Personalization in Collaborative Fusion based Enterprise Information retrieval

Dinesha L, Kumaraswamy S

Abstract: ***Due to data spread across various heterogeneous data stores, information retrieval in Enterprise data stores is always challenging compared to web based retrieval systems. We have proposed a collaborative fusion based information retrieval in [1] using the observation on similar users tends to prefer similar search results. The solution applied three dimensions of user similarity, document similarity and user to documents affinity to a collaborative information fusion based retrieval. The work also proposed active feedback based search result revision to get highly relevant results. But the work did not have any provision for personalization and could not handle the cold start problems. Without consideration for cold start problems, the user to document affinity cannot be modeled accurately as the result, the collaborative fusion process is affected. In this work, we improve our earlier solution of collaborative fusion based information retrieval with consideration for user personalization and solution for cold start problems. The solution is based on query refinement using the information hidden in enterprise messaging systems. A user profile is built as vector of concepts using the information in enterprise messaging systems and this user profile concept vector is used to refine the query in way to personalize the results and avoid cold start problems. Compared to approach in [1], the proposed query refinement based personalization is able to increase the relevancy accuracy by 10% as obtained from experimental results.***

Keywords : IW, NIST, QC, WC

I. INTRODUCTION

A typical enterprise generates large volumes of data about customer, products, sales, transactions, employee etc. These data are stored across various data stores web servers, file servers, wiki, databases etc. With consideration for security, the data is controlled with various levels of access rights. Retrieval of information from enterprise data stores is considerably more difficult than web information retrieval. Following are some of the challenges in enterprise information retrieval as shown in Table 1.

Manuscript published on November 30, 2019.
***** Correspondence Author

Dinesha L*, Research Scholar, Computer Science and Engineering, Sri Siddhartha Academy of Higher Education, Tumakuru, India. Email: ldinesha.ssit@gmail.com
Kumaraswamy S, Professor, Computer Science and Engineering, Sri Siddhartha Institute of Technology, Tumakuru, India. Email: kumar.aruna@gmail.com

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an open access article under the CC-BY-NC-ND license http://creativecommons.org/licenses/by-nc-nd/4.0/

Multiple data formats	Data is available in different content formats like web pages, pdfs, word documents,emails, databases, wiki etc.,
Distributed	The data stores are distributed. With recent adoption of cloud, the data is globally distributed.
Access control	Various access rights in different modes of read/write etc., on the data.

Table 1: Challenges in Enterprise information retrieval

An enterprise information retrieval system must be designed to respond to user request by looking for relevant information in the data stores and respond with ranked relevant results , such that it meet the needs of the users. Different from typical enterprise systems built on indexing and retrieval from index, we proposed an enterprise information retrieval based on user and document similarity in [1]. The approach is most similar to collaborative product recommendation. Based on fusion of three important information of user similarity, document similarity, user to document preference, a collaborative fusion based recommendation was proposed. We proposed a new search idiom called concept. The document similarity was modeled in terms of similarity in the number of concepts shared by the documents and the user similarity is modeled in terms of concept preference similarity between the users. Also a feedback based tuning was proposed to fine tune the search results. The work had following open issues

1. Personalization of search result based on user profile and user search history
2. Cold start problems in Collaborative Fusion.

This work addresses these two problems and evolves the collaborative fusion based information retrieval proposed in earlier work [1]. The contributions in this work are as follows

1. Learn user profiles both statically and dynamically and represent the user profile in terms of concepts. The represented user profile concepts are also used in collaborative fusion to improve the relevancy of the results.
2. Cold start problems are avoided using concept projection based on user profile concepts.

