Semantic Term-Term Coupling-Based Feature Enhancement of User Profiles in Recommendation Systems

Mona Tanwar, Amity Institute of Information Technology, Amity University, Noida, India

Sunil Kumar Khatri, Amity University Tashkent, Tashkent City, Uzbekistan

https://orcid.org/0000-0003-4373-9000

Ravi Pendse, University of Michigan, USA

ABSTRACT

A content-based recommender system is a subclass of information systems that recommends an item to the user based on its description. It suggests items such as news, documents, articles, webpages, journals, and more to users as per their inclination by comparing the key features of the items with key terms or features of user interest profiles. This paper proposes the new methodology using Non-IIDness-based semantic term-term coupling from the content referred by users to enhance recommendation results. In the proposed methodology, the semantic relationship is analyzed by estimating the explicit and implicit relationship between terms. It associates terms that are semantically related in the real world or are used interchangeably such as synonyms. The underestimated features of user profiles have been enhanced after term-term relation analysis, which results in improved similarity estimation of relevant items with the user profiles. The experimentation result proves that the proposed methodology improves the overall search and retrieval results as compared to the state-of-the-art algorithms.

KEYWORDS

Collaborative Filtering, Content-Filtering, Feature Enhancement, Non-IIDness Learning, Recommender Systems, Semantic Term-Term Analysis

INTRODUCTION

In present scenario, information is continuously being generated with a velocity which is much higher than our processing capacity. If utilized properly, it can make a great difference to the world in all spheres. Various e-commerce and information sites have an abundance of products or items such as apparels, footwear, accessories, books, journals, web pages, movies, songs, hotels, restaurants, grocery and so on. From millions of items space, it is tedious and time consuming to find suitable items for the users. Recommendation systems suggest personalized items from the possible options to the users by understanding their requirement, preferences and inclinations.

Recommendation system is a significant application of big data and is involved in online e-commerce sites and business, news sites, social media, mobile applications, online journals and

digital libraries, etc. Exploration of recommender systems has gained attention in many fields such as information retrieval, social networking, data mining and machine learning. Efforts are being made to improve the recommendation's accuracy by considering factors such as social relationships, user reviews/comments, cross-domain recommendations, etc. apart from grouping similar categories of users or items.

The user's behavioural information such as purchase patterns, ratings, likes/dislikes and feedback/comments/reviews are analysed to make recommendations. Useful recommendations are essential not only to ensure good user experience but also to ensure good business from the vendor's perspective. An efficient recommender system should recommend appropriate products or services to appropriate people.

In content-based recommender systems, for content-based filtering of text-based contents like journals on the Web and Digital Libraries, news, articles, web pages, etc. keywords or terms are used for item description and for building a user profile to indicate the user's inclination. The user's feature vector is built considering the content that has been of interest to the user in the past. Content-based filtering or cognitive filtering recommends items by comparing the content of the items to a user profile. In the state-of-the-art approach, the user inclination features are retrieved based on the occurrence frequency of terms. Better representation strategies are required to incorporate content-based recommender systems with semantic intelligence which goes beyond the simple syntactic evidence of user inclinations provided by terms (Lops et al., 2011).

The proposed approach is based on Non-IIDness learning which refers to understanding, modelling, analysing and representing non-IID data (not independent and identically distributed data). Coupling and heterogeneity are the important aspects of Non-IIDness. Usually, the terms of the feature vectors are considered to be independent but there are couplings between the terms. If the couplings between the features or terms are analysed, unravelled, mathematically formulated and estimated, better recommendation results are expected. There are intricate semantic couplings between the terms which if incorporated in the user profiles, better recommendations can be achieved. For instance, many terms are semantically related in the real world though on similarity estimation simply based on co-occurrence of terms, the relation may not be inferred (Cheng et al., 2013). Also, there are few terms which are synonyms or are used interchangeably but have differing feature weights in the user profiles as the occurrence of these terms in the items of interest of the users differs. In that case, the items which should semantically match better to the user profiles do not get their due weightage since the relevant features have underestimated weights in the user profiles. Hence, along with the co-occurrence frequency (explicit relation) of terms, the implicit relationship between terms also needs to be estimated to infer the semantic closeness between terms. In the proposed work, the above issues have been addressed by learning the term coupling relationships (i.e., intra-relations and inter-relations) from the items of interest of the users to infer the semantic relationship between features or terms of the user profiles. The highly semantically related terms are coupled, and the semantic relationships are embedded in the user profiles. This is done by enhancing the underestimated relevant features that are highly semantically related to other relevant features based on the assumption that highly semantically related terms should have comparable or similar significance. During experimentation, it has been observed that the appropriate items match better with user profiles after enhancement of features.

BACKGROUND AND RELATED WORK

The recommendation techniques that exist can be broadly categorized into content-based techniques (Lops et al., 2011), collaborative filtering techniques (Linden et al., 2003), hybrid techniques (Balabanovic & Shoham, 1997) and personalized techniques. The content-based recommendations are based on the description of items preferred earlier by the user. The attributes of a user specific profile which represents the inclination of the user are matched up with the attributes of the objects

or the items to make recommendations. In user-user collaborative filtering, the like-minded users or the users having the same interests are identified by the kind of ratings they give to the items or by what items they purchase, view or like. Recommendations are made based on the assumption that like-minded users would like the same items. The recommendations on e-commerce sites such as Amazon are based on item-item collaborative filtering where products that are mostly purchased along with the products of user interest are recommended to the user. Top ranked items or products are also recommended (Linden et al., 2003). The model-based collaborative filtering techniques are based on latent factor and matrix factorization. For instance, Netflix had announced a contest to improve its recommendation system. The solution was given using matrix factorization which reduced the RMSE i.e., Root Mean Square Error by 10 percent (Koren et al., 2009). Many hybrid techniques which are a fusion of other recommendation techniques also exist and non-personalized techniques such as recommending the top-rated products irrespective of what inclination the user has are also in practice.

