ComEx Miner: Expert Mining in Virtual Communities

Akshi Kumar
Dept. of Computer Engineering
Delhi Technological University
Delhi, India

Nazia Ahmad
Dept. of Computer Engineering
Delhi Technological University
Delhi, India

***Abstract*— The utilization of Web 2.0 as a platform to comprehend the arduous task of expert identification is an upcoming trend. An open problem is to assess the level of expertise objectively in the web 2.0 communities formed. We propose the "ComEx Miner System" that realizes Expert Mining in Virtual Communities, as a solution for this by quantifying the degree of agreement between the sentiment of blog and respective comments received and finally ranking the blogs with the intention to mine the expert, the one with the highest rank score. In the proposed paradigm, it is the conformity & proximity of sentimental orientation of community member's blog & comments received on it, which is used to rank the blogs and mine the expert on the basis of the blog ranks evaluated. The effectiveness of the paradigm is demonstrated giving a partial view of the phenomenon. The initial results show that it is a motivating technique.**

Keywords- expert; web 2.0; virtual community; sentiment analysis.

I. INTRODUCTION

Expert identification is an intricate task because experts and their expertise are rare, expensive, unevenly disseminated, hard to qualify, continuously varying, unstable in level, and often culturally isolated and oversubscribed. The expert seekers behavior further complicates this, as they typically have improperly articulated requirements, are ignorant of expert's performance history, and are not well equipped to differentiate between a good and a bad expert.

Web 2.0 [1] is an evolution from passive viewing of information to interactive creation of user generated data by the collaboration of users on the Web. The proliferation of Web-enabled devices, including desktops, laptops, tablets, and mobile phones, enables people to communicate, participate and collaborate with each other in various Web communities, viz., forums, social networks, blogs. Thus, evidently the Internet now forms the basis for the constitution of virtual communities. According to the definition of Howard Rheingold in [2], virtual communities are social aggregations that emerge from the Net when enough people carry on public discussions long enough, with sufficient human feeling, to form webs of personal relationships in cyberspace.

Thus, the expected alliance of these active areas of research, namely, Expert Identification & Web 2.0, fills the gaps that exist in the diversified Web. In response to the identified need to better exploit the knowledge capital accumulated on the Web 2.0 as a place for people to seek and share expertise, the operative challenge is to mine experts in virtual communities. Expert identification in virtual communities is noteworthy for the following reasons [3]. Firstly, virtual communities are knowledge pools where members communicate, participate and collaborate to gather knowledge. Intuitively, we tend to have more confidence on an expert's text. Secondly, virtual communities allow interaction of novices with experts, which otherwise in real world is tedious and expensive.

Instigated by the challenge to find experts in the virtual communities, we propose a **Com**munity **Ex**pert Mining system called the ComEx Miner system, where, firstly we build an interest similarity group, an online community which is a virtual space where people who are interested in a specific topic gather and discuss in depth a variety of sub-topics related to the topic using blogs. We further propose to mine the sentiment of the each group member's blog along with the sentiment of their respective comments. This is based on the intuition that the blogger and the commenter talk about the same topic or product, treated as feature for opinion orientation identification and if the blog's sentiment about a topic/product matches with the commenter's sentiment about the topic/product this implies that blogger's knowledge about the topic/ product is acceptable as people agree to what has been talked about in the blog. This degree of acceptance matching would then help to rank the blog and mine the expert with highest blog rank.

The main components of the ComEx Miner are:

Interest Mining Module: This module puts forward an algorithm for Interest Group construction by uncovering shared interest relationships between people, based on their blog document entries. The key point of constructing this Collaborative Interest Group is the calculations of interest similarity relations and application of the K-means clustering technique to cluster researchers with similar interests into the same group.

- ***Expert Mining Module:*** The ranking of member's blog within the built group is done on the basis of the score obtained by conjoining the blog & average comment orientation. This helps to identify the expert, the one with the highest ranking blog. The module is further divided into the following sub-modules:

- ***Sentiment Mining Module:*** The goal of this module is to perform sentiment analysis of the group member's blogs and the comments received on the respective blog. It gives the strength of the blogs and strength of their respective comments.
- ***Blog Ranking:*** Once the blog strength and comment strength has been determined; this module ranks the blogs by calculating the blog score, a metric which combines the respective blog's strength to the average comment strength. The blogs are then ranked as per the blog score and the expert is identified as the one with the highest rank.

The paper is organized into 4 sections. Section 2 highlights the related and background work pertinent to the research carried. Section 3 illustrates the proposed ComEx Miner System expounding the methodology used to mine the expert from a virtual community, followed by section 4 which demonstrates the results and analysis of proposed paradigm with the help of sample data. Finally, the conclusion lists out the key contributions of the research work presented.

II. Related Work

We seek *guidance* from people who are familiar with the choices we face, who have been helpful in the past, whose perspectives we value, or who are recognized experts [4]. Expert finding addresses the task of identifying the right person with the appropriate skills and knowledge [5]. There are various approaches related to expert identification & expertise search available in literature. Bogers et al. [6] used two methods: content based expert finding using academic papers and expert finding using social citation network between the documents and authors for finding experts. Breslin et al. [7] introduced a concept of re-using and linking of existing vocabularies in the semantic web, which can be used to link people based on their common interest. They described that a framework made by the combination of popular ontologies FOAF, SIOC, SKOS could allow one to locate an expert in a particular field of interest. Metze et al. [8] proposed a system to provide exchange of information by determining experts who can answer a given question. They provided a prototype expert finding system which enables individual within a large organization to search for an expert in certain area. Schall and Dustdar [9] addressed the problem of expertise mining based on performed interactions between people. Their approach comprised of two steps: Firstly, of offline analysis of human interaction considering tagged interaction links. Secondly, composition of ranking scores based on performance. Huh et al. [10] presented a grid enabled framework of expertise search (GREFES) engine, which uses online communities as sources for expert on various topics. They also suggested an open data structure SNML (Social Network Markup Language) for sharing community data. Smirnova and Balog [11] have argued that in real world, the notion of best expert depends on the individual performing the search. They proposed a user oriented model that incorporates user-dependent factor. It is based on the assumption that the user's preferences for an expert is balanced between the time needed to contact the expert and the knowledge value gained after .Li et al.[12] describe a method for finding expert through rules and taxonomies. They have proposed a combination of RDF FOAF facts and RuleML FOAF rules.

Punnarut and Sriharee [13] have introduced a method for finding expertise research using data mining and skill classification ontology. Zhang et al. [14] utilize an online community to find the people who may have expertise for answering a particular question. They analyze the experts by considering interactions of the people in questioning and answering the questions. Tang et al. [15] propose an expertise search system that analyses information from a web community. They use ontology to determine the correlation between information collected from different sources.

