. 1. AgPubRS Process Flow

where f_{ij} is x_{ij}/m_i or x_{ij}, depending on normalization, and d_j is the number of document that contains token j.

Parameters: m – Term-Document Matrix.

Normalize – boolean value, which indicates whether to normalize term frequency by document totals.

Value: A matrix of the same type as m, with values replaced by the tf-idf.

Cosine Similarity is defined as similarity index that is mainly used to determine the angle of cosine between the two vectors p and q, or between all the column vectors of a matrix p. The above matrix may be a Term-Document Matrix, so that the columns will be the documents and the rows will be the terms. The range of cosine similarity is 0 to 1 (including 0 and 1), with 1 indicating the highest similarity between the two documents, and 0 indicating the least or no similarity.

$$\text{Cosine similarity: Cos (p, q)} = \frac{p \cdot q}{(\|p\| * \|q\|)}$$

Parameters: p - Vector or Matrix (for example, Term-DocumentMatrix).

q - Vector with suitable dimension to p. If 'NULL' is used, all column vectors of p are compared. (optional)

Value: The symmetrical matrix with as many rows as there are columns in input.

The process flow of AgPubRS recommendation system is depicted in Fig:1.

3. RESULTS AND DISCUSSION

A recommendation system for agricultural publications "AgPubRS" has been designed, developed and successfully tested with the data from KRISHI Publication and Data Inventory Repository. A total of 739 (656 from IASRI and 83 from NAARM) publications from IASRI and NAARM institute have been used for similarity checking of 20 newly arrived papers in the database.

The cosine similarity value ranges from 0 to 1, with 1 indicating the highest similarity and 0 being the least. No threshold value was set on cosine values to find the top 5 similar publications, hence the authors and co-authors of top five similar publications with respect to newly added publications are selected to send the recommendations. If more than 5 papers have the highest cosine values, then latest papers are selected

as similar papers by using the paper id which is an auto incremented id. If only few papers (less than 5 papers with cosine value above 0) are found similar to the newly added publication, then it is recommended to authors and co-authors of those many papers only. If there is no paper found in the repository having the cosine value more than 0, then no recommendation is sent.
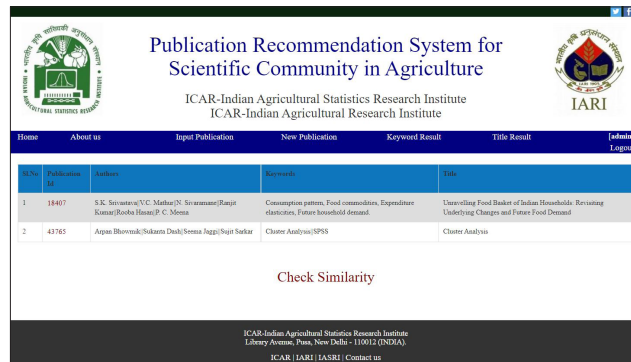

Fig. 2. Similarity Checking for New Publications

Fig.2. depicts the Screen for checking the Similarity of papers. It will run the R code to find the top five publications similar to the new publications. The administrator is presented with the screen depicted in Fig.3, from where he can see the Cosine Similarity Matrix by Title, Keywords and can also view the top five similar publications identified by AgPubRS.
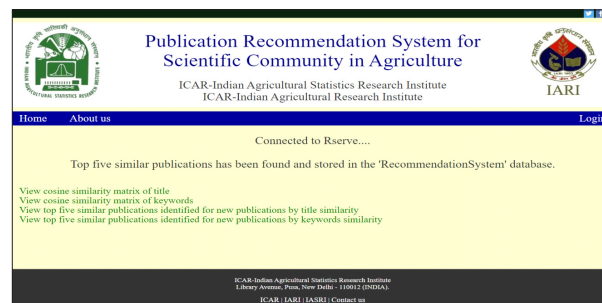

Fig. 3. View top 5 Publications and Similarity after Connecting to R

In Fig.4 and Fig.5,the administrator can view the Cosine Values of top 5 similar papers based on Keyword based similarity checking and Title based similarity checking.

Fig. 6 shows the new publication being recommended to the authors of 5 most similar publications through a dedicated KRISHI Publication Alert mail id. A complete reference and link of the newly published research paper is also included in the alert.