The efficiency of content-based recommender systems highly depends upon how effectively the interest features are being extracted from the content referred by users. Several content-based recommender systems have been proposed in the research domain such as Quickstep system which recommends on-line academic research papers (Middleton et al., 2004). User inclination profiles are constructed by correlating the papers browsed in the past with their classification. The user profile holds a set of values reflecting the user interests and topics. The item profiles are matched by calculating a correlation between the user profile's top three topics of interest. Foxtrot is an extension to the Quickstep system (Middleton et al., 2004). Along with Web page recommendation interface, it implements an interface for profile visualization, an interface for search of papers and an email notification.

Based on similarity of research content of researchers, an interdisciplinary collaborator recommendation method has been presented. The textual features that represent the researcher's inclination or interest area have been calculated. A pair-wise similarity matrix is constructed exploiting the existing social networks with content-based similarity (Araki et al., 2017).

Citeseer uses word or term information and analyses common citations in the papers to perform a scientific literature search for the user (Bollacker et al., 1998). The effect of modelling a group of researcher's respective past work considering their citation and reference papers has been examined in making recommendations of scholarly papers to the researchers. The key part of the model is the enhancement of the researcher's profile derived from their previous work. The researchers have been categorized into junior and senior researchers based on the number of papers they have published (Sugiyama & Kan, 2010).

Many keyword-based content-based recommender systems exist for various categories of items such as movies, jokes, songs, etc. One problem is that this approach lacks intelligence (Lops et al., 2011). Keyword based approaches have limitations because of which more advanced characteristics are needed. For example, if the user likes "French impressionism", the keyword-based approaches will search for sources having terms "French" and "Impressionism". Documents having "Claude Monet" or "Renoir exhibitions" will not be recommended though they would be relevant to the user. Advanced approaches such as semantic analysis or machine learning for user profile construction overcome these limitations. Many content-based recommender systems are interpreted as text-classifiers which are built from training sets categorized as positive or negative. Reliable syntactic evidence of user inclinations is guaranteed by training sets having large number of examples.

The state-of-the-art research on recommender systems is based on the IIDness of the users and items and hence the methods and models for recommendations are IID too. The very low-level non-IID (Non-identical and Non-independent data) information of the items and users has been ignored (Cao, 2014). For Example, the model based collaborative filtering is based on MF or Matrix factorization in which if the lower-level properties of the items are not considered, it would lead to inferior results. A Non-IIDness based CGMF i.e. Coupled Group-based Matrix Factorization model

which incorporates couplings between/within users and items such as intra-couplings, item-couplings and user-item couplings has been proposed and discussed (Li et al., 2015).

A theoretical framework based on non-IIDness has been discussed deeply to understand the intrinsic nature of the problems and the complexities of recommendations (Cao, 2016). Approaches to various complexities associated with recommender systems such as cold-start, sparsity of available data, cross domain recommendations, group-based recommendations and shilling attack related issues have been addressed (Cao, 2016). Behaviour of users in viewing or commenting of products has been modelled to understand the user behaviour well. Coupling relationships between groups of products and items have been modelled for recommendations incorporating the low-level driving forces into ratings estimation (Fu et al., 2015; Yu et al., 2013).

For the proposed work, reference has been taken from the coupled term-term relation analysis for document clustering approach where document similarity has been analysed involving both explicit and implicit semantic couplings (Cheng et al., 2013). In the previous work, we proposed a framework for user profile learning, modelling and enrichment. The user profiles learn from three sources: similar users, similar items and semantic relationship between terms (Tanwar & Khatri, 2019).

The user profiles in most of the existing recommendation techniques lack in going beyond co-occurrence information of terms adding less semantic value to the user profiles. The closely semantically related terms to the features of user profile as well as the terms used interchangeably are often neglected leading to inferior recommendations. This problem has been well addressed in the proposed approach. Semantic information and relationships or couplings between such terms from the content of user interest have been extracted and incorporated in content filtering to make content-based recommendations.

PROPOSED METHODOLOGY

The proposed methodology for incorporating the user-profiles with semantic intelligence based on term-term coupling learning is given by the block diagram as shown in figure 1. The steps involved such as user profile construction, item representation, term-term relation analysis and user feature enhancement have been explained in detail. The block architecture as given in figure 1 depicts the step-by-step process of item representation vector construction from items, user profile building from items referred by user, learning and enhancement of relevant user features by term-term relation analysis and recommending top matching items to active users.