II. RELATED WORK

The existing solutions for personalization and avoiding cold start problems in information retrieval is discussed below Author in [2] proposed a model to represent the user's activities during information retrieval. User set the objectives for a session, based on which the information retrieval must work and the proposed model was implemented in a prototype called METIORE. Natural language is applied in this work to express the objectives. Each document is then evaluated against objectives using Naïve Bayesian rule. The most relevant document matching to user's objectives is given as result. User's social web interactions are explored to extract his preferences and interests and this information is used to build the user profile and search personalization is done using this built user profile in [3]. A vector of terms corresponding to user interest is created and this represents the user profile. To accommodate changing user requirements, temporal weights adjustment is done on user profile. This temporal weight is based on user interest extracted from social media. The user personality is represented in form of class model in [4]. The user's affinity to classes is modeled in form of weight vector. In absence of history information, weight for new class cannot be calculated. Cold start problem is avoided by using a class similarity based algorithm to estimate the weight for new class even in absence of historical information. This solves the cold start problem. The weights for classes are updated periodically on arrival of information. Learning user profile from user interactions with social content and integrating this profile for personalization in information retrieval is proposed in [5]. User profile is built based on user preferences and this information is used to fine tune the retrieval process. Based on user profile additional information is augmented to the initial query of the user. This helps in personalized search. In [6] user profile modeling depending on user's context is used for personalized image retrieval system. The user profile is build based social network attributes such as user comments, ratings, tags and preferences. The user profile model is built using a fuzzy logic based profiling. In this the user initial description of rating concepts and context of his interest are weighted with a score calculated using a fuzzy model. This score is proportional to the preference degree of user to each concept. A personalized information retrieval based on social network is proposed in [7]. A new measure for measuring similarity between users in social network called interestedness measure was proposed in this work. The retrieval results are personalized based on this measure. Use of citation networks to provide a low cost personalized information retrieval is explored in [8]. The solution was implemented on an evaluation framework called PERSON. Web search was personalized using user profile built from twitter data was suggested in [9] The data from twitter is converted to various features using statistical language modeling algorithm. The performance of web information retrieval is improved due to this approach. Use of neural embedding to personalize twitter search is explored in [10]. User interest is learnt effectively using neural embedding approach. The user interest is represented as a vector of weighted words. Different from previous works, authors in [11] propose an approach to represent the documents in a form of personalized document representation suitable for personalized information retrieval. The document representation is based on users' activities in social tagging system. Matrix factorization algorithm is used in this work. The document is represented in form suitable for matching query using the matrix factorization algorithm. Author in [12] used personalized query expansion as a way to improve retrieval efficiency. The user profile is enriched with help of external corpus and this is used for improving the user query with constructs for personalization. Two new algorithms are proposed in this work for query expansion. Word embedding technique with weighted terms is used in those two algorithms. Annotations were used as resources to learn the user profile in [13]. Based on these user profiles query expansion process is fine tuned. In [13] author proposed a novel query expansion framework based on individual user profiles mined from the annotations and resources the user has marked. User profile is learnt in form of work embedding on a connected graph of terms present in the documents. Latent search is done on this connected graph to learn the topic tag associations. Author in [14] proposed a personalized document ranking in retrieval based on social contexts. Social annotations of documents are used as sources to learn the user profile. Work in [15] is based on the concept of similarity in judgment about relevancy by the users in same group. Based on this concept the web search is personalized. The information needed to build the profile is learnt through web click through information. The personalization scheme proposed in this work considers document preferences, user social groups and other users in the network. This work also proposes a measure to quantity the amount of personalization considering both the user and the query.

Active learning based solution to solve the cold start problem in case of recommendation system is proposed in [16]. The prediction of rating for users is based on feedback rating, users past rating and items attributed. In case of new user where past rating is not available, item attributes and feedback rating are used to learn the user profile and personalize the recommendation. Cold start problem is addressed through use of local collective embedding in [17]. The work applies matrix factorization on parameters such as item properties, user past experiences to learn the collective embedding's. Item description is used as only source to predict the recommendation when user past experiences are absent. Cold start problem is solved applying imperialist competitive algorithm in [18]. It is two stage processes. The tag extracted from user past data are clustered using K-medoids algorithm in offline stage. The recommendation for user is done in an online process using the information built in offline mode.

III. PROPOSED SOLUTION

Most of existing solutions proposed for personalization and solving cold start problems are based on integration of social media. User profile is built using the social media information. Based on the user profile, query expansion or result ranking is done to meet challenges in personalization and cold start problems. In enterprise environment, this integration with social media does not work.

In enterprise environment, this gap can be bridged with enterprise user details like department, year of experience and enterprise communication platforms like emails, document ownership and co-authorship information.

The user profile is modeled in terms of significant concepts weighted with temporal coefficient to signify the freshness. A new fusion mechanism is added to collaborative fusion rule proposed in [1] to personalize the search and to solve cold start problems.