In the research presented in the paper, we intend to mine the experts in an online community which is a virtual space where people who are interested in a specific topic gather and discuss in depth a variety of sub-topics related to the topic using blogs. The conformity & proximity of sentimentally orientation of community member's blog & comments received is then used to rank the blogs and mine the expert on the basis of the blog ranks evaluated. The next section furnishes the details of the proposed paradigm.

III. The Proposed ComEx Miner System

In general, an expert is someone who possesses a high level of knowledge in a particular area. This entails that experts are reliable sources of relevant resources and information. An open problem thus arises to assess the level of expertise objectively. We propose the "ComEx Miner System" that realizes Expert mining in virtual communities, as a solution for this by quantifying the degree of agreement between the sentiment of blog and respective comments and finally ranking the blogs with the intention to mine the expert, the one with the highest rank score.

Figure 1 shows the architectural overview of ComEx Miner System proposed in this research.

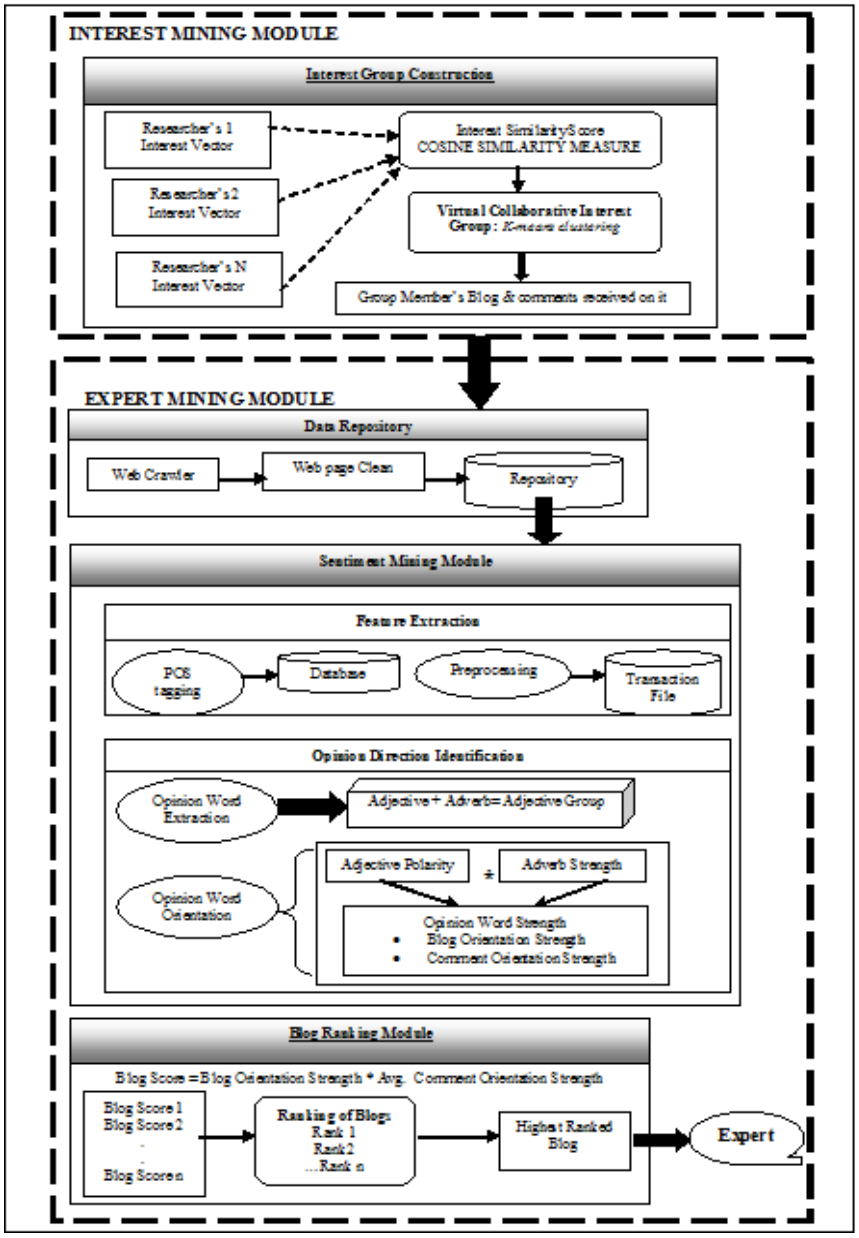


Figure 1. System Architecture of the ComEx Miner System

The following sub-sections expound the details of the ComEx Miner:

A. *Interest Mining Module*

In this module, we focus on the problem of discovering people who have particular interests. The Interest Group construction algorithm is based on interest similarity, which can cluster researchers with similar interests into the same group and facilitate collaborative work.

The following sub-sections expound the details of the Collaborative Interest Group construction [4]:

1) Interest Vector: Each researcher writes blog entries according to his or her interest. The interest vector of the researcher, V_i, is represented as a bag-of-words with frequently used words being assigned high weights. The interest vector is calculated by the equation described below:

and;

$$V_i = (s_{i1}, s_{i2}, s_{i3}, \ldots) \quad (1)$$

And

$$s_{ik} = ef_i(w_k) \times \log\left(\frac{N_u}{uf(w_k)}\right) \quad (2)$$

where s_{ik} means the strength of interest in word w_k; $ef_i(w_k)$ means the number of entries containing w_k in researchers i's site; $uf(w_k)$ means the number of researchers who use w_k; and N_u means the number of researchers.

INTEREST MINING MODULE

Input: Researchers' Blog which contains their research papers

Output: Construction of Collaborative Interest Group

Steps:

1. **Interest Vector**
   - We calculate the interest vector V_i for each researcher i as follows :-

     for each researcher i
       for each frequently-used word w_k in his blog
     {find the values of entry-frequency $ef(w_k)$ & user-frequency $uf(w_k)$ calculate the strength of interest in word w_k (product of ef & log of inverse uf)}
       endfor
     endfor

2. **Interest Similarity Score**
   - We calculate the interest similarity score R_{ij} between researchers i and j using the cosine similarity of V_i and V_j

3. **Collaborative Interest Group Construction**
   We construct the collaborative interest group by using the technique of K-means clustering algorithm. It consists of two basic steps as follows:
   - We find the total number of clusters, denoted by K with the help of researcher groups so formed.
   - And then we assign points to the closest centroid by taking the proximity measure as the distance between two researchers.