In proposed work, the user profile construction, feature vector construction for candidate papers, semantic analysis by inferring intra-relations and inter-relations have been referred from previous work done on proposing a framework for user profile learning and modelling (Tanwar & Khatri, 2019). The proposed algorithm for fetching Term- Term semantic relation is given below:

3.1. Detailed Description of Proposed Algorithm

A.   User Profile Construction

Generally, for profile construction the content-based systems use simple retrieval models like keyword matching or the Vector Space Model (VSM) with the basic tf*idf weighting scheme (Lops et al., 2011). The set of terms or features are retrieved by some standard NLP operations and techniques such as tokenization, stop words removal and stemming. Term frequency (tf) is the frequency of term occurrence with respect to all terms in a document and document frequency (df) is the number of documents containing the particular term with respect to the total number of documents (Lops et al., 2011). Inverse document frequency (idf) is taken as the log of the inverse of the document frequency. Each document is represented as a vector of features or terms weights i.e. tf*idf which indicates the degree of association of the respective features with the document. The profile of each

Figure 1. Block Architecture
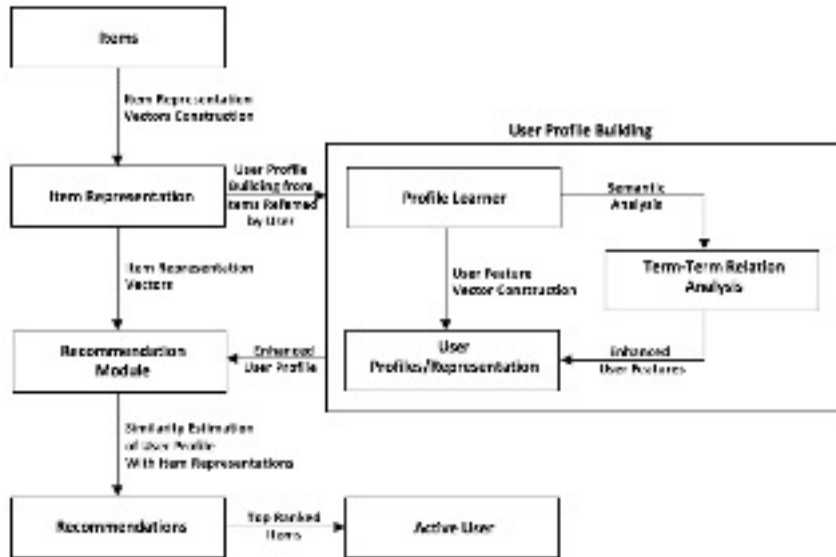
user is represented by a vector of weights of features. In this work the average or mean weight of each feature is estimated from the item representation vectors of items referred by user in past (Lops et al., 2011; Tanwar & Khatri, 2019).

B.   Feature Vector Construction for Candidate Papers

In most content-based recommender systems, items such as Web pages, emails, journals, news articles, etc. are described by textual features extracted from them. There are no attributes with well-defined values contrary to the case of structured data. Like the feature vectors of papers of user interest constructed, the feature vectors of candidate papers are also constructed using the tf*idf scheme (Lops et al., 2011; Tanwar & Khatri, 2019).

C.   User Profile Enhancement

The user feature vectors or user profiles are enhanced by term coupling learning i.e. by extracting the semantic relationship between features. The coupling relationships i.e., Intra-couplings and Inter-couplings between terms or features are estimated for user profile enhancement.

1.   Intra-Relation between terms:

Intra-relation captures the explicit relationship between terms by their co-occurrence frequency in documents. The co-occurrence frequency of terms across all documents is calculated using the popular co-occurrence Jaccard measure (Cheng et al., 2013; Tanwar & Khatri, 2019).

2.   Inter-Relation between terms:

$Begin:$
$Step\,1:$ $Construction\ of\ User\ feature\ vectors$
$for\ each\ set\ of\ documents\,\{d1,\ d2,\ d3,....,\ dn\}\ referred\ by\ user\ `Ux'$
do
$construct\ user\ feature\ vector\ `f_{ufv}^{x}'\ for\ user\ `Ux'\ by\ taking\ mean\ of\ tf{*}idf\ scores\ of\ top\ `y'\ terms$
$Step\,2:$ $Semantic\ Analysis\ -\ Estimating\ Explicit\ Relationship\ between\ features$
$for\ each\ terms\,(ti,\ tj)$
do
$estimate\ the\ Intra\ relation\ between\ terms\ `IaR(ti,\ tj)'\ as$
$$IaR(ti,\ tj) = \begin{cases} 1\ ; & i = j \\ \dfrac{CoR(ti,\ \ tj)}{\sum_{i=1,\ i\neq j}^{n} CoR(ti,\ \ tj)} \ ; & i \neq j \end{cases}$$
$Where,\ \sum_{i=1,\ i\neq j}^{n} IaR(ti,\ \ tj) = 1\ \&\ IaR(ti,\ \ tj) =\ IaR(tj,ti)$
$Step\,3:$ $Semantic\ Analysis\ -\ Estimating\ Implicit\ Relationship\ between\ features$
$for\ each\ terms\,(ti,\ tj)\ via\ common\ terms\ `tk'$
do
$estimate\ the\ Inter\ relation\ between\ terms\ `IeR(ti,\ tj)'\ as$
$$IeR(ti,\ tj) = \begin{cases} 0\ \ ; & i = j \\ \dfrac{1}{|L|} \cdot \sum_{\forall\ tk\in L} R\_IeR(ti,\ tj\ |\ tk); & i \neq j \end{cases}$$
$Where\ `L'\ represents\ the\ number\ of\ link\ terms\ \&$
$L = \{\ tk\ |\ (\ IaR(tk,\ ti)\ >\ 0) \blacklozenge (\ IaR(tk,\ tj)\ >\ 0)\}$
$Step\,4:$ $User\ Feature\ Enhancement$

$$In\ user\ feature\ vector,\ \ f_{ufv}^{x}\ =\ \{ _{W_{t1}{}^{ufv},\ W_{t2}{}^{ufv},\ \cdots,\ W_{tn}{}^{ufv}}\}$$

$$forIeR\left(ti,\ tj\right)_{> T}\ =\ \begin{bmatrix} W_{ti}^{ufv}\ =\ W_{tj}^{ufv}\ ;\quad W_{tj}^{ufv}\ >\ W_{ti}^{ufv} \\ W_{tj}^{ufv}\ =\ W_{ti}^{ufv}\ ;\quad W_{ti}^{ufv}\ >\ W_{tj}^{ufv} \end{bmatrix}$$