A. *User Profile Modeling*

In this work the user profile model is built based on contents from enterprise communication systems like Email, shared board, document authored and co-authored. The user profile model is built based on vector of concepts. The concept is learnt using the following procedure.

The contents such as email, shared board, document authored and coauthored are split to chronological bins with each bins representing a month. In the contents within the bin the standard stop words are removed. Natural language processing is done to identify only nouns and adjectives in document. The identified nouns and adjectives are stemmed and duplicate words are removed to create unique concept. Each concept corresponding to bins are associated with a temporal linear degrading weight with most recent bins concepts given higher weight than others. The temporal linear degrading weight is modeled as below in Fig. 1 linear decreasing function.

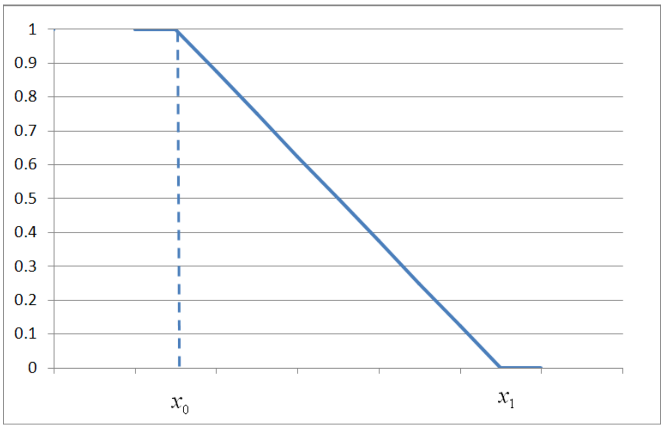

Fig. 1: Linear decreasing function

The x_0 is the most recent bin and the x_1 is the oldest bin. The weight for concepts in bin is calculated as

$wc(x) = f(x) * IW$ - (1)

Where, IW is the configured max weight value for a concept. The concept extraction procedure is detailed in the flowchart Fig. 4.

As the result of concept extraction a vector of concepts is created with weights as shown below

$$UP = \{< C_1, W_1 >, \quad \ldots < C_n, W_n >\} - (2)$$

B. *Query Refinement*

The architecture of the collaborative fusion enhanced for personalized and solving cold start problem is given in Fig. 2. The main change from earlier solution in [1], the search query from user is refined in Fig. 3 instead of directly mapping the search query to concept.

The query to concept mapping is done by selecting the k-gram in the concept list matching to query provided by the user. The query to concept score (QC) is evaluated for each concept in entire candidate list and the concepts are ranked based on the score. Before evaluation, the entire alias for the query is generated.

$$QC(Q,C) = \sum_{i=1}^{N} f(Q_i, C) - (3)$$

Where N is the number of alias for the query Q including Q. $f(Q_i, C)$is the function which provides value of 0 or 1. The function outputs 1 when Q_i *and* C are similar. The QC is calculated for each concept and the concept with the highest QC is taken for refinement. Along with this concept, the concepts from user profile are fused to get the refined query to be used for collaborative fusion procedure discussed in [1].

The fusion is done as follows

1. The intersection of concepts from the user profile and QC ordered is taken.
2. If the intersection list is empty, a new list is constructed by joining the QC ordered list to the top K weighted concepts from user profile which is related to the QC ordered list by value greater than a threshold