IASRI Paper id 43765: Keywords- Cluster Analysis||SPSS				
Similar papers id	43307	34135	31544	29028	8572
Cosine value of similar papers	0.292190078542673	0.359174533328456	0.270327168375459	0.401878021455894	0.304695955524016
Keywords of similar papers	regression analysis||cluster sampling||Two-stage sampling	cluster Sampling||Overlapping Clusters	Non-Hierarchical Cluster analysis||Diallel analysis||SPAR I	Cluster Analysis||Sample Survey Methodology	Random Sampling||Cluster Sampling||Non overlapping clusters

Fig. 4. Keyword Similarity Result

IASRI Paper id 43765: Title- Cluster Analysis				
Similar papers id	37770	34139	34135	8572	7967
Cosine value of similar papers	0.289179108401114	0.555944168553512	0.555944168553512	0.329735898987113	0.289179108401114
Title of similar papers	Web based fuzzy C-Means Clustering Software (wFCM)	Ordered Cluster Sampling	Ordered Cluster Sampling	Some Studies on Non-Overlapping Clusters of Two Units	Web based fuzzy C-Means Clustering Software (wFCM)

Fig. 5. Title Similarity Result

Fig. 6. Recommendation sent via email

4. CONCLUSION

Agricultural research and development are very important for the bright future of agriculture, food security, environment sustainability and reducing the pressure on natural resources. For achieving these, targets, agricultural research is conducted globally in all areas directly or indirectly linked to agriculture. Hence a lot of research papers are published by agricultural researchers every year. Many researchers are working globally in similar areas without any knowledge of what others are doing in these areas. A publication recommendation system for scientific community in agriculture has been designed and developed under this study. It has been developed to provide recommendations of latest published papers to the agricultural researchers working in similar areas. The agricultural researchers working in the similar areas are found by first applying the Term Frequency-Inverse Document Frequency (TF-IDF) for assessing the significance of the term to a document in a cluster. The cosine similarity measure

is then applied that determines the angle of cosine between all the column vectors of a Term-Document matrix. The cosine similarity value ranges from 0 to 1, with 1 indicating the highest similarity and 0 being the least. The authors of top five publications which have got highest cosine similarity values with respect to the newly added publication will get the recommendation of this new publication through email.

The recommender system has been tested using a total of 739 publications data from KRISHI Publication Repository. Similarity of 20 newly arrived papers in KRISHI was successfully found using the methodology adopted in this paper. The recommendations were sent to the authors of top 5 publications using the email alerts. The abstract of the paper is also sent along with the email alert to the authors. The timely recommendations keep the researchers up-date with the latest work carried out in their area of research interest.

REFERENCES

Bai, X., Wang, M., Lee, I., Yang, Z., Kong, X., and Xia, F. (2019). Scientific paper recommendation: A survey. *IEEE Access*, 7, 9324-9339.

Ferrara, F., Pudota, N., and Tasso, C. (2011). A key phrase-based paper recommender system. In *Italian research conference on digital libraries* (pp. 14-25). Springer, Berlin, Heidelberg.

Haruna, K., Ismail, M.A., Bichi, A.B., Chang, V., Wibawa, S., and Herawan, T. (2018). A citation-based recommender system for scholarly paper recommendation. In *International Conference on Computational Science and Its Applications* (pp. 514-525). Springer, Cham.

Hong, K., Jeon, H., and Jeon, C. (2012). UserProfile-based personalized research paper recommendation system. In *2012 8th International Conference on Computing and Networking Technology (INC, ICCIS and ICMIC)* (pp. 134-138). IEEE.

Kaya, B. (2018). User profile based paper recommendation system. *International Journal of Intelligent Systems and Applications in Engineering*, **6(2)**, 151-157.

Kumar, V., Dave, V., Bhadauriya, R., and Chaudhary, S. (2013). Krishimantra: agricultural recommendation system. In *Proceedings of the 3rd ACM Symposium on Computing for Development* (pp. 1-2).

Liling, L.I.U. (2019). Summary of recommendation system development. In *Journal of Physics: Conference Series* (Vol. 1187, No. 5, p. 052044). IOP Publishing.

Low, J.M., Tan, I.K., and Ting, C.Y. (2019). Recent Developments in Recommender Systems. In *International Conference on Multi-disciplinary Trends in Artificial Intelligence* (pp. 38-51). Springer, Cham.

Oba, R.O., and Hsu, L.S. (2017). PostgreSQL: Up and Running: A Practical Guide to the Advanced Open Source Database. O'Reilly Media.

Patil, S., and Ansari, M.B. (2015). User profile based personalized research paper recommendation system using top-K query. *Int. J. Emerg. Technol. Adv. Eng.*, **5**, 209-213.

Philip, S., Shola, P., and Ovye, A. (2014). Application of content-based approach in research paper recommendation system for a digital library. *International Journal of Advanced Computer Science and Applications*, **5(10)**, 37-40.

Sharma, R., Gopalani, D., and Meena, Y. (2017). Concept-based approach for research paper recommendation. In *International Conference on Pattern Recognition and Machine Intelligence* (pp. 687-692). Springer, Cham.

Silge, J., and Robinson, D. (2017). *Text mining with R: A tidy approach*. " O'Reilly Media, Inc.".

Web Resources Accessed:

https://icar.org.in

https://krishi.icar.gov.in/

https://www.rdocumentation.org/packages/RPostgreSQL/versions/0.6-2

https://www.rdocumentation.org/packages/tcR/versions/2.2.4/topics/cosine.similarity

https://www.rdocumentation.org/packages/textir/versions/2.0-5/topics/tfidf

http://www.w3schools.com/css

http://www.w3schools.com/html