$Where,\ T\ is\ inter\ relation\ threshold$

End

Intra-relations are the relations between terms which co-occur in at least one document. Some terms co-relate semantically with each though they do not co-exist in the same documents. For instance, synonyms of terms, they might exist exclusively in separate documents. Similarly, terms having close semantic relation in the real world might occur in different sets of documents. These relationships or couplings between terms can't be inferred by capturing the intra-relations alone but such term's relation with other terms known as the link terms gives a relationship measure between the terms. Such relations are known as inter-relations (Cheng et al., 2013; Tanwar & Khatri, 2019).

D. Proposed User Feature Vector Enhancement

To enhance the user feature vector the implicit relations or inter-relations between the terms have been taken into consideration and on basis of that weights of relevant terms or features have been enhanced. Once the coupling relationships between terms are inferred, the semantic information is incorporated in the user feature vector. The terms having respectively lower weights in the user feature vector but having higher inter-relations with terms having high weightage in the user feature vector are enhanced. For instance term 't_i' has a high inter-relation with another term 't_j' i.e. above a threshold value 'T', and 't_i' has a weight respectively lower than 't_j', in that case the weight of 't_i' is upgraded to the weight of 't_j' in the user feature vector and the vice-versa is also true. This is based on the assumption that if the inter-relation between two terms is high or above a threshold value, the terms have a high semantic relationship. Therefore, they should have comparable or close weights/ significance in the feature vectors.

As represented in the algorithm, 'T' is the threshold value above which the inter-relation values are considered relevant. The value of 'T' would vary according to the requirement. It would depend on a number of factors such as the similarity requirement quotient of the documents, the tf*idf values range of the terms of the user feature vector and other conditions. For higher values of 'T', lesser number of term's inter-relations would be considered in the user feature enhancement. For lower values of 'T', more term's inter-relations would be considered in the user feature enhancement. Hence, there has to be a trade-off in between enhancing the underestimated relevant terms and not enhancing all the values otherwise the purpose would be lost. Future work can be done to formulate and estimate the value of 'T'. After experimentation it has been observed that the proposed enhanced user feature vectors show better similarities with the relevant papers/items which help the relevant items to match better and surface up in the recommendations space.

The proposed approach is based on Non-IIDness based term-term relation analysis (Cheng et al., 2013). Non-IIDness learning refers to understanding, modelling, analysing and representing non-IID data (not independent and identically distributed data). Coupling and heterogeneity are the important aspects of Non-IIDness learning. In the proposed work based on Non-IIDness learning, the co-occurrence relationships/couplings between the features or words recurring in the documents

of user interest are extracted by estimating the intra-relations. Also, the relationships or couplings between such keywords which do not co-occur in the same documents but are inter-related semantically in the real world are extracted by estimating the inter-relations. Based on semantic couplings, the under-rated relevant features in the user profile are enhanced so that the significance of such terms gets a lift. Similarly, the synonyms or terms having the same meaning also get their due weightage. Hence, the documents having these terms which did not match earlier because of under-estimated values match better after feature enhancement. The semantically related documents though having different sets of relevant terms match better after feature enhancement; leading to surfacing up these relevant documents for improved recommendations in the system. The proposed approach leads to better recommendations after quantifying and incorporating the Non-IIDness based semantic closeness/couplings between the features of the user profile.

EXPERIMENTATION RESULTS

A.   Recommending Papers

After having the users enhanced feature vectors and the candidate papers feature vectors, the similarities of possible candidate documents are measured with the user profiles. The top matching papers in terms of high similarity above a threshold are recommended to the respective users. The similarities between the vectors are calculated using the cosine similarity measure (Lops et al., 2011; Sugiyama & Kan, 2010). A threshold-based approach or a top-n approach for recommending items could be implemented alternatively. The candidate papers having similarity with the respective user profiles above a threshold are considered to be recommended to the users in this work.

B.   Experimental Data

For experimentation, the dataset for scholarly paper recommendations has been used (Cao, 2014). This dataset consists of several hundreds of references, citations and publication papers of researchers working on Natural Language Processing (NLP) and Information Retrieval (IR) and have publication lists in DBLP (Sugiyama & Kan, 2010). For this work, the reference papers have been divided into two sections. Nearly 70 percent of papers are used for user profile construction and remaining papers are used for estimating the recommendation accuracy. The feature vectors of the papers having the normalized term frequencies of each term are given in the dataset. Data pre-processing, user's and item's profile construction, term coupling learning, user feature enhancement, similarity calculations and analysis of results have been done using Python Programming language and MySQL database.

C.   Data Pre-processing

Once the feature space of each item or paper is ingested in the database, data pre-processing is done to make the data appropriate for further processing as described below.

I.   Removal of stop words:

The stop words such as 'the', 'of', 'on', 'is', etc. have been removed from the items feature vectors as they are of no importance and only reduce the efficiency of the system by giving inappropriate results. The stop words have been removed by excluding these terms from the item's feature space using the stop words list maintained in the MySQL database.

II.   Removal of single terms

The single letters such as 'a', 'i', 'n', etc. have been removed as such letters have no contribution in the results. The single letters have been eliminated from the items feature space. Since the item profiles do not have these terms after elimination, therefore the user profiles also do not have them.

D.   Term Coupling learning and Similarity Estimation

I.   Feature Vector Construction

The inverse document frequencies have been inferred as described in section III and multiplied by the term frequencies to construct the user and candidate paper feature vectors. The user feature vectors have been built by taking the mean of each feature or term weights given in all document vectors of user interest.

II.   Feature Vector Reduction

The number of features has been reduced. The most relevant terms on the basis of high weights have been filtered out and rest of the terms have been ignored. Experiments have been conducted on top 20 terms. The number of terms has been reduced to improve the response time and effectiveness.

III.   Term Coupling Learning

Once the user feature vectors have been constructed and reduced, the term coupling relations i.e., intra-relations and inter-relations are estimated between the terms of the user profile. The intra-relations (explicit relations) are estimated using the Jaccard measure across all the papers of user interest. On the basis of intra-relations, the inter-relations (implicit relations) between terms which are not directly intra-related but intra-related to at least one common term are estimated. The highly coupled inter-related terms are considered for user feature enhancement.

IV.   User Feature Enhancement

After the coupled term-term relation analysis, the highly inter-related term pairs and their respective weights in user feature vectors are extracted. For each coupled term-term pair in the user profile, the term having comparatively lower weight is enhanced by assigning the weight of the other coupled term having higher weight.

V.   Document Similarity Estimation

The enhanced user feature vectors and the candidate paper vectors are normalized before the similarities are measured using the cosine similarity measure. The cosine similarity between the actual state-of-the-art tf-idf user feature vectors (Lops et al., 2011) and the candidate paper vectors are also measured for comparison between the similarity of the candidate papers with the actual user profiles and the enhanced user profiles at different threshold values. On the basis of similarity estimations, the recommendation accuracy has been analysed as discussed in the following sections.

E.   Evaluation Criteria

The recommendation accuracy has been estimated by three criteria: Recall, Precision and F-Score or F-measure. Recall measures the number of relevant documents retrieved with respect to the total number of relevant documents and precision measures the fraction of relevant documents retrieved from the total retrieved documents at a threshold value. F-score or F-measure or F_1 score generalizes the analysis on the basis of Precision and Recall to one figure. The recall and precision values at different threshold values of document similarities are measured for both the actual and enhanced user feature vectors. The recall and precision values for the candidate papers are tested at two different values of 'T' also, i.e., the inter-relation threshold. F-scores of the same have also been estimated. For this experiment, precision is the preferred or most relevant metric as precision indicates what fraction of retrieved documents is relevant. Higher Precision would indicate a higher chance of the recommendation being appropriate. Appropriate recommendations can keep a user more engaged and interested.

F. Experimental Results

It has been observed that the recall, precision, and F-score values improve mostly for the enhanced user feature vectors as compared to the actual user feature vectors. This means that the documents that are relevant to the user's interest show enhanced similarities which would lead to better recommendations. The experimentation results are presented in Table 1 to Table 3.

Table 1. Recall Values

Content-Similarity	>.1	>.13	>.16	>.19	>.23	>.26	>.30	>.33	>.36
Recall-(tf-idf matching)	87.14%	81.43%	68.57%	60%	52.86%	47.14%	35.71%	34.29%	34.29%
Recall-Enhanced(>.20)	87.14%	82.86%	72.86%	70%	62.86%	47.14%	38.57%	35.71%	35.71%
Recall-Enhanced(>.30)	88.57%	88.57%	74.29%	70%	64.29%	51.42%	38.57%	35.71%	35.71%

Table 2. Precision Values

Content-Similarity	>.1	>.13	>.16	>.19	>.23	>.26	>.30	>.33	>.36
Precision-(tf-idf matching)	33.7%	42.54%	47.52%	50%	54.41%	64.71%	65.79%	66.67%	66.67%
Precision-Enhanced(>.20)	32.62%	43.94%	46.36%	55%	64.71%	66.67%	67.5%	73.53%	73.53%
Precision-Enhanced(>.30)	34.44%	44.28%	48.15%	59%	64.29%	67.92%	67.92%	73.53%	75.86%

Table 3. F-Score Values

Content-Similarity	>.1	>.13	>.16	>.19	>.23	>.26	>.30	>.33	>.36
F-Score-(tf-idf matching)	0.49	0.56	0.56	0.55	0.54	0.55	0.46	0.45	0.45
F-Score(>.20)	0.47	0.57	0.57	0.62	0.64	0.55	0.49	0.48	0.48
F-Score(>.30)	0.5	0.59	0.58	0.64	0.64	0.59	0.49	0.48	0.49

G.   Result Analysis

It is clear that by user feature enhancement, the precision and recall is improving mostly for relevant papers. For this experiment, Precision is the most preferred metric as it indicates the chances of appropriate recommendations. It has been observed that the precision and recall enhanced for the candidate documents in the range of (> 0.1 to > 0.36) in this case (where the document similarities lied in the range of 0.002 to 0.42 with most of the values lying in between 0.09 and 0.39). Below this range and above this range little difference in the results has been observed. So, it has been analysed that for very low values and very high values of similarity threshold, the enhanced features do not bring much of a difference to the actual similarity values. The values are most enhanced at similarity thresholds '> 0.19' and '> 0.23' clearly visible from the Tables 1 to Table 3 and from Figure 2 to Figure 4.