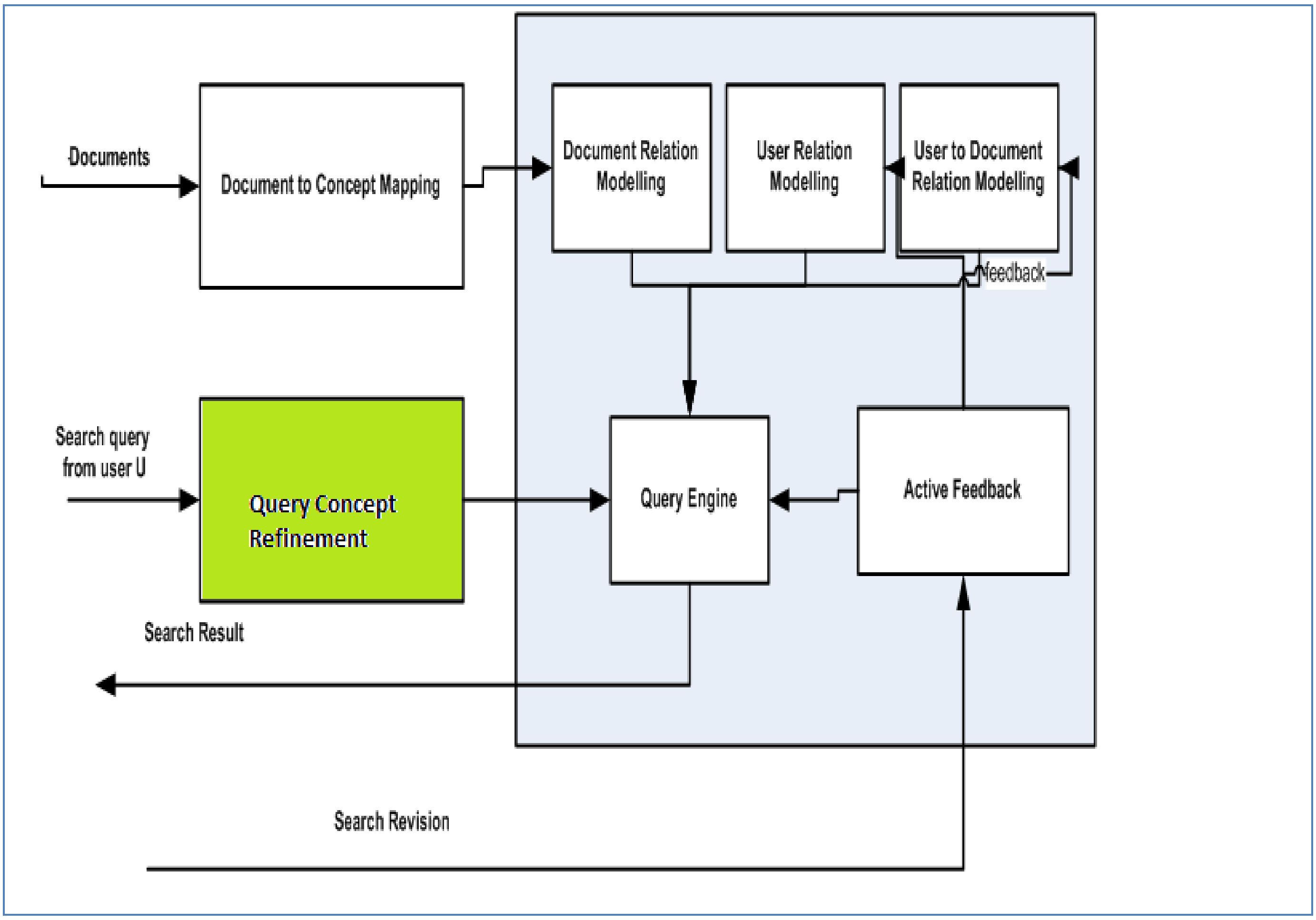

Fig. 2: Collaborative Fusion

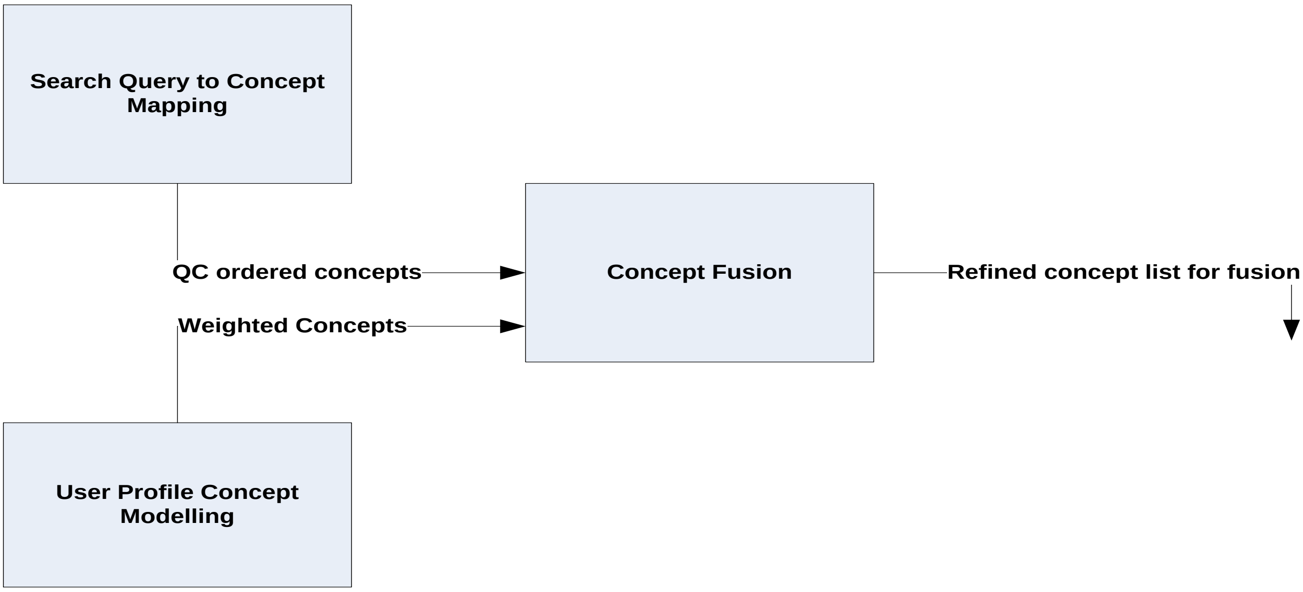
Search Query to Concept Mapping
QC ordered concepts
Weighted Concepts
Concept Fusion
Refined concept list for fusion
User Profile Concept Modelling

Fig. 3: Query Refinement