2) Interest Similarity Score: A similarity score represents how similar the interests of a pair of researchers are. If researcher i and j have similar interests, their interest vectors should be similar. Thus, we calculate the similarity score between them, R_{ij}, using the cosine similarity of V_i and V_j as described below.

$$R_{ij} = \frac{V_i \times V_j}{|V_i||V_j|} \quad (3)$$

All elements of V_i and V_j are positive and thus the range of R_{ij} is 0 to 1.

3) Collaborative Interest Group Construction: Construction of an interest group is done to cluster the researchers with similar interests into the same group and facilitate collaborative work. Collaborative Interest Group Construction is done by using the technique of K-means clustering algorithm [16] where K is a user-specified parameter and it refers to the total number of clusters required.

Each point is then assigned to the closest centroid, and each collection of points assigned to a centroid is a cluster. The centroid of each cluster is then updated based on the points assigned to the cluster. We keep repeating this procedure again and again and update steps until no point changes clusters, or equivalently, until the centroids remain the same.

Basic K-means algorithm
1: Select K points as initial centroids.
2: **repeat**
3: Form K clusters by assigning each point to its closest centroid.
4: Re-compute the centroid of each cluster.
5: **until** centroids do not change.

a) Finding total number of clusters, denoted by K:

The value of K is found out by first forming the researcher groups. Total number of researcher groups formed is equal to the total number of researchers and researchers belonging to a particular group can carry out the co-operative work among themselves. Each group will have its respective threshold value which will decide the membership of a particular researcher in that group. T_i denotes the threshold for group i and is found out by averaging all the similarity scores corresponding to researcher i.

Membership criteria:

$$\text{If } R_{ij} >= Ti, \text{ then researcher j belongs to group i}$$
$$\text{else, researcher j belongs to some other group} \quad (4)$$

Now, once all the researcher groups have been formed, then the value of K is equivalent to the minimum number of groups required to cover all the data points.

b) Assigning Points to the Closest Centroid: To assign a point to the closest centroid, we need a proximity measure that quantifies the notion of 'closest' for the specific data under consideration. We use the proximity measure as the distance between any two researchers, denoted by d_{ij} and is given as:

$$d_{ij} = 1 - R_{ij} \quad (5)$$

where d_{ij} denotes the distance between researchers i and j R_{ij} denotes the similarity score between researchers i and j.

c) Centroids and Objective Functions: The next step is to re-compute the centroid of each cluster, since the centroid can vary, depending on the proximity measure for the data and the goal of clustering.

Once the virtual collaborative interest similarity group is put together, the next step is to identify the expert from this group. To realize this task, the sentiment of each group member's blog along with the sentiment of their respective comments is analyzed for opinion strengths. As mentioned previously the degree of acceptance matching would then help to rank the blog and mine the expert with highest blog rank.

B. Expert Mining Module

The expert mining module is divided into three sub-modules; namely, the Data Repository module which collects the web pages from the member's blog & comments, cleans them and then stores them in the repository, the Sentiment Mining Engine that receives these cleaned web pages from the repository and then provides orientation strengths of blogs & respective comments by extracting opinion features and opinion words and the Blog ranking module which finally

EXPERT MINING MODULE

Input: Member's blog and comments on each blog

Output: Expert, one with the highest ranking blog

Steps:

1. **Data Repository**
   - Web Crawler: Crawls the member's blog and respective comments, collects them as web pages.
   - Web page Cleaning: Remove HTML tags
   - Stores in the "Repository"

2. **Sentiment Mining Module**
   - Feature Extraction: POS Tagging ; Preprocessing
   - Opinion Direction Identification:
     - Opinion Words Extraction
     - Opinion Words Orientation
       - ✓ Adjective Polarity
       - ✓ Adverb Strength
       - ✓ Opinion Word Strength

Blog Orientation Strength & Comment Orientation Strength

Blog Ranking Module

- Blog Score = Blog Orientation Strength* Avg. Comment Orientation Strength
- Rank the blogs by their Blog Score .

ranks the blogs on the basis of combined orientation strength of blog & comments to mine the expert as the one with the highest ranking blog.

The details of each of these sub-modules are given in the sections below.

1) Data Repository: This sub-module deals with collecting the web pages and storing them in the repository. Firstly the web crawler periodically crawls the member's blog and respective comments to collect them as web pages. Thereafter, these pages are cleaned up to remove the HTML tags and then are organized properly to be stored in the " Repository".

2) The Sentiment Mining Module: This sub-module deals with providing the actual orientation strenghts of both member's blog & comments received on it. The Sentiment Mining Module receives the web pages from the repository, i.e., if there are k members in a group then k blogs and their respective n comments will be processed to finally calculate the opinion strengths using the following three steps:-

a) Feature Extraction: This is most basic and crucial step for providing orientation strength by identifying those features that the bloggers & commenters have expressed their opinion on. Such features are known as Opinion Features. We make use of both the data mining and NLP Techniques to perform the task of feature extraction. We extract the opinion features with the help of POS Tagging and Preprocessing techniques.

- *POS Tagging (Part of Speech Tagging)*

  POS Tagging is done to find out the features of the product that have been written about. As we know, features are usually noun or noun phrases in the review sentences. Therefore, we use NL Processor linguistic Parser [17] to parse each text, to split texts into sentences and to produce POS Tag for each word (whether the word is a noun, verb, adjective etc.) NL Processor generates XML output and deals only with explicit features, which are the features that occur explicitly as nouns or noun phrases. Each sentence is then saved in a Database along with the POS Tag information of each word in the sentence.

- *Pre-Processing*

  In this sub-step, a transaction file is created which consists of pre-processed noun/noun-phrases of the sentences in the database. Here pre-processing includes the deletion of stop words, stemming and fuzzy matching.

b) Opinion Direction Identification: In this step, we find out the opinion direction using the opinion features extracted in the previous step. To find the opinion direction, we will first extract the opinion words in the text and then find out their orientation strengths. It includes the following sub-steps:

- *Opinion Words Extraction*

  In this sub-step, we extract the opinion words from the text given by the member's in their respective blog & by the commenter's in their comments on that blog. Opinion words are the words that people use to express their opinion (either positive, negative or neutral) on the features extracted in the previous steps. In our work, we are considering the opinion words as the combination of the adjectives along with their adverbs. We have called them collectively as an Adjective-Group (AG). Although, we can compute the sentiment of a certain texts based on the semantic orientation of the adjectives, but including adverbs is imperative. This is primarily because there are some adverbs in linguistics (such as "not") which are very essential to be taken into consideration as they would completely change the meaning of the adjective which may otherwise have conveyed a positive or a negative orientation.

  For example;

  One user says, "***This is a good book***" and;

  Other says, "***This is not a good book***"

Here, if we had not considered the adverb "not", then both the sentences would have given positive review. On the contrary, first sentence gives the positive review and the second sentence gives the negative review. Further, the strength of the sentiment cannot be measured by merely considering adjectives alone as the opinion words. In other words, an adjective cannot alone convey the intensity of the sentiment with respect to the document in question. Therefore, we take into consideration the adverb strength which modify the adjective; in turn modifying the sentiment strength.

Adverb strength helps in assessing whether a document gives a *perfect* positive opinion, *strong* positive opinion, a *slight* positive opinion or a *less* positive opinion.

For example;
One user says, ***"This is a very good book"*** and ;
Other says, ***"This is a good book"***

The Algorithm used for extraction of Opinion Words is given below:

For each sentence in the review database
If (it contains a product feature, *extract all the Adjective-Group* i.e. adjectives and their adverbs as opinion words)
For each feature in the sentence
The nearby adjective and adverb is recorded as its effective opinion (which modifies the noun / noun phrase which is a product feature)

- *Opinion Words Orientation*
  In this sub-step, we find out the orientation strength of the opinion word. As our opinion word consists of adjective + adverb, therefore to find out the orientation of the opinion word, we first find out the polarity of the adjective in the opinion word and then identify the strength of its corresponding adverb in the opinion word which modifies the adjective. Finally, the product of the adjective polarity and the adverb strength gives us the strength (orientation) of the opinion word. The details for finding adjective polarity, calculating adverb strength and deducing the final opinion word strength are as follows:

a) Adjective Polarity

Here, we will identify the semantic orientation for each of the adjective. As we know, words that have a desirable state (e.g. good, great) have a positive orientation, while words that have an undesirable state (e.g. bad, nasty) have a negative orientation. In general, adjectives share the same orientations as their synonym and opposite orientations as their antonyms. Using this idea, we propose a simple and effective method by making use of the adjective synonym set & antonym set in WordNet [18] to predict the semantic orientation of adjectives. Thus, our method is to use a set of seed adjectives whose orientations we know, & then grow this set by searching in the WordNet. The complete procedure for predicting adjective polarity is given below: Procedure "*determine_ polarity"* takes the target adjective whose orientation needs to be determined and the adjective seed list as the inputs.

1. Procedure **determine_polarity** (target_adjective w_i , adjective_ seedlist)
2. begin
3. if (w_i has synonym s in adjective_ seedlist)
4. { w_i's orientation= *s*'s orientation;
5. add w_i with orientation to adjective_ seedlist ; }
6. else if (w_i has antonym a in adjective_ seedlist)
7. { w_i's orientation = opposite orientation of *a*'s orientation;
8. add w_i with orientation to adjective_ seedlist; }
9. end

Note:

1) *For those adjectives that Word Net cannot recognize, they are discarded as they may not be valid words.*
2) *For those that we cannot find orientations, they will also be removed from the opinion words list and the user will be notified for attention.*
3) *If the user feels that the word is an opinion word and knows its sentiment, he/she can update the seed list.*
4) *For the case that the synonyms/antonyms of an adjective have different known semantic orientations, we use the first found orientation as the orientation for the given adjective.*

b) Adverb Strength

We collect all the adverbs which are used to modify the adjectives from English lexicon. Based on the different emotional intensity expressed by the adverb, we mark the negative adverbs with a negative score and other positive adverbs with different score in different sentiment level. The score is ranging from -1 to +1 and a higher score expresses a stronger sentiment. For example, we consider that the adverb "**extremely**" has higher strength than "**more**" does, but lower than that of "**most**". Consequently, "**most**" is marked with 0.9, "**extremely**" with +0.7, and "**more**" with +0.3. Negative adverbs, such as "not", "never", "hardly", "seldom", are marked with a negative score accordingly.

c) Opinion Word Strength

It is calculated by the product of adjective polarity i.e. $P(adj_i)$ and the adverb strength i.e. $S(adv_i)$ and is given by the following formula:

$$S(OW_i) = P(adj_i) \bullet S(adv_i) \quad (6)$$

where, $S(OW_i)$ represents the sentiment of i^{th} opinion word , $P(adj_i)$ represents the polarity of i^{th} adjective and $S(adv_i)$ represents the strength of i^{th} adverb. The value of $P(adj_i)$ is either -1 or +1 and the value of $S(adv_i)$ ranges from -1 to +1.

Therefore, the strength of each opinion word i.e., $S(OW_i)$ will also lie in the range of -1 to +1.

Note:

Sometimes, there is no adverb in the opinion word, so the S(adv) is set as a default value 0.5 . When there is no adjective in the opinion word, then the P(adj) is set as +1.

d) Blog & Comment Orientation Strength: After extracting all the opinion words from the blog and finding their respective strength, the overall strength of a Blog B is calculated by averaging the strength of opinion words as shown below:

$$S(B) = \frac{1}{|OW(B)|} * \sum_{i=1}^{|OW(B)|} S(OW_i) \quad (7)$$

Researcher / Entry	i	j	k	n	m
1.	W_1, W_{16}, W_3, W_2, W_{17}, W_9, W_{24}, W_{25}	W_{14}, W_8, W_6, W_7, W_{17}, W_{21}, W_{25}	W_{11}, W_7, W_2, W_9, W_{19}, W_{21}, W_{25}	W_{13}, W_{13}, W_{10}, W_{14}, W_{21}, W_{22}	W_{10}, W_{15}, W_2, W_{21}, W_{23}, W_{24}
2.	W_4, W_2, W_3, W_{14}, W_{11}, W_{18}, W_{21}, W_{23}	W_1, W_{16}, W_{11}, W_7, W_{18}, W_{17}, W_6, W_{23}	W_{14}, W_{10}, W_4, W_9, W_{19}, W_{20}	W_{11}, W_{13}, W_6, W_5, W_{20}, W_{21} W_{22}, W_{25}	W_{14}, W_{16} W_9, W_8, W_{18}, W_{23}, W_{24}
3.	W_1, W_2, W_6, W_{13}, W_{20}	W_7, W_3, W_{18}, W_8, W_{17}, W_{24}	W_9, W_{19}, W_{11}, W_{10}, W_{17}, W_{23}	W_{13}, W_{14}, W_{18}, W_{12}, W_{20}, W_{22}	W_{15}, W_{19} W_1, W_{16}, W_{20}, W_{23}, W_{24}
4.	W_1, W_2, W_4, W_8, W_{15}, W_{10}	W_6, W_6, W_7, W_{17}, W_{22}	W_{12}, W_9, W_{19}, W_{16}, W_{24}	W_{17}, W_{13}, W_2, W_{20}, W_{21}, W_{22}	W_{11}, W_{17}, W_6, W_{15}, W_{24}, W_{25}
5.	W_1, W_2, W_5, W_3, W_{19}	W_7, W_{18}, W_{15}, W_2, W_{18}, W_6, W_{17}, W_1	W_{19}, W_9, W_{17}, W_{10}, W_{10}	W_{18}, W_7, W_{13}, W_{13}, W_{20}, W_{23} W_{24}	W_3, W_{13}, W_{22}, W_{23} W_{24}, W_{25}

TABLE I. SAMPLE BLOG ENTRIES OF 5 RESEARCHERS [3]

where; |OW(B)| denotes the size of the set of opinion words extracted from the blog and $S(OW_i)$ denotes the sentiment strength of i^{th} opinion word. As the overall strength of the blog is calculated by averaging the strength of the opinion words, therefore the strength of the review i.e. S(B) will also lie in the range of -1 to +1; where, S(B) = -1 indicates a strong negative opinion, S(B) = +1 indicates a strong positive opinion and S(B) = 0 indicates a neutral opinion.

Similar to blog orientation, comment orientation, S (C), of each comment received on a particular blog is determined. Once the orientation of every comment is known, the average comment orientation, Avg. S (C), is calculated (dividing the total comment orientation by the no. of comments).

3) Blog Ranking Module: This module takes as input the blog orientation strength and the average comment orientation strength for each member to compute the blog score.

Blog Score = Blog Orientation Strength* Avg. Comment Orientation Strength

The member's blogs are are then ranked on the basis of this computed blog score. Finally, the expert is identified as the one with the highest ranking blog score.

IV. ILLUSTRATION

To clearly illustrate the use and effectiveness of the proposed system, a case study is presented to describe a typical scenario and examine the result of each module of the approach.

A. Interest Mining Module: To demonstrate the Interest mining module we directly take the sample data calculations of interest vector and interest similarity from [3], where there are 5 researchers viz. i, j, k, n & m. Therefore, $N_u = 5$ and there are 5 entries in each of the researcher's blog site. The following table I shows the blog entries of each of the Researcher i, j, k, n & m.

The key point of constructing this Collaborative Interest Group is the calculations of interest similarity relations and application of the K-means clustering technique to cluster researchers with similar interests into the same group.

Interest Vector calculations: *We have the interest vector corresponding to each of the researcher i, j, k, n & m represented as V_i, V_j, V_k, V_n, V_m. The vectors using equation (2) is shown below:*

For Researcher i:

$$V_i = (0.8874, 0.4846, 1.1938, 1.3979, 0.6989)$$

For Researcher j:

$$Vj = (0.8874, 1.9897, 0.7959, 0.4845, 0.6655)$$

For Researcher k:

$$V_k = (1.9897, 0.6655, 0.1938, 0.6988, 1.9897)$$

For Researcher n:

V_n = (1.9897, 0.1938, 0.8874, 0.2907, 0.8874)

For Researcher m:

V_m = (1.1938, 0.4436, 0.3876, 0.4845, 0.1938)

Interest Similarity Score calculations: *The calculated values of Similarity Score between each of the 2 researchers:*

R_{ij} = 0.7063; R_{ik} = 0.7110; R_{in} = 0.7502; R_{im} = 0.8064;

R_{jk} = 0.6688; R_{jn} = 0.6132 ; R_{jm} = 0.7424;

R_{kn} = 0.8786; R_{km} = 0.8140; R_{nm} = 0.9169

As all the elements of both the vectors taken at a time to calculate the similarity score are positive, thus the range of similarity score is between0 to 1.

This indicates that:

The value of 1 means that the 2 researchers have exactly similar interests and;

The value of 0 means that the 2 researchers do not have any similar interests at all.

Therefore, we can say that:

The researchers n & m have almost similar interests (as R_{nm}= 0.9169, approx 1)

The researchers k & n have similar interests to a very great extent (as R_{kn} = 0.8786)

The researchers "k & m" and "i & m" have quite a lot similar interests (as R_{km} = 0.8140 and R_{im} = 0.8064)

- The researchers "j & k" and "j & n" have quite less similar interests (as R_{jk} = 0.6688 and R_{jn} = 0.6132)

a) Collaborative Interest Group Construction

We construct the collaborative interest group by using the technique of K-means clustering algorithm with the help of two basic steps. We first construct the researcher groups by finding the membership of each of the researcher using the formula defined in equation (4). This step would give us the total number of clusters required, denoted by K. And then we assign points to the closest centroid by taking the proximity measure as the distance between two researchers using the formula defined in equation (5).

CONSTRUCTION OF RESEARCHER GROUPS

1) Membership for group i

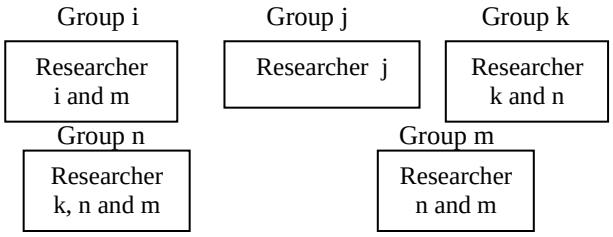
Step 1: Calculate the threshold for this group i.e. T_i

$$T_i = \frac{1}{5}[R_{ii} + R_{ij} + R_{ik} + R_{in} + R_{im}]$$
$$= \frac{1}{5}[1 + 0.7063 + 0.7110 + 0.7502 + 0.8064]$$
$$= 0.79478$$

Step 2: Deciding the members for group i

As we can see, $R_{ii} > T_i$ and $R_{im} > T_i$, therefore Researcher i and Researcher m belong to group i.

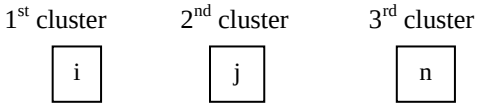
We find Membership for group j, group k, group n, group m in a similar way and the following Researcher Groups are formed with their respective members:



CONSTRUCTION OF CLUSTERS

1) Total number of clusters

Now as we know total number of clusters i.e. K is equivalent to the minimum number of groups required to cover all the data points. Therefore, K=3. In other words, we can say that there are total three number of clusters required with the centroid as i, j, and n respectively.



2) Assigning points to the closest Centroid

In this step we assign points (researcher m and k) to the closest centroid by taking the proximity measure as the distance between two researchers. Therefore using the formula defined in equation (5), we calculate the distance of these two researchers with each of the above researchers:

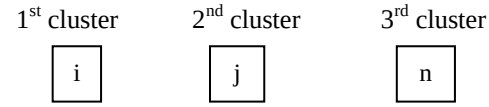
d_{ki} = 0.289; d_{kj} = 0.3312; d_{kn} = 0.1214

Since d_{kn} is minimum, therefore researcher k belongs to the 3rd cluster with centroid as n.

d_{mi} = 0.1936; d_{mj} = 0.2576; d_{mn} = 0.0831

Similarly, Since d_{mn} is minimum, therefore researcher m also belongs to the 3rd cluster with centroid as n.

So, after the first iteration we have the following clusters:



Now, the 2nd iteration begins. We recompute the centroid of the 3rd cluster.

Distance between each of the two researchers is as follows:

d_{ij} = 0.2937; d_{in} = 0.2498; d_{jn} = 0.3868;

d_{km} = 0.186; d_{ki} = 0.289; d_{kj} = 0.3312;

d_{kn} = 0.1214; d_{mi} =0.1936; d_{mj} = 0.2576; d_{mn} = 0.0831

Assuming n to be the centroid:

$S1 = d_{nm} + d_{nk}$ = 0.1214 + 0.0831= 0.2045

Assuming m to be the centroid:
$S2 = d_{mk} + d_{mn} = 0.186 + 0.0831 = 0.2691$

Assuming k to be the centroid:
$S3 = d_{km} + d_{kn} = 0.186 + 0.1214 = 0.3074$

Since S1 is minimum, therefore n remains the centroid.

B. Expert Mining module:

As described in section III, the expert mining module is divided into three sub-modules, namely the data respository; sentiment mining & blog ranking modules, here we demonstrate them & examine their effectiveness. We consider a group with 4 members and analyze their blogs & cooments received on them. Our final task is to determine the expert from this group of 4 members.

BLOG 1:-

> The colours are boring. The headlights are not very strong and rear seats are less comfortable. There's hardly any boot space. The ride is not too bad, but there is a little stiffness and it crashes over sharp bumps. Ground clearance is very poor and is unstable at high speeds above 100km/h.

Comments:

1) Yes there's very little boot space.
2) I agree.
3) There are only 3 colours available.
4) I think the ride is quite good.
5) Instability at high speeds is a major drawback.
6) According to me, the seats are very comfortable.
7) Ground clearance is a very common issue in India!

The Sentiment mining module works in the following manner:

Feature Extraction: Each of the sentences along with their POS tag information is saved in the Repository. Sample XML for the blog 1 about car above:

```
<S>
<ART><WC = 'the'> the</W></ART>
<N><WC = 'colours'> colours </W></N>
 <V><WC = 'are'> are </W></V>
<A><WC='boring'>boring</W></A>
</S>

<S>
<ART><WC = 'the'> the</W></ART>
<N><WC = 'headlights'> headlights </W></N>
<V><WC = 'are'> are</W></V>
<AG><WC='not'>not</W><WC='very'>very</W><WC=
'strong'> strong</W></AG>
<CONJ><WC='and'>and</W></CONJ>
<N><WC = 'rear-seats'> rear- seats</W></N>
<V><WC = 'are'> are</W></v>
<AG><WC='less'>less</W><WC='comfortable'>comfortable</W>
</S>

<S>
<P><WC = 'There'> there</W></P>
<V><WC = 'is'> is</W></V>
<AG><WC = 'hardly'> hardly</W><WC = 'any'> any</W></AG>
<N><WC = 'boot space'> boot space</W></N>
</S>

<S>
<ART><WC = 'the'> the</W></ART>
<N><WC = 'ride'> ride </W></N>
<V><WC = 'is'> is</W></V>
<AG><WC = 'not'> not</W><WC = 'too'> too</W><WC =
'bad'> bad</W></AG>
<WC = ','>, </W>
<CONJ ><WC = 'but'> but</W></CONJ>
<P><WC = 'There'> there</W></P>
<V><WC = 'is'> is</W></V>
<ART><WC = 'a'> a</W></ART>
<A><WC = 'little'> little</W></AG>
<N><WC = 'stiffness'> stiffness</W></N>
<CONJ><WC = 'and'> and</W></CONJ>
<N><WC = 'it'> it</W></N>
<V><WC = 'crashes'> crashes</W></V>
<P><WC = 'over'> over</W></P>
<A><WC = 'sharp'> sharp</W></A>
<N><WC = 'bumps'> bumps</W></N><WC='.'>.</W>
</S>

<S>
<N><WC = 'Ground clearance'>Ground clearance</W></N>
<V><WC = 'is'> is</W></V>
<AG><WC='very'>vary</W><WC='poor'>poor</W>
<CONJ><WC = 'and'> and</W></CONJ>
<V><WC = 'is'> is</W></V>
<A><WC='unstable'>unstable</W></A>
<P><WC = 'at'> at</W></P>
<N><WC = 'high speeds'> high speeds</W></N>
<P><WC = 'above'> above</W></P>
<N><WC = '100km/h'> 100km/h</W></N>
</S>
```

To determine the opinion word orientation, we establish the Adjective Polarity & **the Adverb Strength:** For adjective polarity, we use a set of seed adjectives whose orientations we know, & then grow this set by searching in the WordNet. We consider the following initial Adjective Seed-List, shown in Table II (with positive & negative orientations):-

TABLE II. SEED LIST OF ADJECTIVES

Positive Orientation	Negative Orientation
Great	Sharp
Blend	Dirty
Amazing	Sick
Compact	Unfortunate
Affordable	Bad
Reasonable	Boring
Excellent	Nasty
Big	Wrong
Fast	Poor
Comfortable	Awful
Strong	Scary
Beautiful	Dull
Impressive	Inferior
Good	Unstable
Exciting	Jerky
Stiff	Noisy
Variety	Common
Smooth	Okay okay

High	Bulky
Value-for-money	Low
Spacious	Drawback
Effective	
Major	
Attractive	
Stylish	
Streamlined	
Maneuverable	
Better	
Value for money	

We manually mark the strengths of a few frequently used adverbs with values ranging from -1 to +1 based on our intuitions. We consider the most frequently used adverbs (for our illustration) along with their strength as below in table III:-

TABLE III. ADVERB STRENGTHS

Adverb	**Strength**
Complete	+1
Most	0.9
Extremely	0.8
Absolutely	0.7
Too	0.7
Very	0.6
Indeed	0.6
More	0.4
Much	0.3
Reasonably	0.2
Any	0.1
Quite	-0.2
Pretty	-0.3
Little	-0.4
Less	-0.6
Not	-0.8
Never	-0.9

Opinion Strength Calculations: The strength of each opinion word is given by the formula defined in equation (7)

Opinion Words (for blog):

1. boring -1 * +0.5 = -0.5
2. not very strong -0.8 * +0.6 * +1 = -0.48
3. less comfortable -0.6 * +1 = -0.6
4. hardly any -1 * +0.1 = -0.1
5. not too bad -0.8 * +0.7 * -1 = +0.56
6. little stiff -0.4 * -1 = +0.4
7. sharp -1 * +0.5 = -0.5
8. unstable -1 * +0.5 = -0.5

Total Blog Orientation Strength = S (B_1) =
(- 0.5 - 0.48 - 0.6 - 0.1 + 0.56 + 0.4 - 0.5 - 0.5) / 8 = - **0.215**

Opinion Words (for comments):

1. very little +0.6 * -0.4 = -0.24
2. quite good -0.2 * +1 = -0.2
3. major drawback -1 * +1 * +0.5 = -0.5
4. very comfortable +0.6 * +1 = +0.6
5. very common +0.6 * -1 = -0.6

Average Comment Orientation Strength = Avg. S (C_1) =
(- 0.24 - 0.2 - 0.5 + 0.6 - 0.6) / 5 = -0.188/5 = **-0.0376**

Blog Score$_1$ = S (B_1) *Avg. S (C_1) = **-0.215** * **-0.0376** = **+0.008**

BLOG 2:-

> The drive is reasonably smooth but gets jerky at higher speeds. Only manual transmission is available and that too is a little poor. The diesel model has a very noisy engine even for a new car. There is a very good variety of colours and a reasonably high mileage. All in all, it's value for money and a good buy.

Comments:

1) Jerky drive.
2) Mileage is good.
3) Its an okay okay buy.
4) Colour choices are good.
5) Engine is a little noisy.
6) Transmission is good.
7) I found it to be a smooth car.

Now we calculate the S (B_2) & Avg. S(C_2) to compute blog score $_2$

Opinion Words (for blog):

1. reasonably smooth 1* 0.2= 0.2
2. jerky -1 * 0.5=-0.5
3. little poor -1*-0.4= 0.4
4. very noisy -1*0.6= -0.6
5. very good 1*0.6= 0.6
6. reasonably high 1*0.2=0.2
7. good 1*0.5= 0.5
8. value for money 1*0.5=0.5

Total Blog Orientation Strength = B (S_2) =
(0.2+ (-0.5) + 0.4 + (-0.6) + 0.5 +0.2+ 0.5 + 0.5) / 8= +1.2/8 = **+0.15**

Opinion Words (for comments):

1. Jerky -1 *0.5= -0.5
2. Good 1* 0.5= 0.5
3. Okay okay -1* 0.5= -0.5
4. Good 1* 0.5= 0.5
5. Little noisy -0.4* -0.6= 0.24
6. Good 1* 0.5= 0.5
7. Smooth 1*0.5= 0.5

Average Comment Orientation Strength = Avg. S (C_2) =
(-0.5 +0.5 + (-0.5) + 0.5 + 0.24 +0.5 + 0.5) / 7= (+1.24)/7 = **+0.1771**

Blog Score$_2$ = S (B_2)* Avg. S(C_2) = **0.15 * 0.1771**= **+ 0.02656**

BLOG 3:-

> This car is a complete blend of great power and style, with exciting features. It has very good fuel efficiency and engine is pretty impressive too. It's very spacious for its size and the drive is absolutely smooth. It has got beautiful interiors and the compact dimensions make it an excellent traffic warrior.

Comments:

1) Good review!
2) Interiors are indeed attractive.
3) Engine is a little noisy guys.
4) Car is quite stylish!
5) The car is spacious but bulky too!

Now we calculate the S (B_3) & Avg. S (C_3) to compute blog score $_3$

Opinion Words (for blog):

1. complete blend +1 * +1 * +0.5 = +0.5
2. great +1 * +0.5 = +0.5
3. exciting +1 * +0.5 = +0.5
4. very good +0.4 * +1 = +0.4
5. pretty impressive -0.3 * +1 = -0.3
6. very spacious +0.4 * +1 = +0.4
7. absolutely smooth +0.7 * +1 = +0.7
8. beautiful +1 * +0.5 = +0.5
9. compact +1 * +0.5 = +0.5
10. excellent +1 * +0.5 = +0.5

Total Blog Orientation Strength = B (S_3) =

(+ 0.5 + 0.5 + 0.5 + 0.4 - 0.3 + 0.4 + 0.7 + 0.5 + 0.5 + 0.5) / 10 = **+0.42**

Opinion Words (for comments):

1. good +1 * +0.5 = +0.5
2. quite stylish -0.2 * +1 = -0.2
3. indeed attractive +0.6 * +1 = +0.6
4. little noisy -0.4* -0.6= +0.24
5. spacious +1 * +0.5 = +0.5
6. bulky -1 * +0.5 = -0.5

Average Comment Orientation Strength = Avg. S (C_3) =

(+ 0.5 - 0.2 + 0.6 + 0.24 + 0.5 - 0.5) / 5 = (+1.14)/5 = **+0.228**

Blog Score$_3$ = S (B_3)* Avg. S (C_3) = +0.42 * +0.228 = +**0.09576**

BLOG 4:-

> The size of this car is never big and this makes its price pretty reasonable and affordable. It does not demand any maintenance and its performance and safety are also amazing. Not much of car service is required. The cooling is very effective and this car is not very smooth on hilly terrains.

Comments:

1) Yes, very low maintenance required.
2) The ride is not very smooth.
3) Affordable price .Just right for middle class families.
4) Streamlined shape.
5) Cooling is not good especially for Delhi summers.
6) Its a good buy.
7) Better cars are available in the market.

Now we calculate the S (B_4) & Avg. S (C_4) to compute blog score$_4$

Opinion Words (for Blog):

1. never big 1* -0.9= -0.9
2. pretty reasonable 1 *-0.3=-0.3
3. Affordable 1*0.5=0.5
4. any 0.5*-0.1= -0.05
5. amazing 1*0.5=0.5
6. not much 0.5*0.3*-0.8=-0.12
7. very effective 1* 0.6= 0.6
8. not very smooth 1* -0.8* 0.6= -0.48

Total Blog Orientation Strength = B (S_4) =
-0.9 + (-0.3) +0.5 + (-0.05) +0.5 + (-0.12) +0.6+ (-0.48) / 8 = (-0.25) / 8 = **-0.03125**

Opinion Words (for Comments)

1. very low -1*0.6= -0.6
2. not very smooth 1*-0.8*0.6=-0.48
3. Affordable 1*0.5=0.5
4. Streamlined 1*0.5=0.5
5. not good 1*-0.8=-0.8
6. good 1*0.5=0.5
7. Better 1*0.5=0.5

Average Comment Orientation Strength = Avg. S (C_4) =
(-0.6 + (-0.48) +0.5 +0.5 + (-0.8) + 0.5 +0.5) / 7 = (0.12) / 7 = **+0.0171**

Blog Score$_4$ = S (B_4)* Avg. S (C_4) = -0.03125* 0.171= **-0.0005**

TABLE IV. BLOG RANKING

Blog	***Blog Score***	***Blog Rank***
Blog1	+0.008	3
Blog 2	+0.02656	2
Blog 3	+0.09576	1
Blog 4	-0.0005	4

Thus, comparing all the blog strengths, according to our approach, the highest blog score is for blog 3 and therefore the Expert is blogger 3!

Limitations:

1) It covers comments only written in English.
2) No abbreviations or acronyms can be accounted for.
3) It does not cover interrogative sentences.
4) A negative adjective and a negative adverb convert into a positive opinion word.
5) A positive adjective and a negative adverb also convert into a negative opinion word.
6) The method has no way of detecting and dealing with emoticons.

V. CONCLUSION

We proposed a novel ComEx Miner System for mining experts in virtual communities. This work is exploratory in nature and the prototype evaluated is a preliminary prototype. The major contributions of this research are:

i. Constructing a collaborative interest group known as the virtual community which will cluster researchers with similar interests in a same group and thereby facilitate collaborative work.

ii. Accessing the expertise from the virtual community using sentiment analysis of each group member's blog & comments received on it. Their combined orientation strength determined the blog score which enabled to rank the blogs and identify the expert as the one with the highest blog rank.

The practice result proves that this algorithm has the characteristics of highly effective group arranging and identifying expert. This study is just one step in this direction. Due to the complex nature of framework, it is impossible to consider and incorporate all the factors that could have an impact on the effectiveness and efficiency of this system. . More research needs to be done in order to validate or invalidate these findings, using larger samples.

REFERENCES

[1] L. Colazzo, A. Molinari and N. Villa., "Collaboration vs. Participation: the Role of Virtual Communities in a Web 2.0 world". International Conference on Education Technology and Computer, 2009, pp.321-325.

[2] H.Rheingold, "The Virtual Community: Homesteading on the Electronic Frontier", revised edition. The MIT Press, 2000.

[3] Kumar, A. & Bhatia, MPS., "Community Expert based Recommendation for solving First Rater Problem". International Journal of Computer Applications (IJCA), Vol. 37, No.10, 7-13, January 2012, Foundation of Computer Science, USA.

[4] Kumar, A. & Jain, A., "An Algorithmic Framework for Collaborative Interest Group Construction". Recent Trends in Networks and Communications, LNCS-CCIS, Springer, Volume 90, Part 3, 2010, pp.500-508.

[5] Balog, K., Bogers, T., Azzopardi, L., de Rijke, M., van den Bosch, A.: "Broad expertise retrieval in sparse data environments". In: SIGIR 2007, pp. 551–558.

[6] Bogers, T., Kox, K., van den Bosch, A. "Using Citation Analysis for Finding Experts in Workgroups". In: Proc. DIR. (2008)

[7] John G. Breslin, Uldis Bojars, Boanerges Aleman-Meza, Harold Boley, Malgorzata Mochol, Lyndon J. B. Nixon, Axel Polleres, and Anna V. Zhdanova. "Finding experts using Internet-based discussions in online communities and associated social networks". First International ExpertFinder Workshop, Berlin, Germany, January 16, 2007, 2007.

[8] F. Metze, C. Bauckhage, T. Alpcan, K. Dobbrott, and C. Clemens, "A community-based expert finding system," in IEEE Int. Conf. on Semantic Computing (ISCS 2007), Irvine, CA, September 2007.

[9] Schall, D. & Dustdar, S. "Dynamic Context-Sensitive PageRank for Expertise Mining". SocInfo 2010, pp. 160-175

[10] Huh, E.-N., Lee, P. and Newby, G. "A Framework of Online Community based Expertise Information Retrieval on Grid". OGF Document Series GFD-I.164, January 2010.

[11] E. Smirnova and K. Balog. "A User-oriented Model for Expert Finding". In: 33rd European Conference on Information Retrieval (ECIR 2011), LNCS 6611,2011,pp. 580-592.

[12] Jie L"i, Harold Boley, Virendrakumar C. Bhavsar, and Jing Mei. Expert finding for eCollaboration using FOAF with RuleML rules". Montreal Conference on eTechnologies (MCTECH), 2006.

[13] Punnarut, R. & Sriharee, G. "A Researcher Expertise Search System using Ontology-Based Data Mining". Proceedings of Seventh Asia-Pacific Conference on Conceptual Modelling, APCCM'10, 2010, Volume 110, pp. 71-78.

[14] Zhang, J., Ackerman, M.S., Adamic, L. "Expertise Networks in Online Communities: Structure and Algorithms". The 16th international conference on World Wide Web, 2007, PP.7-16.

[15] Tang, J., Zhang, J., Zhang, D., Yao, L. and Zhu, C. "ArnetMiner: An Expertise Oriented Search System for Web Community. Semantic Web Challenge". In Proceedings of the 6th International Conference of Semantic Web (ISWC'2007).

[16] Pang-Ning Tan, Michael Steinbach and Vipin Kumar. "Introduction to Data Mining". Pearson Education.

[17] POS Tagger: http://www.infogistics.com/textanalysis.html

[18] WordNet: http://wordnet.princeton.edu/

AUTHORS PROFILE

Akshi Kumar is a PhD in Computer Engineering from University of Delhi. She has received her MTech (Master of Technology) and BE (Bachelor of Engineering) degrees in Computer Engineering. She is currently working as a University Assistant Professor in Dept. of Computer Engineering at the Delhi Technological University, Delhi, India. She is editorial review board member for '*The International Journal of Computational Intelligence and Information Security'*, Australia, ISSN: 1837-7823; *'International Journal of Computer Science and Information Security'*, USA, ISSN: 1947-5500; '*Inter-disciplinary Journal of Information, Knowledge & Management'*, published by the Informing Science Institute, USA. (ISSN Print 1555-1229, Online 1555-1237) and '*Webology*', ISSN 1735-188X. She is a life member of Indian Society for Technical Education (ISTE), India, a member of International Association of Computer Science and Information Technology (IACSIT), Singapore, a member of International Association of Engineers (IAENG), Hong Kong, a member of IAENG Society of Computer Science, Hong Kong and a member of Internet Computing Community (ICC), AIRCC. She has many publications to her credit in various journals with high impact factor and international conferences. Her current research interests are in the area of Web Search & Mining, Intelligent Information Retrieval, Web 2.0 & Web Engineering.

Nazia Ahmad is doing M.Tech (Master of Technology) in Computer Technology & Application from Delhi Technological University, Delhi, India and has done her B.Tech (with Distinction) also in Computer Science and Engineering. She is currently working as an Assistant Professor in Dept. of Computer Engineering at the Delhi Technological University, Delhi, India.