The recall values have either enhanced or are the same at few places which means that at few places the features lift is not enough to cross that particular threshold. The precision values also either increase or remain the same except at two places: at similarity '> 0.1' and at '> 0.16' in case of $T > 0.20$ (as given in Table 2) though the recall value is the same for the first case and improves for the second case (as given in Table 1). The reason is that the number of retrieved documents increased (also increasing the number of relevant documents retrieved) in this case since the similarity of some documents got enhanced which were not referenced by the respective users but had the enhanced feature terms. These all papers are of the same domain so the similarity of many papers got influenced but it did not affect much since the enhanced values could not reach the threshold. The results would be more evident where there are documents of multi-disciplinary fields.

H.   Performance Evaluation

It is quite evident from the results that the enhanced user feature vectors are enhancing the similarities of the relevant documents since precision and recall both are improving. The similarity threshold values lie in the range from 'above 0.1' to 'above 0.36'.

I.   Impact of Inter-Relations

For higher inter-relations between terms, the weightage of the terms having inferior weights enhance with respect to the higher weighted coupled terms. Since the terms are more relevant as they are among the top 'k' terms of the user feature vector and have high semantic relation with relevant terms but have underestimated weights, enhancing their weights make sense. By inferring inter-relations and enhancing the user feature vector, the similarity of documents having these relevant terms gets enhanced. The terms which are related semantically in real world but do not come into consideration during document similarity estimation get a lift. Hence, documents having these terms match accurately with the user feature vectors resulting in the recommendation of documents which are according to the user's interest but did not match appropriately before enhancement.

II.   Impact of 'T'

The results are more enhanced for appropriate documents or papers in case of value of $T > 0.30$ though more number of inter-relations are retrieved for feature enhancement in the case of $T > 0.20$ and hence more feature values are enhanced in case of $T > 0.20$. The reason is that when the vectors are normalized before similarity measure, in case of $T > 0.20$, we have a higher denominator quotient which reduces the weight of feature terms which are not enhanced or are of less weight comparatively, though the values of precision and recall improve at both values of 'T'.

At values of cosine-similarity thresholds '> 0.1', '> 0.13', '> 0.16', '> 0.23' and '> 0.26' as given in Table 1, the recall values are higher at T > 0.30 as compared to T > 0.20. Similarly, the precision values are higher at cosine-similarity '> 0.1', '> 0.13', '> 0.16', '> 0.19', '> 0.26', '> 0.30' and '> 0.36' for T > 0.30 as compared to T > 0.20.

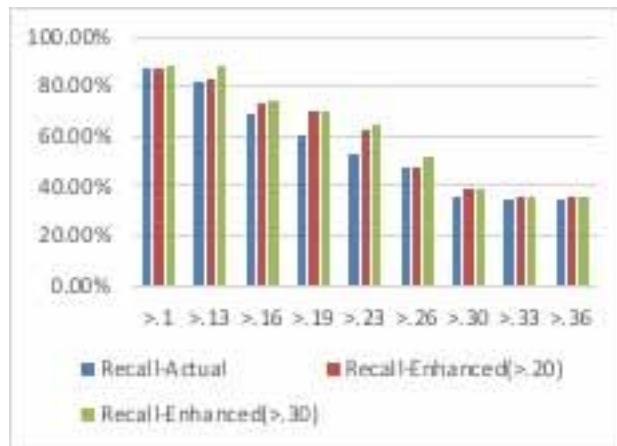**Figure 2. (a) Bar-Chart for Recall values Figure 2. (b) Line-Chart for Recall values**

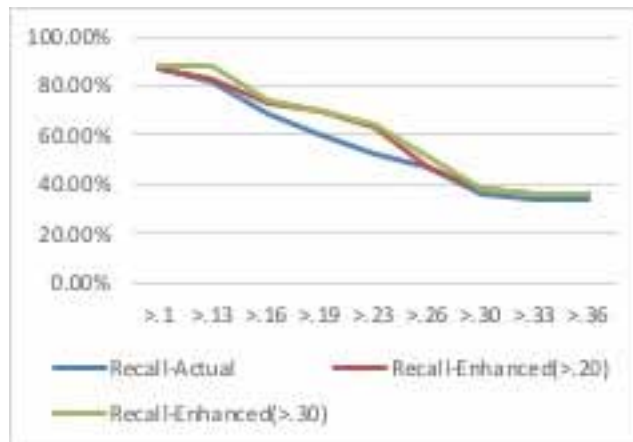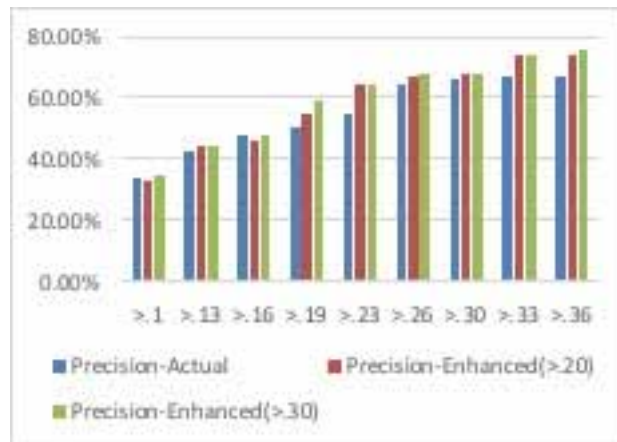**Figure 3. (a) Bar-Chart for Precision values Figure 3. (b) Line-Chart for Precision values**

Figure 4. (a) Bar-Chart for F-score values Figure 4. (b) Line-Chart for F-score values

From Table 1, it is quite evident that recall for T > 0.30 is improving at many places as compared to recall values for T > 0.20 though more user features are being enhanced in case T > 0.20. The recall for T > 0.30 has improved at all the places with respect to the actual values. Though, the recall value for T > 0.20 at Cosine-Similarity > 0.26 is the same as the recall value for the actual user profile i.e., 47.14%. In future, work can be done to mathematically formulate a suitable value of 'T' specific to the feature vectors available for similarity calculation.

III. Impact of feature size

The feature size can play a crucial role in the results since more the number of features, better is the chance of enhancement. The space for inter-relation calculation increases. Hence, the feature size should not be small enough to not capture the inter-relations and lose the relevant user specific information nor should it be large enough to accommodate many inter-relations of each word which would superficially increase the term weights. For 35 features in user profile, there are 35 features for inter-relations estimation. Because of such large space of features, some features formed inter-relations with multiple terms leading to superficial exaggerated enhancement of the features. For instance, the term 'term' got one of the most enhanced features for one of the users. For 10 feature's set, very few inter-relations have been captured making no significant enhancements.

In this work, we considered a feature vector of size 15 and 20 features. One more condition was imposed on the user feature sets, the features which existed in at least 15 to 20 percent of the documents used to build the user feature vector are only considered as the features in the user feature vector construction.

IV. Impact on Recommendations

This will lead to better recommendations as the terms of significance have been enhanced in the user profile. Therefore, the similarities of documents which have these terms are likely to improve and meet the threshold values for making recommendations. Also, the appropriate documents or items which are of more relevance semantically to the user profile will surface up for recommendations. This is evident from the Tables 1, 2 and 3. The Recall and Precision is improving which indicates that more relevant documents are meeting the threshold values. These are the documents that the

researchers or users have referred to. By user feature enhancement it is clear that more number of documents that they referred has met the criteria as compared to before.

V.   Impact on synonyms

This will also help in enhancing few terms which are used as synonyms of certain relevant terms but since they are used in few documents they have lower weights in the feature vectors. For instance, 'sentiment-analysis' and 'opinion-mining' are the same but sentiment-analysis is widely used in comparison to opinion-mining. By this approach if the weight of opinion-mining is underestimated in the user feature vector, it would be enhanced since it will form higher inter-relation with sentiment-analysis.

CONCLUSION

Big data analytics plays an important role in day-to-day life to extract useful knowledge from structured and unstructured data but work on unstructured data has challenges associated (Tanwar et al., 2015; Duggal et al., 2015). Various information sites and digital libraries are overloaded with text-based content. Filtering personalized information according to user requirement and inclination from the large available information space has limitations. The dynamic and heterogeneous nature of the sources adds up to the complexity. As a result, the need for effective user profile learning, modelling and semantic enrichment becomes crucial. In this paper, the user interest feature vectors have been enhanced semantically so that the relevant documents meet the similarity criteria for recommendations. The relevant terms are enhanced in the user feature vectors by coupling the explicitly and implicitly related features. The semantic information is incorporated in the user profiles so that the interest areas of users are accurately represented. During experiments, better recommendation predictions have been observed. Improved precision recall and f-scores have been observed for semantically enhanced user profiles. In future, we intend to continue working on extracting semantic information from content referred by users and incorporating the semantics in user interest profiles for better recommendations. We also intend to explore if multidimensional user interest areas can be identified by semantic analysis of the content referred by users.

ACKNOWLEDGMENT

REFERENCES

Araki, M., Katsurai, M., Ohmukai, I., & Takeda, H. (2017). Interdisciplinary Collaborator Recommendation Based on Research Content Similarity. *IEICE Transactions on Information and Systems*, E100-D(4), 785–792. doi:10.1587/transinf.2016DAP0030

Balabanovic, M., & Shoham, Y. (1997). Fab: Content-Based, Collaborative Recommendation. *Communications of the ACM*, *40*(3), 66–72. doi:10.1145/245108.245124

Bollacker, K. D., Lawrence, S., & Giles, C. L. (1998). CiteSeer: An AutonomousWeb Agent for Automatic Retrieval and Identification of Interesting Publications. In *Proceedings of the Second International Conference on Autonomous Agents* (pp. 116–123). ACM Press. doi:10.1145/280765.280786

Cao, L. (2014). *Non-IIDness Learning in Behavioral and Social Data. The Computer Journal, 57(9), 1358−70*. doi:10.1093/comjnl/bxt084

Cao, L. (2016). Non-IID Recommender Systems: A Review and Framework of Recommendation Paradigm Shifting. *Engineering*, *2*(2), 212–224. doi:10.1016/J.ENG.2016.02.013

Cheng, X., Miao, D., Wang, C., & Cao, L. (2013). Coupled term-term relation analysis for document clustering. In *Proceedings of the 2013 International Joint Conference on Neural Networks (IJCNN)*. IEEE. doi:10.1109/IJCNN.2013.6706853

Duggal, R., Shukla, B., & Khatri, S. K. (2015). Big Data Analytics in Indian Healthcare System – Opportunities and Challenges. *National Conference on Computing, Communication and Information Processing (NCCCIP-2015)*. https://doi:NCCIP2015/NERIST/02/03-05-2015/CP28

Fu, B., Xu, G., Cao, L., Wang, Z., & Wu, Z. (2015). Coupling Multiple Views of Relations for Recommendation. In Advances in Knowledge Discovery and Data Mining. PAKDD 2015. Lecture Notes in Computer Science, 9078. Springer. doi:10.1007/978-3-319-18032-8_57

Koren, Y., Bell, R., & Volinsky, C. (2009). Matrix Factorization Techniques for Recommender Systems. *Computer, 42*(8), 30-37. 10.1109/MC.2009.263

Li, F., Xu, G., & Cao, L. (2015). Coupled Matrix Factorization Within Non-IID Context. In Advances in Knowledge Discovery and Data Mining. PAKDD 2015. Lecture Notes in Computer Science, 9078. Springer. doi:10.1007/978-3-319-18032-8_55

Linden, G., Smith, B., & York, J. (2003). Amazon.com recommendations: Item-to-item collaborative filtering. *IEEE Internet Computing*, *7*(1), 76–80. doi:10.1109/MIC.2003.1167344

Lops, P., de Gemmis, M., & Semeraro, G. (2011). Content-based Recommender Systems: State of the Art and Trends. In F. Ricci, L. Rokach, B. Shapira, & P. Kantor (Eds.), *Recommender Systems Handbook* (pp. 73–105). Springer. doi:10.1007/978-0-387-85820-3_3

Middleton, S. E., Shadbolt, N. R., & De Roure, D. C. (2004). Ontological user proðling in recommender systems. *ACM Transactions on Information Systems*, *22*(1), 54–88. doi:10.1145/963770.963773

Sugiyama, K., & Kan, M. Y. (2010). Scholarly paper recommendation via user's recent research interests. In *Proceedings of the 10th annual joint conference on Digital libraries (JCDL'10)* (pp. 21-25). ACM. doi:10.1145/1816123.1816129

Tanwar, M., Duggal, R., & Khatri, S. K. (2015). Unravelling Unstructured data: A wealth of information in big data. In *Proceedings of 2015 4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions)*. IEEE. doi:10.1109/ICRITO.2015.7359270

Tanwar, M., & Khatri, S. K. (2018). A Framework for User Profile Enrichment in Content-based Recommender Systems by inferring the Semantic Couplings. In *Proceedings of 2018 Fifth International Symposium on Innovation in Information and Communication Technology (ISIICT)*. IEEE. doi:10.1109/ISIICT.2018.8613290

Yu, Y., Wang, C., Gao, Y., Cao, L., & Chen, X. (2013). A Coupled Clustering Approach for Items Recommendation. In JAdvances in Knowledge Discovery and Data Mining. PAKDD 2013. Lecture Notes in Computer Science, 7819. Springer. doi:10.1007/978-3-642-37456-2_31

Mona Tanwar is presently pursuing PhD from Amity University, Noida. She has experience in the field of Academics, Data Science and Machine Learning and has worked for organisations such as Google NYC, US (On Contract basis) and Adglobal360 (Gurugram, India). Presently, she is working as Data Scientist for Hubspot, US. Her research area is Machine Learning, User Profile Learning and Modeling, User Engagement and Data Science for leveraging Emails for improved business KPI's and metrics.

Sunil Kumar Khatri is Director of Campus at Amity University in Tashkent, Uzbekistan. He is a Fellow of IETE, Sr. Life Member of IEEE, CSI and IASCSIT and Member of IAENG. He is Secretary in SREQOM, Past Convener, EAC, IEEE UP Section Executive Council and Past Vice-Chairman of CSI Noida Chapter. Dr. Sunil Kumar Khatri is Editor IJSAEM, Springer Verlag. He is in Editorial Board of several Journals from USA, Egypt, Hong Kong, Singapore and India. He has twelve edited books, eleven guest edited special issues of international journals, eleven patents filed and more than 225 papers in international and national journals and proceedings. His areas of research are Artificial Intelligence, Software Reliability and Testing, and Data Analytics.

Ravi Pendse serves as the Vice President for Information Technology and Chief Information Officer at the University of Michigan (USA). He is an Executive Officer and provides university-wide leadership and strategic direction for information technology. He is also a professor in the Department of Electrical Engineering and Computer Science. Dr. Pendse has secured more than $21 million in external research grants, developed and taught courses in the areas of computer architecture, networking, and cybersecurity, earned several teaching awards, and published numerous scholarly articles co-authored with students. His research interests include areas of Internet of things, cybersecurity, and future of work. Dr. Pendse holds a B.S. in electronics and communication engineering from Osmania University in Hyderabad (India). He received his M.S. and Ph.D. in electrical engineering from Wichita State University. He serves as a board member on the Global Customer Advisory Board of Cisco Systems, board member of Unizin and as an independent director on the board of High Touch Technologies.