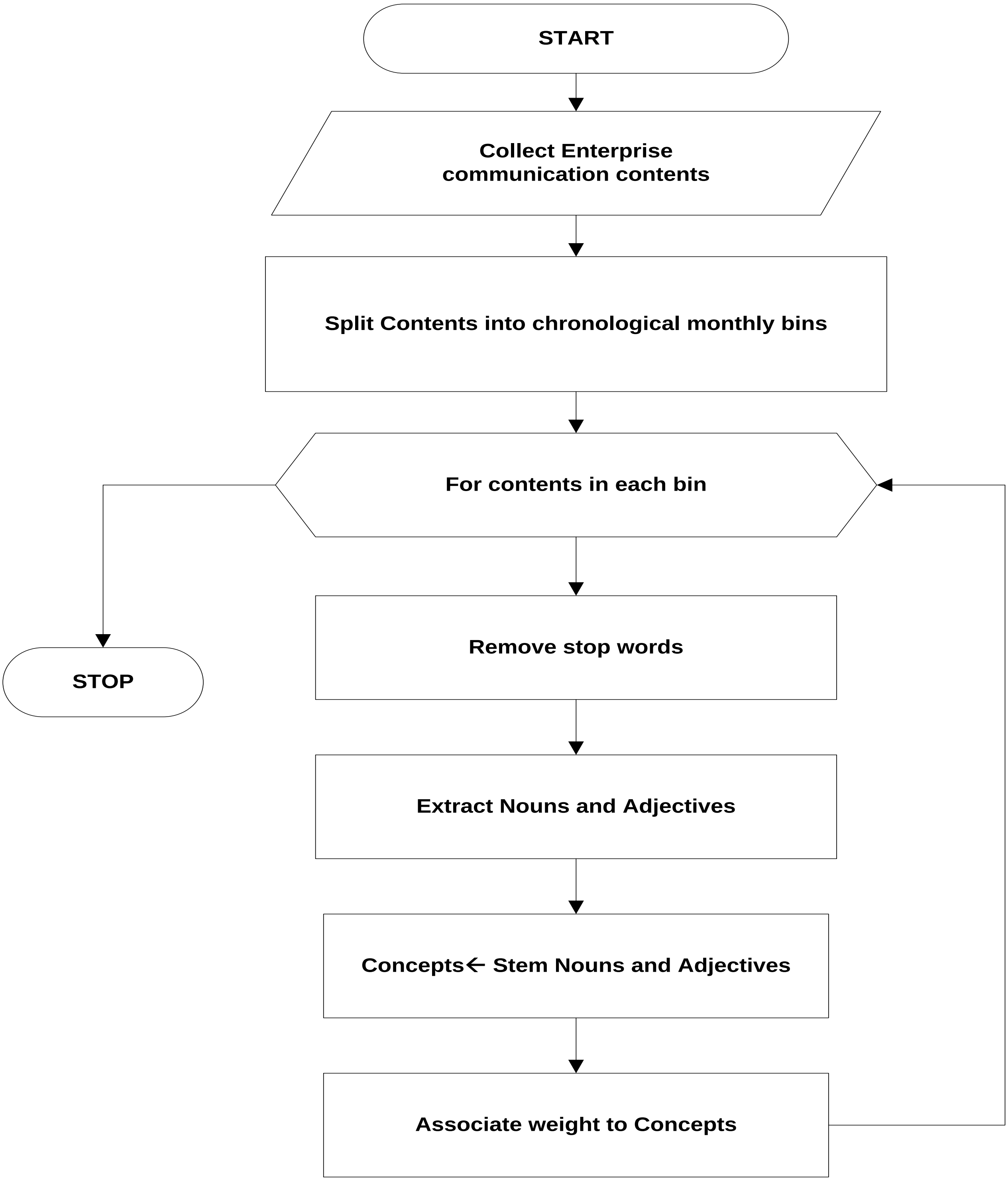
START
Collect Enterprise communication contents
Split Contents into chronological monthly bins
For contents in each bin
Remove stop words
STOP
Extract Nouns and Adjectives
Concepts← Stem Nouns and Adjectives
Associate weight to Concepts

