Appl Intell (2013) 39:1–13
DOI 10.1007/s10489-012-0389-1

A semantic social network-based expert recommender system

Elnaz Davoodi · Keivan Kianmehr · Mohsen Afsharchi

Published online: 12 October 2012

Abstract This research work presents a framework to build a hybrid expert recommendation system that integrates the characteristics of content-based recommendation algorithms into a social network-based collaborative filtering system. The proposed method aims at improving the accuracy of recommendation prediction by considering the social aspect of experts' behaviors. For this purpose, content-based profiles of experts are first constructed by crawling online resources. A semantic kernel is built by using the background knowledge derived from *Wikipedia* repository. The semantic kernel is employed to enrich the experts' profiles. Experts' social communities are detected by applying the social network analysis and using factors such as experience, background, knowledge level, and personal preferences. By this way, hidden social relationships can be discovered among individuals. Identifying communities is used for determining a particular member's value according to the general pattern behavior of the community that the individual belongs to. Representative members of a community are then identified using the eigenvector centrality measure. Finally, a recommendation is made to relate an information item, for which a user is seeking an expert, to the representatives of the most relevant community. Such a semantic social network-based expert recommendation system can provide benefits to both experts and users if one looks at the recommendation from two perspectives. From the user's perspective, she/he is provided with a group of experts who can help the user with her/his information needs. From the expert's perspective she/he has been assigned to work on relevant information items that fall under her/his expertise and interests.

Keywords Semantic information extraction · Social network analysis · Expert recommender system · Knowledge management

E. Davoodi
Department of Computer Science, Institute for Advanced Studies in Basic Sciences, Zanjan, Iran
e-mail: elnazood@gmail.com

K. Kianmehr (✉)
Department of Electrical and Computer Engineering, Western University, London, Canada
e-mail: kkianmeh@uwo.ca

M. Afsharchi
Department of Electrical and Computer Engineering, University of Zanjan, Zanjan, Iran
e-mail: afsharchim@znu.ac.ir

1 Introduction

In the past decade, lots of major internet retailers have begun to build recommender systems to personalize content to show to their users through an information filtering process. The principal objective of any recommendation system is that of complexity reduction for the human being, sifting through very large sets of information and selecting those pieces that are relevant to the active user's needs. Thus, all of significant recommender systems create a profile for each user and then use one of recommendation approaches to make recommendations about information items a user might be interested in. Generally, these recommendation approaches are classified into three groups based on the way that the user models are constructed, the employed prediction methods, and also the type of items to be recommended [1]. These three groups are content-based [2], collaborative filtering [3], and hybrid methods [4]. Content-based methods provide recommendations by comparing characteristics of content contained in an item to

characteristics of content that the user is interested in (e.g. keywords describing the items, such as movie genre, artists, etc., for movies). In other words this approach treats items as objects with many attributes, some of which are shared across items [5, 6]. The goal of the system is to discover which attributes about items a user likes, and to find out which attributes an item has. However, the problem with this approach is that it is limited to dictionary-bound relations between the keywords used by users and the descriptions of items and therefore does not explore implicit associations between users [7]. Collaborative filtering is an approach to make recommendations by looking for users who share the same rating patterns with the active user (the user whom the prediction is for). Then the ratings from those like-minded users are used to calculate a prediction for the active user [8, 9]. The problem with this approach is that first we have to decide on a rather arbitrary basis over the number of like-minded users. Second, the prediction model built by the system might not generalize in different contexts, that is similar users in the context of movies may be dissimilar users in the context of music albums. To overcome the certain limitation of these two kinds of recommender approaches, some hybrid approaches have been created by combining the content-based and collaborative filtering recommenders. In general, the goal of recommendation systems is to provide personalized recommendations of items to users based on their previous behavior, item descriptions, and user profiles, however recommendations are not made in isolation. An approach that has recently received much attention is to use the social structure of users' informal relationships as an additional source of information in recommender systems. It seems more rational to deliver recommendations within an informal community of users and a social context [10]. The social component of a recommendation is sometimes more important than the user behavior in understanding the decision making process of the user [18]. The social embedding of the recommendation systems is determined by factors such as experience, background, knowledge level, beliefs and personal preferences [11].

In fact, social networks (SN) are an attractive way to model the interaction among the people in a group or community. It provides a rich source of information regarding users' communities and their correlation. The information provided by a social network can be integrated into the engine of a recommender system to build more rational prediction model that takes into consideration not only the item descriptions and user behavior in isolation but also the social behavior of the user. The basic and the most time consuming step in the process of social network analysis (SNA) is to build the social network. Once the social network is constructed, different measures can be applied to study the characteristics of the social network and hence it is necessary to decide correctly and clearly on elements of the model (actors) and the interactions between them (ties or connections). Indeed, in common methods of the social network construction, the semantic associations among individuals are not considered and their relationships may be falsely built. Discovering semantic relationships among entities of a social network which leads to a semantic-based social network is a promising solution to this problem. To discover semantic relationships, in this research we make use of *Wikipedia*.

In addition, identifying and classifying experts is an emerging research area that has already attracted the attention of many research groups. One objective in exploring the experts is to facilitate the process of finding the right people whom we may ask a specific question and who will answer that question for us. Finding the appropriate experts in the knowledge management system is not a straightforward task due to the complexity and diversity of the expertise and the knowledge needs. Further, classifying the type of knowledge that different people have is a challenging problem. An expert carries a type of knowledge, namely tacit knowledge that is gained through experience and learning over time and is hard to be codified. One example of tacit knowledge is experience. People gain knowledge through experience and they are not often aware of the knowledge they possess or how it can be valuable to others. Effective transfer of tacit knowledge generally requires extensive personal contact and trust which is not feasible all the time. Tacit knowledge usually resides in the expert's brain. Therefore finding relevant experts for a particular task is challenging.

This research work presents a hybrid expert recommendation system which is indeed a semantic social network-based collaborative strategy that it also maintains the content-based profiles for each user. One advantage of this approach is that users can be recommended an item not only when this item is rated highly by users with similar profiles, but also directly, i.e., when this item gets highly scored against the user's profile. In the domain of the expert recommendation system, our proposed system discovers communities of experts and accordingly assists users to effectively find groups of experts who carry users' desired tacit knowledge. In this context, the social structure of the experts' relations, captured in a semantic social network, is used as the social component of the recommendation system. The semantic social network of experts is constructed based on factors such as experience, background, knowledge level, and personal preferences of experts.

The rest of this paper is organized as follows. Section 2 describes the architecture of the proposed social network-based recommender system, its components and algorithms. In Sect. 3, the results obtained from the experiments conducted to test different aspects of the expert recommender system are demonstrated. In addition, results from the user

study that has been conducted to validate the expert recommender system are reported. Finally, the paper is concluded in Sect. 4.

2 The proposed framework

Proposed in this paper is a general framework that attempts to detect communities of experts in a social network and to build a recommendation system based on the information extracted from expert communities. The ultimate goal of the system is to recommend experts who have the appropriate knowledge with regards to the user information needs. In the proposed framework, the expert recommendation system is built in four major phases, as depicted in Fig. 1:

1. building textual profiles for experts based on the information extracted from the web,
2. extracting background knowledge to construct a semantic kernel, based on the *Wikipedia* concepts to enrich the experts' profiles,
3. constructing a semantic-based social network of experts and detecting experts communities,
4. building the expert recommendation system and recommending an expert community or several active members of a community who can respond to a user's specific information need.

In the first phase, a profile is constructed for each individual expert in a specific domain. Each profile is in a textual format, namely a text document, and contains a variety of information types, such as expertise and experience, related to a particular expert. This information is collected from different online sources on the web. In the second phase, the main task is to embed background knowledge derived from *Wikipedia* into a semantic kernel, which is then used to enrich the experts' profiles constructed in the first phase and convert them into semantic-based profiles. In the third phase, a social network is constructed according to the similarities among the experts' profiles, which were enriched with the semantic knowledge in the second phase. Afterwards, communities are detected in the constructed social network. In addition, for each community, using a social network centrality measure, an individual is chosen as the representative of the community. In the last phase, a prediction is made to recommend an expert community that has required expertise to fulfill the user's specific information need.

A case study has been conducted to assess and evaluate the effectiveness and usability of the proposed expert recommendation system in the real world.

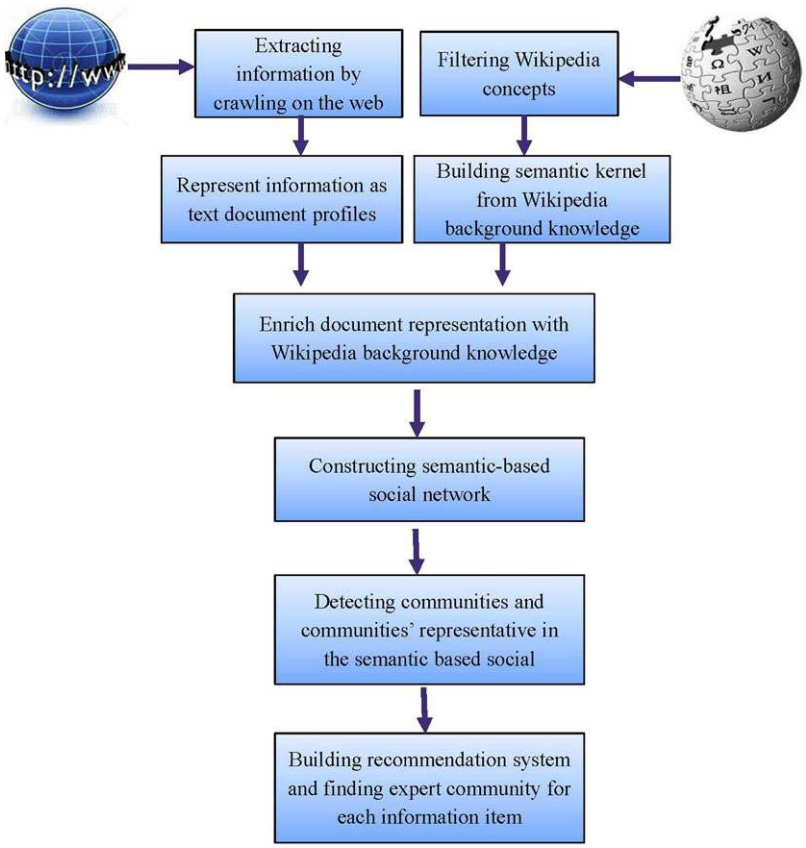

Fig. 1 The model architecture

2.1 Constructing an expert's profile

In order to build a social network of individuals, the first step is to extract relevant information about individuals and create a profile for each of them. Eventually, the social structure of individuals, with respect to the semantic-based similarities among their profiles, is modeled in a social network. To build an expert's profile, different types of relevant information needs to be collected. The manual entering mean for each expert is a very time consuming task and obviously not feasible. Therefore, profile information is extracted by crawling the web and extracting information to create a profile from relevant web pages to the corresponding individual. Profiles constructed in this manner can contain any relative information to each individual such as work experience, educational history, social and political activities, abilities and specialties, interests, etc. Next, constructed profiles will be enriched by background knowledge extracted from *Wikipedia*.

2.2 Semantic enrichment of an expert's profile

The "Bag of Words" (BoW) model has been widely used in traditional text analysis and information retrieval methods. In this model, a text (such as a document) is represented as a vector with a dimension corresponding to each word of the dictionary, containing all the words that appear in the corpus. The value associated to a given term reflects its frequency of occurrence within the corresponding document (term frequency or *tf*) and within the entire corpus (inverse document frequency or *idf*). Apparently, the BoW approach is limited since it only uses the set of terms explicitly mentioned in the document and ignores relationships between important terms that do not co-occur literally. For instance, if two documents use different collections of core words to represent the same topic, they may be falsely assigned to different clusters due to the lack of shared core words, although the core words they use are probably synonyms or semantically associated in other forms. For example, if two documents are about wireless communication, but one of them uses cellphone and the other one uses mobile phone as core word, they may be falsely assigned to different clusters in spite the fact that both of them have the same topic and use synonym core words.

The most common way to solve this problem is to enrich document representation with the background knowledge. There exist some ontologies like Word Net and Mesh [12–14, 16, 17] which were used as the external sources for embedding background knowledge to text documents, but these ontologies are manually built [15]. Data mining techniques have also been applied for this purpose. For instance, Eyharabide *et al.* in [20] proposes an agent-based method in which personal agents gather information about users in a user profile. Further, they semantically enrich a user profile with contextual information by using association rules, Bayesian networks and ontologies in order to improve agent performance. At runtime, they learn which the relevant contexts to the user are based on the user's behavior observation. Then, they represent the relevant contexts learnt as ontology segments. However, the coverage of the above methods is too restricted and their maintenance requires extreme effort as well. For these reasons, as a more feasible solution, *Wikipedia* has been recently used for text representation enrichment [19]. *Wikipedia* is a well-formed document repository such that each article only describes a single topic. The title of each article is a succinct phrase which is considered as a concept. Equivalent concepts are related to each other by redirected links which are specified by the same page on the web. Meanwhile, each article (concept) belongs to at least one category and categories are organized in a hierarchical structure. All these features make *Wikipedia* a proper ontology which excels other ontologies to embed semantic information in text documents and therefore improve the similarity measure based on the document content.

2.2.1 Extracting background knowledge from Wikipedia

To extract semantic knowledge from *Wikipedia*, a content-based method is applied to enable system find proximity between *Wikipedia* concepts, thus connections between concepts can be established. In this method, each *Wikipedia* article is represented by a *tf/idf* vector. The similarity between concepts are measured by computing the cosine similarity of their corresponding vectors. Then, a symmetric concept-concept matrix, called semantic kernel S, is created to present similarities among all pairs of *Wikipedia* concepts. Each element $S_{i,j}$ of this matrix determines the cosine similarity between a pair of concepts with indexes i and j, where $i, j \in \{1, 2, \ldots, c\}$ and c is the total number of concepts. If a row and a column refer to the same concepts or two synonym concepts, the similarity value is 1. Note that queries on synonym concepts are redirected to the same page by *Wikipedia*. Further, the more similar two corresponding concepts are, the higher the value of the corresponding entry is. This kernel represents semantic relationships among all *Wikipedia* concepts according to similarities of their corresponding articles. The semantic knowledge embedded in the semantic kernel is then integrated into the textual profiles obtained from previous phase to enrich the representation of the documents. Figure 2 shows semantic kernel of *Wikipedia* concepts which is a $c \times c$ symmetric matrix where c is the number of *Wikipedia* concepts and $S_{i,j}$ denotes the semantic similarity between two *Wikipedia* concepts.

2.2.2 Integrating background knowledge into experts' profiles

To integrate the semantic knowledge, embedded in the semantic kernel, into the text document profiles, first a connec-

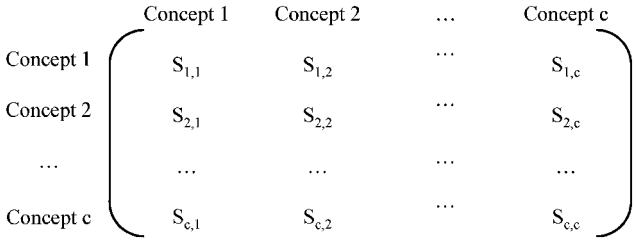

Fig. 2 The semantic kernel

tion should be established between text document profiles and the *Wikipedia* concepts. For this purpose, a proper mapping method is required. In our work, two mapping schemes are adapted: concepts match and information item relatedness. Concept match scheme maps the text document profiles to the *Wikipedia* concepts directly and in information item relatedness scheme, information items are used as features to connect text document profiles to *Wikipedia* concepts.

2.2.2.1 Concept match scheme In concept match scheme, all textual profiles are scanned in order to discover relationships between the *Wikipedia* concepts and each profile. This way, a document-concept matrix is calculated so that each entry in this matrix shows the similarity between a document and a concept. The *tf/idf* method is applied to calculate the similarity between an expert profile and a concept. In order to apply *tf/idf* method, expert profiles are considered as a collection of documents and each concept is considered as a phrase query which can be assumed a short text document. Documents and queries are presented as vectors. Dimensions represent all the words that appear in all documents. In addition, all operations that are applied for documents in *tf/idf* approach, like porter stemmer or removing stop words, now are employed for concepts that are assumed as query phrases. Finally the cosine similarity is used to measure similarity between corresponding vectors of document profiles and *Wikipedia* concepts. The resulted document-concept matrix (*DC*) in which a row entry represents a profile, columns are *Wikipedia* concepts, and each element $DC_{i,j}$ denotes the cosine similarity between a document i and a concept j of *Wikipedia*, where $i \in \{1, 2, 3, \ldots, n\}$, $j \in \{1, 2, \ldots, c\}$, n is the number of documents, and c is the number of concepts.

Once the document-concept similarity matrix is built, the background knowledge can be integrated into the text documents by embedding the semantic kernel, constructed by the process which was described in the previous section, into the similarity matrix. By combining these two matrices linearly, the semantic relationships between concepts and documents are measured. The advantage of this linear combination, that computes the semantic relationship between a concept and a document, is that not only the occurrence of the concept in the document is taken into the consideration, but also occurrences of all other concepts, in that particular document, that may have some relations with the given concept are considered. In fact, when the semantic relationship between a concept and a document is calculated, the occurrences of other concepts in that document that are more similar to the given concept have more impact on the semantic relationship between the concept and the document than the occurrences of other dissimilar concepts. By performing the linear combination of the semantic kernel (*S*) and the document-concept matrix (*DC*) a new semantic-based document-concept similarity matrix (*SDC*) is generated. In the semantic-based document-concept similarity matrix *SDC*, each entry $SDC_{i,j}$, which shows the semantic relationship between a corresponding document with index i and a corresponding concept with index j, is calculated by multiplying the ith row of the document-concept similarity matrix (*DC*) and the jth column of the semantic kernel *S*.

$$SDC_{i,j} = \sum_{j=1}^{c} DC_{i,j} \times S_{j,i},$$

where c is the number of concepts. As can be seen in the above equation, the semantic relationship between the jth concept and the ith document is influenced by the weights of similarities between the jth concept and all other concepts. In other words, the weight by which a concept impacts the semantic relationship between the jth concept and the ith document is equal to the similarity between that particular concept and the jth concept. Figure 3 shows the semantic-based document-concept similarity matrix resulted from the linear combination of a document-concept matrix *DC* and a semantic kernel *S* where n and c indicate the number of profiles and *Wikipedia* concepts respectively.

2.2.2.2 Information item relatedness scheme In this approach, *tf/idf* method is applied to represent profiles with respect to information items as features and as a result a document-information item similarity matrix (*DI*) will be generated in which each entry indicates the cosine similarity between corresponding information item and profile. Then, for document enrichment with the semantic knowledge embedded in the semantic kernel, a connection should be built to connect information items to *Wikipedia* concepts. Since information items and *Wikipedia* concepts are both short phrases, the occurrences of concepts in information items are very rare. Indeed, a concept and an information item may have some common words, but the whole concept phrase does not have any exact match in the information item set. Therefore, if only the exact match occurrence of information items in *Wikipedia* concepts is considered, the similarity matrix will be a sparse matrix; however there are many common words between information items and *Wikipedia* concepts.

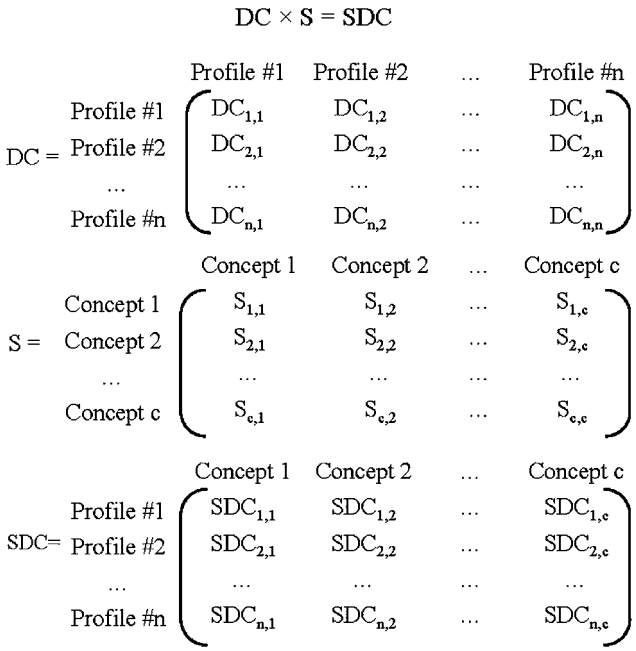

Fig. 3 Linear combination of document-concept similarity matrix (*DC*) and semantic kernel (*S*) produces semantic document-concept similarity matrix (*SDC*)

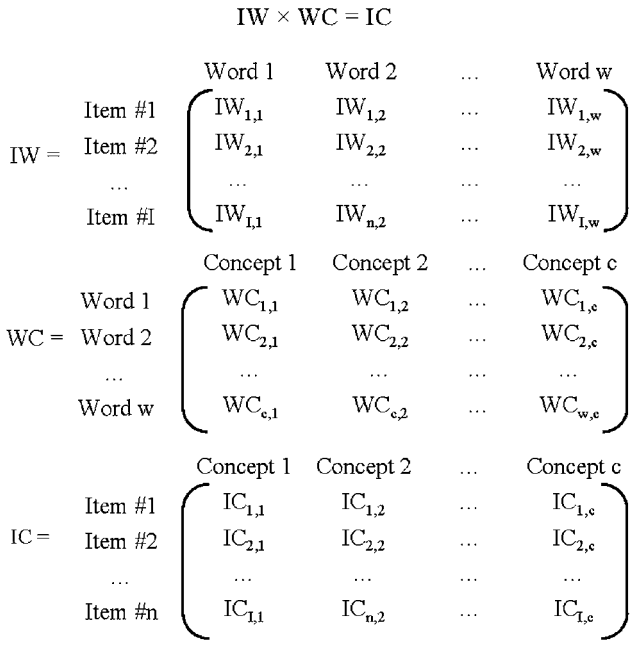

Fig. 4 Linear combination of Information item-word similarity matrix (*IW*) and word-concept similarity matrix (*WC*) produces information item-concept Similarity Matrix (*IC*)

To overcome this problem and make an accurate connection between information items and *Wikipedia* concepts, an alternative method is applied in which *Wikipedia* concepts are first broken down into words. Once the *Wikipedia* concepts are broken down into words, a collection of distinct words are created. Then a connection should be built between theses words and information items and as a result an information item-word similarity matrix (*IW*) is constructed in which each entry demonstrates the similarity between an information item and a word that is a part of at least one *Wikipedia* concept. To measure this similarity, *tf/idf* method with cosine similarity measure is applied. The set of information items is treated as a set of documents and each word is treated as a query. In addition, a connection should be established to connect words to *Wikipedia* concepts and consequently a word-concept similarity matrix (*WC*) is built in which each entry shows the cosine similarity between corresponding word and concept. The result of the linear combination of these two matrices are denoted by information item-concept similarity matrix (*IC*). Each entry of this matrix is calculated in the following manner.

$$IC_{i,j} = \sum_{j=1}^{w} IW_{i,j} \times WC_{j,i},$$

where w shows the total number of distinct words of *Wikipedia* concepts and I is the total number of information items. Figure 4 shows the linear combination of *IW* and *WC* matrices which generates the information-concept(*IC*) matrix.

Afterwards, text document profiles, which are represented by information items as features, can be connected to *Wikipedia* concepts by linearly combination of document-information items similarity matrix (*DI*) and information item-concept similarity matrix (*IC*). The relatedness document-concept similarity matrix (*DC*) built in this way clearly suffers from the lack of the semantic knowledge. Therefore, there is a need to integrate the semantic knowledge embedded in the semantic kernel into the document-concept similarity matrix and as a result, the semantic-based document-concept similarity matrix (*SDC*) is generated.

The semantic-based document-concept similarity matrices constructed by the aforementioned methods will then be used to construct a semantic social network of experts.

2.3 Constructing the semantic social network of experts

Once the process of enriching document representation is completed and the semantic-based expert profiles are constructed, the semantic similarities between all pairs of profiles have to be computed. In order to compute the semantic proximity of documents to each other, an operation widely used in social network analysis, namely folding, is applied. In social network analysis, it is possible to derive two one-mode networks from a two-mode network by applying the folding operation, which operates directly on the similarity matrix that corresponds to the social network. Assume the semantic-based document-concept similarity matrix (*SDC*), which in nature represents a two-mode network of documents and concepts, rows represent documents and columns

represent concepts. Multiplying this similarity matrix, by its transpose, will produce a new square similarity matrix that represents a one-mode social network; rows and columns both represent documents. The folding operation generates a symmetric matrix whose elements reflect the influence of the documents in the original two-mode network. In other words, each entry quantifies the semantic relationship between a pair of profiles. This recently generated similarity matrix is used to construct a one-mode social network of profiles. We will treat this network as a social network of experts because in our framework every expert is represented by a semantic-based profile.

2.4 Detecting expert communities and their representatives

According to the semantic-based similarity between individuals measured in the earlier stage, the social network of experts is constructed in which the relationships between semantic-based experts' profiles specify the weight of edges between corresponding nodes. Detecting communities is typically thought of as a group of nodes with more interaction amongst its members than between its members and the remainder of the network. Such clusters of nodes are often interpreted as organizational units in a social network. Different clustering algorithms can be applied for this purpose. In this study, the aim is to detect communities of experts such that there are stronger similarities between cluster members in terms of expertise, knowledge, and experience than between cluster members and other members of network. For this study, the k-means clustering algorithm is chosen to detect the communities of experts. Further, two measures, homogeneity and separateness, are used to evaluate different clustering solutions produced by k-means algorithm when different input values for k are used. The main goal in the clustering process is to optimize two main objectives: minimizing the number of clusters and maximizing cluster quality. The latter object combines two sub-objectives, namely maximizing within cluster similarity (homogeneity) and maximizing between clusters dissimilarity (separateness). All these objectives are conflicting. For instance, as the number of clusters decreases, more values are expected per cluster and hence the quality of the cluster is negatively affected. In order to achieve an acceptable compromise, k-means algorithm is applied with various numbers of clusters (k), and for each clustering solution, its homogeneity and separateness are measured. The final clustering solution is a trade off between maximizing homogeneity and minimizing separation.

In order to apply k-means algorithm to cluster the social network, each node (expert) is represented by a vector whose features are the semantic-based similarities to all other actors in the network. Given the collection of vectors that represents the semantic-based similarities among nodes, as the input data to k-means algorithm, the goal is to find a clustering solution that maximizes the semantic-based similarity within a cluster; and minimizes the semantic-based similarity among different clusters. Euclidean distance is used as the distance measure in k-means algorithm. The use of Euclidean distance in text document clustering and classification [21] is very common though other distance measures could be used. In a clustering solution with several clusters, the homogeneity and separation are calculated as the average homogeneity and separation of clusters as follows.

Assume $S(x, y)$ is the semantic similarity between two nodes x and y in the social network. Since the semantic-based similarity matrix is symmetric, then $S(x, y) = S(y, x)$ and consequently the edges between nodes in the network are undirected. A notation which is used for computing the separation is $S(i, j_n)$ that refers to the semantic similarity between nodes i and j, where j belongs to the cluster whose index (label) is n.

The homogeneity measurement of the kth cluster in a clustering solution is computed as follows:

$$Hom_k = \frac{1}{m} \sum S(i, j),$$

where m is the size of the kth cluster. As above equation denotes, the homogeneity of a cluster is the average similarity of its members. Further, the homogeneity of a clustering solution is defined as the average homogeneity of all clusters in that clustering solution. The formulation is shown below:

$$Homogeneity = \frac{1}{C} \sum_{\text{for each cluster } k} Hom_k,$$

where C is the number of clusters in a clustering solution.

Similarly, the separation of the kth cluster in a clustering solution is calculated as follows:

$$Sep_k = \arg \min_{\forall i \in k, \forall 1 \le n \le C} S(i, j_n),$$

where C is the number of clusters in a clustering solution. The separation of a cluster is defined as the minimum similarity that exists between an element of a cluster and an element of other remaining clusters. Similar to the homogeneity measurement for a clustering solution, the separation of a clustering solution is defined as follows:

$$Separation = \frac{1}{c} \sum_{\text{for each cluster } k} Sep_k.$$

Usually clustering solution can be summarized by introducing a representative member. A good representative member is the one whose average similarity to other members within the same cluster is the highest and whose average similarity to other non-mate elements is the least compared to the average similarities of the same cluster elements

to other non-mate elements. In this work, since each cluster represents an expert community in the experts' social network, the representative member of each cluster is in fact an expert who summarizes a particular community very well in terms of the knowledge, experience, and expertise that the community carries on.

Different methods can be used to find a cluster representative. For example, a very simple approach is to select the closest member to the center of a cluster as the representative of that cluster. However, we have decided to use a centrality measure that sounds more rational to find the member who perfectly summarizes the knowledge carried on by a community. The centrality measure that is used for this purpose is called *eigenvector centrality*, and is widely used in social network analysis. According to the *eigenvector centrality*, a node is central to the extent that its neighbors are central. In other words, in a clique the individual most connected to others within the cluster and other clusters, is the leader of the cluster. Members who are connected to many otherwise isolated individuals will have a much lower score in this measure than those that are connected to groups that have many connections themselves. In domain of this work, the *eigenvector centrality* follows that an expert well-connected to well-connected experts can carry on valuable types of knowledge and experience much more widely than one who only has connections to lesser important experts in a network or community. Experts with higher scores of eigenvector centrality could be critical when rapid communication is needed to find the right people whom we may ask a specific question and who will answer that question for us.

2.5 Building expert recommendation system

Recommendation systems are designed to allocate information items to individuals quickly and to achieve this goal, the similarity among thousands or even millions of data have to be computed. In this work, a hybrid approach that integrates the content-based characteristics into a social network-based collaborative filtering system, is proposed to recommend the most appropriate information items to communities of experts. Information items are specified in forms of user's questions for which a user is seeking the right experts. By using information retrieval approaches, the most similar information items are recommended to community members based on the similarity between the taste and preference of the community representative, and information items by means of cosine similarity. Other similarity measures that are commonly used in information retrieval methods could be similarly applied.

The social network component of the proposed system captures the social aspect of the experts' behaviors. Experts collaborate with their peers, whom they trust, on different knowledge areas to obtain new expertise and improve their own knowledge and experience. For a user who is looking for an expert for her/his information needs, a representative of an expert social community will be a better choice than an individual expert who has been recommended based on only the expert's profile. If more than one expert is required, more members of the same expert community are recommended. In the expert social network, experts are connected to each other based on the existing relationships among them, e.g. common expertise and experience. This social structure is best modeled in a social network and is often difficult to understand without being modeled in a network.

Thus, in a collaborative filtering recommendation system, to recommend information items to an expert, its community members are the most appropriate experts to be used. Indeed, users' preferences, interests, needs, experiments, etc. can be used to find similarity between experts in a social network and communities can be formed based on these similarities. Since text document profiles of communities' representatives contain some information about their experiences, abilities and skills, and other related information, finding experts for each information item according to the profiles information is feasible and reasonable. When an expert is detected by this way, an information item is assigned to the community that the given expert is its representative. Further, because all community members are semantically similar, all community members are experts in the same topic.

3 A case study

Conference mining and expert finding are highly investigated by knowledge discovery researchers for making useful recommendations to researchers. However, they mostly ignore semantics-based intrinsic structure of the words and relationships between conferences. As a new attempt to improve conference mining task Daud *et al.* in [22] consider semantics-based intrinsic structure of words and relationships presented in conferences (richer text semantics and relationships) by modeling from Group Level. They propose group topic modeling methods based on Latent Dirichlet Allocation (LDA). Their empirical evaluation shows that their proposed Group Level methods significantly outperformed Document Level methods for conference mining and expert finding problems. However, in this paper we propose a general framework that applies a semantic-based method to expert finding that could be used in conference mining as well.

In this section, we present a framework of an expert recommendation system for paper review process of the academic conferences, where the conference chair usually needs to assign appropriate experts (academic researchers)

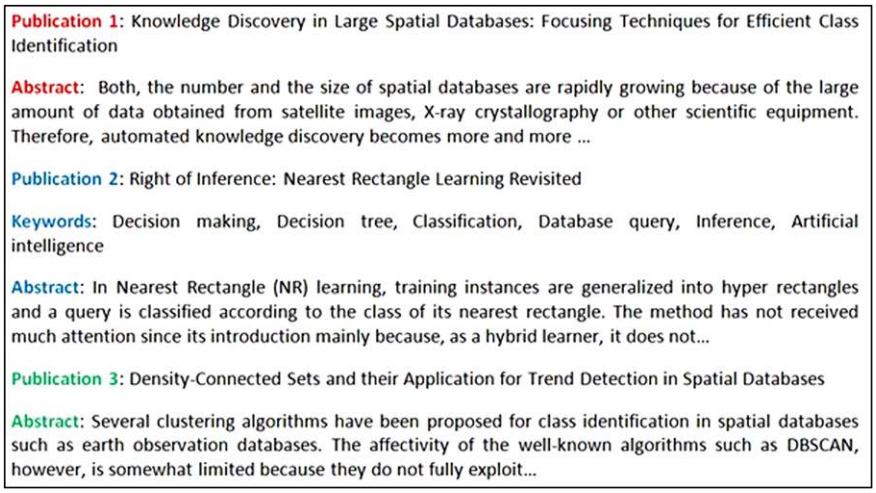

Fig. 5 An example of a researcher profile

to review the submissions. For this purpose, we have chosen 315 academic researchers in the field of computer science who are program committee members of the 16th ACM SIGKDD[1] conference. In addition, 62 keywords listed under the conference topics have been used as information items for which the program chair is seeking relevant researchers. This set of keywords covers a wide range of scientific topics in the field of knowledge discovery and data mining. The main goal of this case study is to assess the effectiveness of semantic social network-based expert recommender system in the task of assigning papers to members of the program committee for the review process of the conference.

As described in previous sections, in the first stage relevant information to researchers has been collected from online sources. For this purpose, a crawler has been programmed in C programming language that automatically collects experts' information and constructs their profiles. Experts' profiles contain their research interests, experiences and specialties that can be reflected through their publications. To gather this information automatically, the DBLP[2] bibliography has been used as a source by the system to automatically extract the list of publications corresponding to each researcher. For each publication, some related information such as the list of keywords and the abstract are retrieved from either digital libraries or Google scholar. Figure 5 demonstrates a part of an example profile.

Once the profiles are constructed, they will be enriched by adding the semantic knowledge. To extract the semantic knowledge from *Wikipedia* articles, we have automatically extracted *Wikipedia* pages and a tree structure of *Wikipedia* thesaurus has been constructed. In this structure, concept pages are located in the leaves of the tree while internal nodes indicate category pages. In addition, it should be mentioned that since the number of *Wikipedia* articles in the area of Computer Science is too large, not only do they need much space and computation to be processed, but also their relevance to this area is decreased by further proceeding in the tree structure. To cope with these problems, we cut the tree rooted in the *Computer Science* category at the level of 7. Figure 6 shows a small part of this tree constructed automatically from *Wikipedia* thesaurus.

For the clustering analysis and detecting communities, Weka 7.3[3] (which is a data mining and machine learning tool) was utilized. In addition, for detecting representative individuals in each community, ORA,[4] a dynamic meta-network assessment and analysis tool, was used. Once the social network of experts was constructed, communities were detected by using ORA features. The eigenvalue of all community members were also calculated and reported. For each community, the member with the highest eigenvalue was reported as the representative of that community. In our experiment, we have evaluated the results from two perspectives: first we report the k-means clustering results that lead to choose the best clustering solution, and second we discuss the effectiveness of the proposed semantic social network-based recommendation system.

3.1 Clustering analysis

Recall that two mapping schemes were applied to integrate the semantic knowledge into researchers' profiles and then the social networks of researchers were built based on the semantic relationships among them. Therefore, different clustering solutions, generated by k-means algorithm using different values of k in the range of 10 to 40, in both methods were generated and evaluated. The quality of solutions is evaluated based on homogeneity and separation measures.

[1] http://www.kdd.org/kdd2010.

[2] http://www.informatik.uni-trier.de/ley/db/.

[3] http://www.cs.waikato.ac.nz/ml/weka/.

[4] http://www.casos.cs.cmu.edu/projects/ora/index.html.

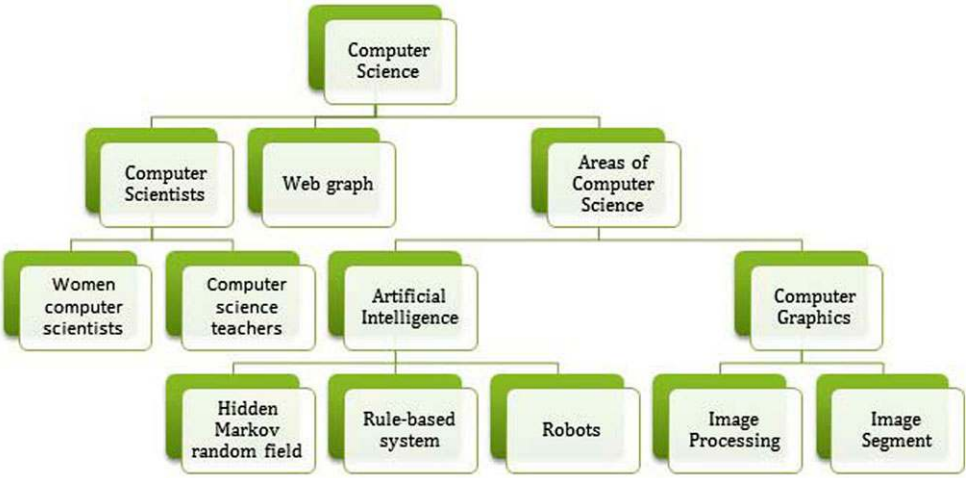

Fig. 6 A portion of developed *Wikipedia* tree structure

Because of the fact that there is not a clustering solution with the maximum homogeneity and minimum separation (conflicting objectives), a solution that trades of between maximizing the homogeneity and minimizing the separation is selected. The range of cluster numbers is chosen based on the number of researchers as well as the number of information items such that the average number of researchers in each cluster varies in a reasonable range. Further, the number of information items that is assigned to each cluster should vary in an acceptable range.

In Fig. 7, the number of clusters for different clustering solutions, while two different schemas where applied, are plotted on the horizontal axis against the values of homogeneity and separateness on the vertical axes. Figure 7(a) shows the clustering analysis on the social network that was built based on the concept match schema. The result of clustering analysis on the social network obtained by employing the information item relatedness schema is presented in Fig. 7(b). Moreover, Fig. 8 indicates the result of clustering analysis on the social network built without considering semantic-based relations between social entities.

The best clustering solution for the constructed researcher semantic social networks that are obtained by applying two enrichment schemes are the solutions with 12 clusters. However, according to the results of clustering analysis of social networks, without considering semantic-based relationships the best solution which maximizes the homogeneity and minimizes the separation is the one with 10 clusters. According to the clustering analysis of the two semantic-based social networks, both of them have the same patterns with respect to the homogeneity and the separation of different clustering solutions with just a small difference in the average value of both metrics. These results demonstrate that both concept match and information item relatedness schemes have the same efficiency in terms of the community detection. Therefore, only the identified communities of the social network constructed based on the concept

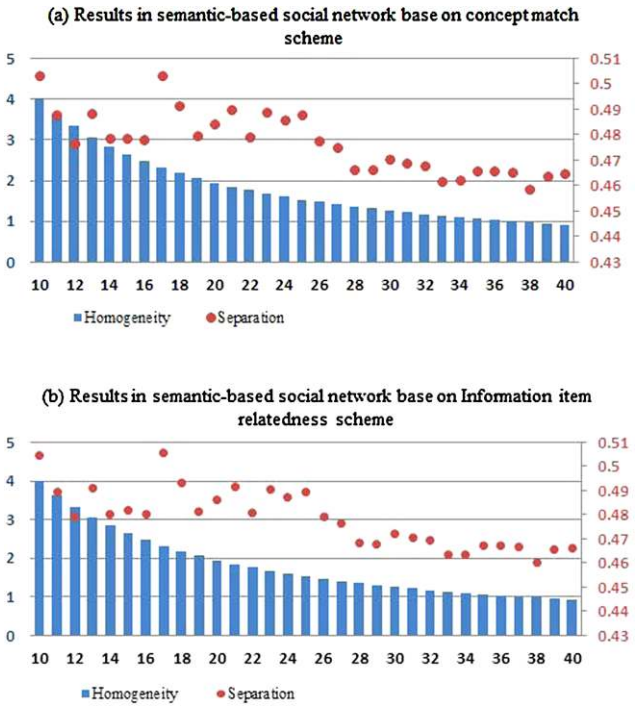

Fig. 7 Homogeneity and separation results of different clustering solutions in the semantic-based social network

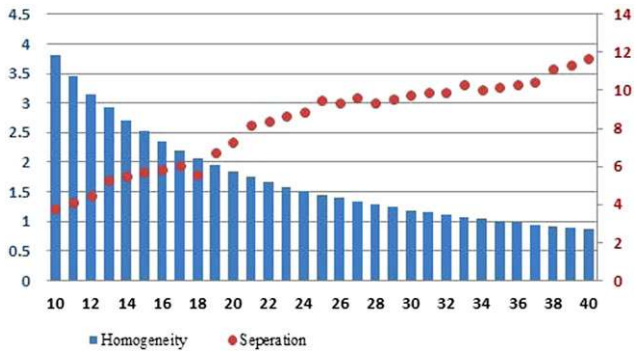

Fig. 8 Homogeneity and separation results of different clustering solutions in the social network without considering the semantic

match scheme are used in evaluating the performance accuracy of the system in the recommendation phase.

3.2 Recommendation phase

We have evaluated our recommender system from two perspectives. In the first one, we have conducted experiments to evaluate the impact of the social structure of experts' relationships captured in a social network as the social component of the recommendation system. While in the second one, the performance of recommender system that considers representative members of semantic-based social networks have been compared with the performance of the same recommender system in social network without considering semantic relationships.

3.2.1 The impact of the social structure of experts' relationships

We have conducted two sets of experiments in order to investigate the performance accuracy of the recommendation system with and without the social network component. When the social network is not used, recommendations are made based on the similarity between researchers profile and information items. In other words, the importance of individuals in their community is neglected. In the second set of experiments, the system utilizes the social network of experts through the process. Recommendations are made based on the similarity between communities representatives and information items. In this approach, the most appropriate experts are selected from a community whose representative has more expertise and knowledge about the requested item based on his/her profile information. In both approaches, if more than one expert is required, the system automatically suggests the second most relevant expert.

To measure the accuracy of our system, a set of 23 researchers were chosen to form a test set. Then, a questionnaire, for each researcher in the test set, was designed to discover preferred information items that a researcher is interested in. The questionnaires would contain 15 items from relevant to irrelevant. Questionnaires designed for different researchers were different from each other because we prepared them based on the recommended items by our system to researchers. Researchers were asked to score information items based on their relevancy to researchers' interests.

To evaluate the accuracy of the recommended items, a metric called precision at n or $P@n$ was used. This precision is defined as the fraction of retrieved instances that are relevant. The precision metric takes all retrieved items into account, but it can also be evaluated at a given cut-off rank, considering only the topmost results are returned by the system. We consider k-top most relevant items that the system recommends to researchers and investigate how many of them are actually relevant considering the researchers real interests given in the questionnaire. In using of $P@n$, we set n to 1, 3 and 5. For example, $P@1$ indicates the percentage of researchers who are recommended relevant information when only one information item is considered. The same method is applied to evaluate the accuracy of the prediction when information items are recommended to members of communities whose representatives have expertise and knowledge relevant to information items for which we are looking for experts.

Type of Measure	Result (%)	
	Recommendation without Social Network Component	Recommendation with Social Network Component
P@1	82.6	83.4
P@3	78.2	75.0
P@5	75.6	78.4

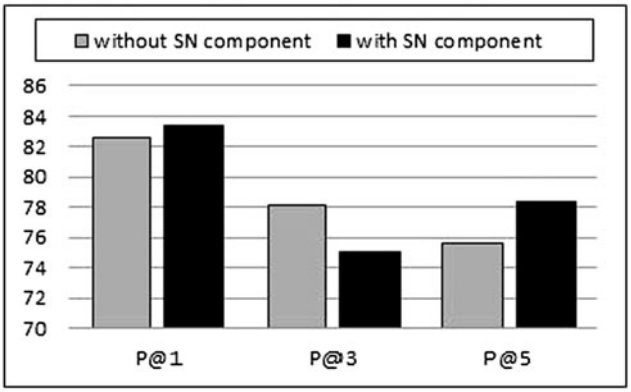

Fig. 9 The prediction accuracy of two recommendation models

Figure 9 demonstrates the precision values achieved when the above experiments were conducted. A $P@1$ value of 82.6 %, appeared in the first column of the table shown in Fig. 9 indicates that 19 out of 23 researchers in the test set, are recommended with relevant information item when only one information item is considered. In addition, the $P@1$ value of 83.4 %, shown in the second column of the table, means that the first recommended information item to 83.4 % of representatives are relevant. In other words, 10 out of 12 (12 is the number of communities achieved in the precious experiment) representatives are recommended with relevant item when only one information item is considered.

As described earlier, the proposed recommendation system helps users, who are looking for the most appropriate expert in a specific domain, choose representative member of each community to fulfill their information needs. In fact, a representative member can represent the knowledge and expertise of all members within the same community better than any other member in his/her community since his/her similarity to mate elements is the highest among all other mates. Thus, whenever a user searches for an expert who has relevant expertise to a specific information domain, a reliable choice is to trust to a community representative who

is recommended by the system. In addition, if more than one expert is needed, other community members can be recommended according to their importance indicated by eigenvector centrality measure; community members with higher eigenvector centrality are more reliable in that specific domain. Figure 9 summarizes the performance results shown in the result Table in Fig. 9. As can be seen, the performance of recommendations with the social network component slightly outperforms the performance of recommendation system without the social network component. Indeed, considering three types of precisions that were calculated in each experiment, only P@3 value for recommendation system without the social network component is higher than its corresponding value in the second experiment. Therefore, based on the comparison made between the results, the use of social network seems to be reasonable in that it improves the prediction accuracy of the recommendation model.

3.2.2 The performance of the recommender system

According to the above experiments, representative members of communities can best represent the knowledge and expertise of all members within a community. Thus, whenever an expert with relevant expertise to a specific information domain is needed, a reliable choice is to trust the representative of the community whose members are experts in that field. In addition, if more than one expert is needed, other community members can be recommended according to their importance indicated by the eigenvector centrality measure; community members with higher eigenvector centrality are more reliable in that specific domain. In this experiment, we have utilized this feature to evaluate the effectiveness of the semantic-based social network against the social network in which hidden relationships among social actors are ignored. Therefore, the representative members of all communities in both semantic-based social network (SSN) and social network (SN) without considering semantic relations are considered as our test set. To evaluate the accuracy of the assigned items $P@n$ was used. In using of $P@n$, we set n to 1, 3 and 5. Figure 10 demonstrates the results of this set of experiments. As can be seen, a significant improvement obtained by considering semantic relationships among individuals. In other words, this experiment shows the effectiveness of the semantic-based social network of researchers when there is a need to assign papers to relevant researchers for the review process.

It is worth mentioning that precision and recall (the fraction of relevant instances that are retrieved) are usually applied together to evaluate the performance of an information retrieval system, however in this work, only the precision was used. The main reason is that for calculating recall criterion, all relevant information items should be known. Since it was not possible to access all relevant items for each researcher, it was decided to restrict each individual to choose

Type of Measure	Result (%)	
	Recommendation in Semantic-based Social Network (SSN)	Recommendation in Social Network (SN)
P@1	83.4	60
P@3	75	53.3
P@5	78	50

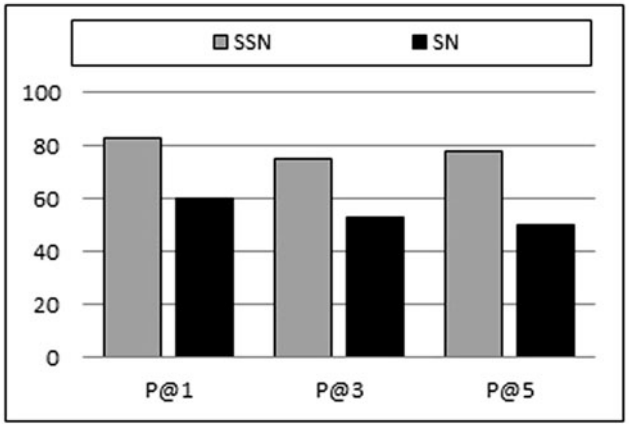

Fig. 10 The comparison between the performance of SSN and SN models

his/her relevant items from the list of 15 keywords. Indeed, relevant items for each researcher indicated in the questionnaire may not be all relevant keywords for each person. Hence, only the precision measure was employed to evaluate the performance of the system.

4 Conclusion

We have proposed a hybrid method for an expert recommendation system that integrates the characteristics of content-based recommendation algorithms into a social network-based collaborative filtering system. For this purpose, experts profiles are semantically enriched by the external knowledge extracted from *Wikipedia*. Hidden relationships can be discovered among experts' semantic profiles and accordingly a social network of experts can be constructed. In the resulted semantic-based social network, communities are detected by clustering analysis and representative members of communities can be detected by applying social network analysis measures. Recommendations are made based on the relevancy of an information item, for which a user is looking for experts, to the knowledge carried by representatives of groups. The proposed framework was tested in a typical application domain with a real data set. Experimental results show that not only does the presence of social components has a positive impact in increasing the accuracy of recommendation, but also discovering hidden relations among actors influence the accuracy of predictions in social communities.

References

1. Adamic LA (1999) The small world web. In: Proceedings of the third European conference on research and advanced technology for digital libraries, ECDL '99, London, UK, pp 443–452
2. Kalles D, Papagelis A, Zaroliagis C (2003) Algorithmic aspects of web intelligent systems. In: Zhong N, Liu J, Yao Y (eds) Web intelligence, vol 15. Springer, Berlin, pp 323–344
3. Herlocker J, