SOCIETY FOR SCIENCE AND EDUCATION
UNITED KINGDOM
TMLAI TRANSACTIONS ON MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE
ISSN 2054-7390

A Model- Based Research Material Recommendation System For Individual Users

Nikhat Akhtar
Research Scholar Ph.D (Computer Science & Engineering)
M.Tech, B.Tech (Computer Science & Engineering)
Department of Computer Science & Engineering,
Babu Banarasi Das University, Lucknow, India
dr.nikhatakhtar@gmail.com

ABSTRACT

As there is an enormous amount of online research material available, finding pertinent information for specific purposes has become a tedious chore. So there is a requirement of the research paper recommendation system to facilitate research scholars in finding their interested and relevant research papers. There are many paper recommendation systems available, most of them are depending on paper assemblage, references, user profile, mind maps. This information is generally not easily available. The majority of the prevailing recommender system is based on collaborative filtering that rely on other user's proclivity. On the other hand, content-based methods use information regarding an item itself to make a recommendation. In this paper, we present a research paper recommendation method that is based on single paper. Our method uses content-based recommendation approach that employs information extraction and text categorization. . It performs the profile learning by using naive Bayesian text classifier and generates recommendation on the basis of an individual's preference.

Index Terms **—** Data Extraction, Text Classification, Profile Learning, Recommendation, Information Extraction.

1 Introduction

Large amount of time of the research people invests in internet in searching their interested papers because there are online research papers easily available in enormous amount. Most of the online research papers storehouse like journals and conference proceedings present their research papers corresponding to the year of publication and volumes , which make it strenuous to get accompanying research papers. Even most of the conferences and journals are not indexed in the most popular search engine among researchers like Google Scholar. For successful search of research papers, a user must have knowledge about correct link to that journal so that searching can possibly by year of publication, volumes and numbers that is very time consuming task. The most authentic way to overcome this problem is the research paper recommender system. Research paper recommendation system aspires to recommending apropos research papers to researchers, according to their personal preferences [1].The most promising recommender system is that which has the capability to gratify its user's information retrieval requirements. As different people may be interested in different ways like some of them may be

DOI: 10.14738/tmlai.52.2842
Publication Date: 13th March, 2017
URL: http://dx.doi.org/10.14738/tmlai.52.2842

interested in the first published research paper of their interest other may like to read more latest paper of their research interest. Most of the researchers like to cite the research papers which use the same methods to solve the similar problems. Many of the research paper recommender systems are using User profiles [2,3,4] and some are based on the citation relation [1]. User profiles based recommender systems necessitate that the researchers are already registered with their profiles and the research papers are recommended to the researchers on the basis of similarities between their profiles. There are some shortcomings in these methods as new and not registered researchers cannot take advantage from them. Some other types of recommender systems require keywords from their researchers to represent their interest, In that type it is just like a search engine which require a search query from the researcher to extract desired papers. This method is useful, but creating queries for searching new research articles can be a strenuous task [5].

The major imperfection of this perspective is that it requires rendering user's information retrieval requirements in the most suitable query by the user in order to fetch only literally require research paper [6]. This is a research paper recommendation system that uses content-based filtering which is similar to our method. Their approach needs only a single as input and constructs many queries with the help of words in that paper, these queries are then submitted to an existing web system that contain research papers. They consider the title, abstract and body as target section for query generation and used the title and abstract section for candidate papers generation. We propose an approach that can recommend related papers based on the topics the target paper is addressing and its main idea by considering the full paper.

Many recommender systems employ a collaborative filtering approach, in which computerized matching is performed. These systems contain a repository of users individual choices and they search users whose choices tally consequentially with that given user, and recommends to that user other items enjoyed by their matched user. This method presumes that a users' choices are tally with other users' choices in the system and for this purpose adequate amount of user rating should accessible. Items that have not been rated by adequate number of uses cannot be recommended properly. Consequently, collaborative filtering methods continuously favor to recommend widespread title, eternizing congruity in reading choices. Further, as adequate data belonging to other users are requisite to make a recommendation, this method arises perturb related to privacy issues in accessing other data. On the other hand, content-based recommendation learns examples to distinctively characterize each user without comparing it to other's [7]. The recommended items are based on information about the item itself rather than on the preferences of other users.

The certainly content-based filtering provide initial objective for the system by its users itself. Machine learning for text-categorization has been applied to content-based recommending of web pages [8] and newsgroup messages [9]; however, to our knowledge has not previously been applied to research papers recommending. We are trying to probe content-based research paper recommending by implementing automated text-categorization methods to semi-structured text extracted from the web, . For paper recommendation; title, abstract, introduction and related works sections of candidate paper are considered .Our system uses a database of papers information extracted from web pages in the ACM digital library. Researchers provide 1-10 ratings for a selected set of training papers; the system, then learns a profile of the user using a Bayesian learning algorithm and produces a ranked list of the most recommended additional titles from the system's catalog.

The first part of this paper briefly presents some of the research literature related to the existing approaches to designing the recommender system. The other parts introduce the proposed system architecture, the technologies that have been used. The paper is concluded by presenting the conclusions and future work.

2 Related Works

There are three essential viewpoints that are used in design of recommendation systems, content-based, collaborative filtering and hybrid [10]. The content-based recommender systems habituated scheme of analogous up the hitherto collected attributes of a user profile with those of the items volition, in access a episodic result [11]. Another content-based filtering method creates practicable recommendation instituted exclusively on the description of the items themselves and not the enamored users. These situations for the preparatory stage of extremely digital libraries and identical information retrieval systems [6]. The proportionality matching in this viewpoint is straightforward so far as it's matching an item with other items, unless in order for the recommendation to make comprehension, there is the requirement for beginning item that is known to be of some liking for the user. The collaborative filtering recommender systems [19] recommend items based on the past liking of identifying users. The recommendation is based on the imagination that items identified by users with analogous profiles with a substantial user, are highly believable to be liked by the recent. Supposing the users of the digital library are not actively involved by construction reviews or render some feedback concerning the articles, or supposing they do not have full specified profiles (research area, liking), this database would scarcity of vital data in the recommendation process. In spite of these data exists, there is an exalted probability that the recommendation process generates improved outcome [6], [7]. Hybrid recommender systems [7] commonly use a mixture of content based and collaborative filtering recommendation for recommending items. This mixture viewpoint deals with the deficiency of the above mentioned ones, permit for a preparatory content-based recommendation in this matter of a cold start (deficiency of user profiles) [6]. The collaborative-filtering recommendation can be rectified the outcome by adding context-respective information in the content-based approach.

Today scenario recommender systems are much prevalent in commercial applications these days, recommender systems for the academic research have also obtain liking. We are observing by the emergence of an agglomeration of research papers about this context presented at many conferences and journals.

The Docear is an academic literature accessory to search, organize, and evolve research articles [20]. The recommender system [21] uses content based filtering methods to recommend articles. It permits the users to construct "mind maps" that delineate a user model, which is coinciding with Docear's Digital Library. The authors requisition to have cognizable a reasonable outcome based on the number of clicks enlist via about 31 thousand tested recommendation outcome. In the [22] an individualize academic research paper recommendation system is presented. It recommends episodic articles in the research field of the research users. It is supposed that the researchers interest their personal articles. This system uses, a web crawler to access research papers from two solid digital libraries: IEEE Xplore and ACM Digital Library. It praxis text equality to ensure the equality amid two research papers and collaborative filtering methods to recommend the items. Nascimento et al. Endow another illustration of a content-based recommender system for scientific articles [6]. They insinuate that the majority of the recommender system approaches believes that a large collection of scientific papers is available previously. In this

situation for some digital libraries preference IEEE Xplore, but it does not hold for many other circumstances. Their proposed solution relies on publicly obtainable scientific metadata, literally the title and abstract of the articles. Their designed system accumulates these data by simulating searches on the websites of several publishers. In lieu of using user defined keywords, they create keywords from an exceptional article that is presented by the users. The symmetry of the articles is calculated by using the cosine similarity based on the vector space model [23]. The same symmetry measure is used in our designed recommender system.

The outcome obtains by Nascimento et al. Werejustly positive, demonstrating that it is sufficient to ponder only the title and abstract of the articles for recommendation intention. One more approach used by some academic paper recommender systems praxis the paper's citations for recommending articles. In [12] it is presented one more hybrid recommendation system. Its purpose is an impressive substitute for academic discover engines by not exclusively keep faith on keyword analysis, but besides using citation analysis, explicit and implicit ratings, author's analysis, and source analysis. The famous academic discover engine CiteSeerX also uses the citations to quest analogous scientific papers . The stemming algorithm proposed by Sadiku and Biba in [24] has been experimenting with deal out documents written in Albanian about literary, chemistry,biology, and history. We observe a purity [25] enhancement when using the stemming algorithm in similitude when it was not used. They also demonstrate that the outcome were exacerbated when categorize documents of respective fields.

3 Proposed system

3.1 Information Extraction and Database Formation

Initially, a Google Scholar topic search is perpetrated to acquire a catalogue of research paper-description URL's of predominantly apposite papers. Then the system downloads each of these links and performs a basic pattern-based information-extraction to educe data about paper title. Information extraction is the errand of detecting particular segments of information from a paper, through acquiring appropriate structured [26] data from unstructured text. Peculiarly, It entails discovering a collection of substring from the paper, for every collection of identifying slots, considered as fillers. For extracting information from web pages other than natural language text, it may call as wrapper [12, 13, 14]. The prevalent slots considered by our recommender system are: title, authors, abstracts, keywords, related titles and references. Many other slots are also extracted like a conference/journal, ISBN, but are currently not consider by our recommender system. The extractor utilizes a simple pattern matcher uses pre-filler , filler and post-filler patterns for every slot, as mentioned in [26].The a that are available in every slot are then refined into a nonhierarchical sack of words and specimen depicted as a vector of sacks of words. A research paper's title and authors are added to its own related-title too, as a paper is definitely related to itself, and this permits overlay in these slots with paper related to it.

3.2 Profile Learning

Then, the user chooses and appraises a collection of training papers. By piercing for certain titles or authors, the user can shun inspecting the entire database. The user is catechized to bestow a discrete 1-10 appraising for every adopted title. The inductive learner presently employs by our recommender system is a sack-of-words naïve Bayesian text classifier [16] elongated to tackle a vector of sacks instead a solitary sack. Current experimental results [17,18] demonstrate that this method of text categorization

accomplishes same or some time better than many emulating approaches. Our system not strives to prognosticate the veracious numerical rating for the title, but it prefers simply a total ordering of these titles in sequence of proclivity. This process is then revised like a probabilistic binary categorization issue of prognosticating the probability that a research paper would be estimated as propitious instead inauspicious, where a user estimating about 1-5 is expounded as negative and 6-10 as positive. The veracious numerical rating for the training specimens are employed to examine the training specimens when reckoning the parameters of the paragon. Explicitly, we use a multinomial text model [18], in that a research paper is framed as an ordered excerpt of word incidents worn from the aforementioned lexis, L. The "naive Bayes" conjecture states that the probability of every incident is dependent on the paper set, but independent of the word's lexicon and physical position. For every set, s_j, and word, $w_k \in$ L , the probabilities, P (s_j) and P ($w_k | s_j$) should be evaluated from training data. Then the posterior probability of every set for a document, D, is calculated by Bayes rule:

$$\mathrm{P}(\mathrm{s_j} | \ \mathrm{D}) = \frac{P(s_j)}{P(D)} \prod_{i=1}^{|\mathrm{D}|} p(a_i | \ \mathrm{S}_j) \tag{1}$$

Where a_i is the ith word in the document, and IDI is the span of the document in words. After all, for any particular document, the preparatory P(D) is unvarying, If all the coveted factors formed ranks instead of a probability calculate, then this facet can be disregarded. A ranking is induced by sorting documents by their odd ratio, $P(s_1|D)/ P(s_0|D)$, where $s_1$represents the propitious set and s_0 represents the inauspicious set. An instance is categorized as positive if the odds are higher than 1, otherwise negative.

In our manifestation, as research papers are represented as a vector of a document, d_m, one for every slot, where l_m represents the m[th] slot. The probability of every word given the category and the slot, $P(w_k | s_j , l_m)$, should determine and the posterior category probabilities for a research paper, R, calculated by:

$$\mathrm{P}(\mathrm{s_j} | \ \mathrm{R}) = \frac{P(s_j)}{P(R)} \prod_{m=1}^{L} \prod_{i=1}^{|\mathrm{d_m}|} p(a_{mi} | \ \mathrm{S}_j, l_m) \tag{2}$$

Where L is the number of slots and a_{mi} is the i^{th} word in the m^{th} slot.

Parameters are evaluated from the training instances as follows. Every N training, research paper, R_e $(1 \le e \le N)$ are provided two real weights, $0 \le \alpha_{ej} \le 1$, depending on scaling it's users rating $r,(1 \le r \le 10)$: a propitious weight, $\alpha_{e1} = (r-1)/9$, and an inauspicious weight, $\alpha_{e0} = 1- \alpha_{e1}$. If a word emerges n times in an instance R_e, it is enumerated as occurring $\alpha_{e1}n$times in a propitious instance and $\alpha_{e0}n$ times in an inauspicious instance. The representation criterion is therefore evaluated as follows:

$$\mathrm{P}(\mathrm{s_j}) = \sum_{e=1}^{N} \alpha_{ej} \ / \ N \tag{3}$$

$$\mathrm{P}(w_k / s_j, l_m) = \sum_{e=1}^{N} \alpha_{ej} n_{kem} / L(s_j, l_m) \tag{4}$$

Where n_{kem} is the total of the number of the time word w_k arrives in instance $R_e$in slot l_m, and represents the total weighed length of documents in set s_j and slot l_m.

$$L(s_j, l_m) = \sum_{e=1}^{N} \alpha_{ej} | \ d_m | \tag{5}$$

The calculation complexity of the ensuing training algorithm is linear in the quantity of the training data. An augmented implementation provides definitively better performance even further. A profile can be partly elucidated by presenting the attributes much emblematic of a propitious or inauspicious rating. Strength estimates up to what extent a word in a slot is to arrive at propitiously rated research paper than an inauspiciously rated one, calculated as:

$$Strength\ (w_k, l_j) = log\ (P(\ w_k | \ s_1, l_j) / P(w_k | \ s_0, l_j) \tag{6}$$

3.3 Generating and Revising Recommendation

When a profile is learned, it is used to prognosticate the preferred ranking of the lingering research papers relied

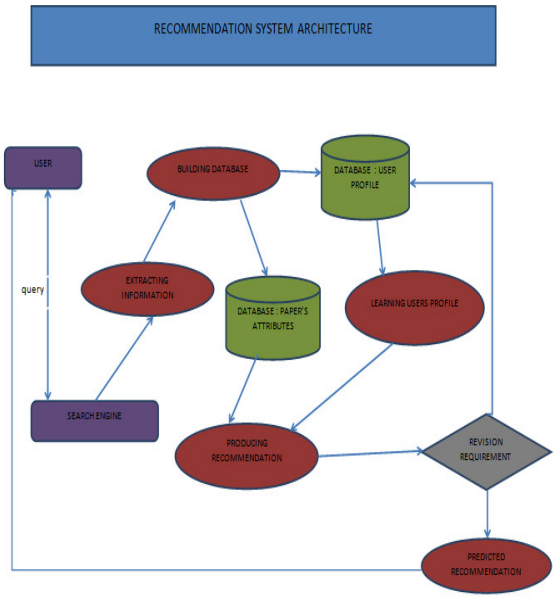

Figure. 1. The Model- Based Research Material Recommendation System

on the posterior probability of a propitious categorization, and highest ranking recommendation is provided to the user.Subsequently reviewing the recommendation, the user can allot their own rating to instances they suppose to be not correctly ranked and cling to the system to generate upgraded recommendations. This cycle can be iterated various times in order to accomplish the best result.This cycle can be repeated several times in order to produce the best results. Also, as new examples are provided, the system can track any change in a user's preference and alter its recommendations based on the addition

4 Conclusion

During writing a research paper, academicians and researchers must have to spend copious amount of time and efforts to acquire the latest and relevant research paper of their interest. This paper proposes an efficient research paper recommendation approach without user profiles. It takes a single paper as

input and then useful data are extracted and submitted to paper database to retrieve the similar research paper. These similar research papers, forms the training sets for our proposed system. These sets of training are categorized by the naïve Bayesian text classifier as positive or negative. On the basis of this the recommendation is generated.

5 Future Work

Future work in this field can be the implementation of grouping of specific subject research papers. We plan to demeanor a study in which each user selects its own training examples, get recommendations and provides personal rating after reading selected research papers. Another plan is to identify more better data extraction method by using deep learning and other machine learning methods to generate a more upgraded database repository for our candidate research papers.

REFERENCES

[1] Y. Liang, Q. Li and T. Qian, "Finding Relevant Papers Based on Citation Relations", Springer-Verlag Berlin Heidelberg, (2011), pp. 403–414.

[2] K. W. Hong, H. Jeon and C. Jeon, "User Profile-Based Personalized Research Paper Recommendation System", 8th International Conference on Computing and Networking Technology (ICCNT), IEEE (2012), pp. 134-138.

[3] K. Sugiyama and M.-Y.Kan, "Scholarly Paper Recommendation via User's Recent Research Interests",In Proc. of the 10th ACM/IEEE Joint Conference on Digital Libraries ,(2010), pp. 29–38.

[4] C. Wang and D. M. Blei, "Collaborative Topic Modeling for Recommending Scientific Articles", In Proc. of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,(2011), pp. 448–456.

[5] C. Nascimento, A. H. F. Laender, A. S. da Silva and M. A. Gonçalves, "A Source Independent Framework for Research Paper Recommendation".ACM, (2011) June 13–17, Ottawa, Ontario, Canada.

[6] M. Balabanovic and Y. Shoham. Fab: Content-based, collaborative recommendation. Communications of the Association for Computing Machinery, 40(3)"66"72,1997.

[7] M. Pazzani, J. Muramatsu, and D. Billsus.Syskill& Webert: Identifying interesting web sites. In Proceedings of the Thirteenth National Conference on Artificial Intelligence, pages 54"61, Portland, OR, August 1996.

[8] K. Lang. NewsWeeder: Learning to iternet news. In Proceedings of the Twelfth International Conference on Machine Learning, pages 331{339, San Francisco, CA,1995. Morgan Kaufman.

[9] G. Adomavicius, and A. Tuzhilin. "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions." Knowledge and Data Engineering, IEEE Transactions on 17, no. 6 (2005): 734-749.

[10] M.J. Pazzani and D. Billsus, "Content-based recommendation systems", in: P. Brusilovsky, A. Kobsa, W. Nejdl (Eds.), The Adaptive Web, Lecture Notes in Computer Science, vol. 4321, Springer-Verlag, 2007, pp. 325–341.

[11] Wendy Lehnert and Beth Sundheim. A performance evaluation of text-analysis technologies. AI Magazine, 12(3),81-94, 1991.

[12] DARPA, editor. Proceedings of the 6th Message Understanding Conference, San Mateo, CA, 1995. Morgan Kaufman

[13] N. Kushmerick, K. Weld, and R. Doorenbos. Wrapper induction for information extraction. In Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence, pages 729{735, Nagoya, Japan, 1997.

[14] M. E. Cali_ and R. J. Mooney. Relational learning of pattern-match rules for information extraction. In Proceedings of the Sixteenth National Conference on Articial Intelligence, Orlando, FL, July 1999.

[15] T. Mitchell. Machine Learning. McGraw-Hill, New York, NY, 1997.

[16] T. Joachims. A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization. In Proceedings of the Fourteenth International Conference on Machine Learning, pages 143{151, San Francisco, CA, 1997. Morgan Kaufman.

[17] A. McCallum and K. Nigam. A comparison of event models for naive Bayes text classi_cation. In Papers from the AAAI 1998 Workshop on Text Categorization, pages 41{48, Madison, WI, 1998.

[18] J.L. Herlocker, J.A, Konstan, L.G. Terveen, and J. Riedl, "Evaluating Collaborative Filtering Recommender Systems," ACM Transactions on Information Systems, 22(1), pp. 5-53, 2004.

[19] J. Beel, B. Gipp, S, Langer, and M. Genzmehr, "Docear: An Academic Literature Suite for Searching, Organizing and Creating Academic Literature", Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries (2011), 465–466.

[20] J. Beel, S. Langer, M. Genzmehr, and A. Nürnberger, "Introducing Docear's research paper recommender system", Proceedings of the 13th ACM/IEEE-CS joint conference, (JCDL '13), 2013, pp.459-460.

[21] J. Lee, K. Lee, and J. G. Kim, "Personalized Academic Research Paper Recommendation System.", arXiv preprint arXiv:1304.5457(2013).

[22] P. Lakkaraju, S. Gauch, and M. Speretta, "Document similarity based on concept tree distance.", Proceedings of the nineteenth ACM conference on Hypertext and hypermedia. ACM, 2008, pp. 127-132

[23] B. Gipp, J. Beel, and C. Hentschel, "Scienstein: A Research Paper Recommender System", In Proceedings of the International Conference on Emerging Trends in Computing (ICETiC'09), Virudhunagar (India), January 2009, pp. 309–315.

[24] Nikhat Akhtar, Prof. (Dr.) Devendera Agarwal, "A Literature Review of Empirical Studies of Recommendation Systems" International Journal of Applied Information Systems (IJAIS) USA , Volume 10, No. 2, Pages 6 – 14, December 2015, ISSN 2249 - 0868, DOI : 10.5120/ijais2015451467.

[25] J. Sadiku and M. Biba, "Automatic Stemming of Albanian Through a Rule-based Approach", Journal of International Research Publications: Language, Individuals and Society, Vol. 6, 2012.