Fig. 4: Concept Extraction Procedure

The relation is calculated as joint frequency of occurrence of the concepts in same document. It is given as

$$R(a,b) = \frac{\sum D(a) \cap D(b)}{|\sum D(a) + \sum D(b)|} \text{ - (4)}$$

Where a and b are the concepts and the relation is given as number of documents in which both a and b are jointly present divided by total number of documents in which a is present and b is present. Due to addition of related concepts, the cold start problem is avoided.

IV. RESULTS

The proposed system for personalized information retrieval was compared against [1]. Dataset used for evaluation is TREC Enterprise Track 2007[19]. TREC dataset is most used for testing the information retrieval algorithms. National Institute of Standards and Technology (NIST) maintains this dataset. The goal of TREC enterprise track is to conduct experiments with enterprise data that reflect the experiences of users in real organizations. The dataset have 370715 documents with different types. A set of 50 queries was designed and the expected results and their ranking order is established before the start of the test. The performance is measured in terms of

1. R^2 measurement
2. Hold out Estimated accuracy
3. Precision
4. Recall
5. Ranking Analysis

R^2measurement is one for measure of accuracy of statistical models. It is calculated as

$$R^2 = \frac{\sum_{i=1}^{n}(\check{y} - \bar{y})^2}{\sum_{i=1}^{n}(y - \bar{y})^2} = \frac{SS\ Predicted}{SS\ Total} - (5)[1]$$

Where n is the number of data instances, $\bar{y}$ is mean of actual values of the instances, $\check{y}$ is the predicted value of the data instance i.

Hold out Estimated accuracy is measured as

$$acc_h = \frac{1}{h} \sum_{vi,yi \in Dh} \sigma(vi, yi) - (6)\ [1]$$

Where D_h is the subset of data set D of size h and $\sigma(v,y) = 1$ if v=y and 0 otherwise. $vi$is the predicted value of instance i and yi is the actual value of instance i.

Precision is the ability of the system to present only relevant items. It is measured as

$$P = \frac{|R_a|}{A} - (7)[1]$$

Where R_a is the number of items relieved and A is the total number of items retrieved in response to search query.

Recall is the ability of the system to present all relevant items. It is measured as

$$R = \frac{|R_a|}{|R_m|} - (8)[1]$$

Where R_m is the total number of relevant items in the document set.

Document ranking analysis is done to measure the effectiveness of ranking in the proposed system.

R^2 Measurement values for proposed and the [1] is given below Table2.

X	[1]	Proposed
SS Predicted	10674	11294
SS Total	12412	12412
R^2	85.99%	91%

Table 2: R^2 Measurement

The proposed solution is able to increase the R^2 by 5.01% compared to collaborative fusion proposed in [1].

The Hold out Estimated accuracy for the proposed and [1] is given below Fig. 5.

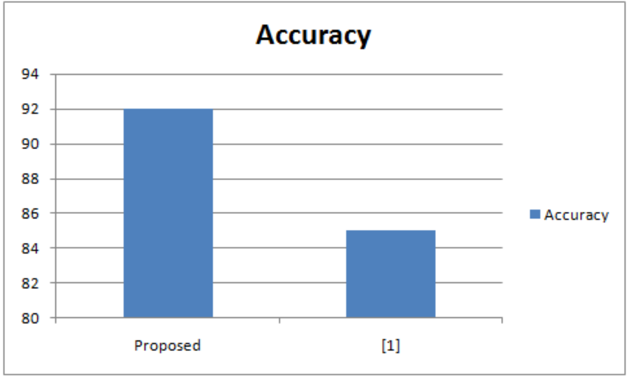

Fig. 5: Hold out accuracy

The precision and recall for the proposed and [1] is below Table 3.

X	Proposed	[1]
Precision	0.91	0.85
Recall	0.85	0.828

Table 3: Precision and Recall

The precision and recall of the proposed solution is higher than fuzzy based solution [1] as shown in Fig. 6.

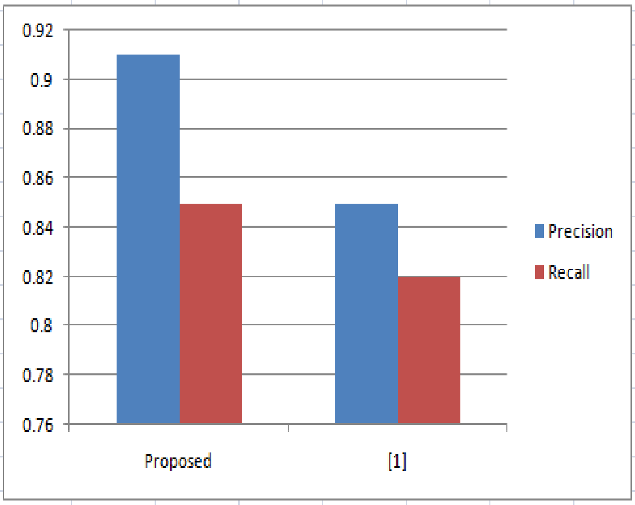

Fig. 6: Precision and Recall comparison

The ranking of the relevant documents for the first 50 queries are split into following categories in intervals of 10.

- 1-10 →A
- 11-20→B
- 21-30→C
- 31-40→D
- >40-> E

Retrieval Number: D8539118419/2019©BEIESP
DOI:10.35940/ijrte.D8539.118419
Journal Website: www.ijrte.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

- Not retrieved → F

The document frequency is calculated in each category and the result is given below Table 4 for both the proposed and solution [1].

X	Proposed	[1]
A(1-10)	83.83%	73.83%
B(11-20)	15.89%	7.44%
C(21-30)	1.13%	2.13%
D(31-40)	1.06%	1.06%
E(41)	1.06%	1.06%
F(Not retrieved)	7.02%	12.02%

Table 4: Relevant Analysis

Highly relevant ranked results are provided as result in the proposed personalization solution compared to collaborative fusion proposed in [1].

V. CONCLUSION

In this work we have enhanced the Collaborative Fusion based information retrieval proposed in [1] with personalization and solution for cold start problems. The solution relied on building the user profile in terms of temporal weighted concepts extracted from enterprise communication systems. The extracted concepts are then fused with search query derived concepts to refine the query. The refined query is able to personalize the search result as it seen from category wise ranking accuracy. Due to personalization, the relevancy of search results improved. Also use of enterprise messaging system to build user profile is able to avoid cold start problems. The relevancy of search results increased by 10% compared to work [1] proving the efficiency of the proposed query refinement based personalization.

REFERENCES

1. Dinesha L, and Kumaraswamy S "Enterprise Information Retrieval using Collaborative Fusion and Active Feedback",IEEE Explore Digital Library, 2019
2. Estimation Algorithm for the Personalized Enterprise Search Engines, Journal of Computational Information Systems 10 (2014) 1903-1910.
3. R. Fakhfakh, G. Feki, A. B. Ammar and C. B. Amar, "Personalizing information retrieval: A new model for user preferences elicitation," *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, Budapest, 2016, pp. 002091-002096.
4. R. Fakhfakh, A. Ben Ammar, C. Ben Amar, "Fuzzy user profile modeling for information retrieval", *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval*, pp. 431-436, 2014.
5. S. A. Tabrizi, A. Shakery, M. A. Tavallaei and M. Asadpour, "Search Personalization Based on Social-Network-Based Interestedness Measures," in *IEEE Access*, vol. 7, pp. 119332-119349, 2019. doi: 10.1109/ACCESS.2019.2935425.
6. S. A. Tabrizi, A. Shakery, H. Zamani, and M. A. Tavallaei, ''PERSON: Personalized information retrieval evaluation based on citation networks,'' Inf. Process. Manage., vol. 54, no. 4, pp. 630–656
7. A. Younus, C. O'Riordan, and G. Pasi, ''A language modeling approach to personalized search based on users' microblog behavior,'' in Advances in Information Retrieval. Cham, Switzerland: Springer, 2014, pp. 72
8. S. Samarawickrama, S. Karunasekera, A. Harwood, and R. Kotagiri, ''Search result personalization in Twitter using neural word embeddings,'' in Big Data Analytics and Knowledge Discovery. Cham, Switzerland: Springer, 2017, pp. 244
9. M. R. Bouadjenek, H. Hacid, M. Bouzeghoub, and A. Vakali, ''PerSaDoR: Personalized social document representation for improving Web search,'' Inf. Sci., vol. 369, pp. 614–633, Nov.
10. D. Zhou, X. Wu, W. Zhao, S. Lawless, and J. Liu, ''Query expansion with enriched user profiles for personalized search utilizing folksonomy data,'' IEEE Trans. Knowl. Data Eng., vol. 29, no. 7, pp. 1536–1548, Jul.
11. D. Zhou, S. Lawless, V. Wade, "Improving search via personalized query expansion using social media", *Inf. Retrieval*, vol. 15, no. 3/4, pp. 218-242, 2012.
12. M. R. Bouadjenek, H. Hacid, M. Bouzeghoub, "Sopra: A new social personalized ranking function for improving web search", *Proc. 36th Int. ACM SIGIR Conf. Res. Development Inf. Retrieval*, pp. 861-864, 2013.
13. A. Dridi and Y. Slimani, ''SonetRank: Leveraging social information for personalized search,'' Social Netw. Anal. Mining, vol. 7, no. 1, p. 16,.
14. Yu Zhu, Jinhao Lin, Shibi He, Beidou Wang, "Addressing the Item Cold-start Problem by Attribute-driven Active Learning",Information Retrieval (cs.IR),2018
15. M. Aharon, O. Anava, N. Avigdor-Elgrabli, D. Drachsler-Cohen, S. Golan, and O. Somekh, "Excuseme: Asking users to help in item cold-start recommendations," in Proceedings of the 9th ACM Conference on Recommender Systems. ACM, 2015, pp. 83–90.
16. M. H. Jafari, G. T. Tabrizi and M. Jalali, "Solving cold start problem in tag-based recommender systems using discrete imperialist competitive algorithm," *2014 International Congress on Technology, Communication and Knowledge (ICTCK)*, Mashhad, 2014, pp. 1-7
17. Martin Saveski and Amin Mantrach. 2014. Item Cold-start Recommendations: Learning Local Collective Embeddings. In Proceedings of the 8th ACM Conference on Recommender Sys- tems (RecSys 14). ACM, New York, NY, USA, 8996. https://doi.org/10.1145/2645710.2645751
18. M. H. Jafari, G. T. Tabrizi and M. Jalali, "Solving cold start problem in tag-based recommender systems using discrete imperialist competitive algorithm," *2014 International Congress on Technology, Communication and Knowledge (ICTCK)*, Mashhad, 2014, pp. 1-7
19. https://dblp.org/db/conf/trec/trec2007

AUTHORS PROFILE

Mr. Dinesha L is Research Scholar in Department of Computer Science and Engineering at Sri Siddhartha Academy of Higher Education, Tumakuru, India. He has published papers in international conference and journals. His area of interest is Information Retrieval, Big Data and Data Mining.

Kumaraswamy S is currently working as an Professor in the Department of Computer Science and Engineering, Sri Siddhartha Institute of Technology, Tumakuru, India. He received his Ph.D from Bangalore University. He has 19 years of teaching experience. He published more than 12 papers. His research interest is in the area of Data mining, Web mining, Semantic web and cloud computing.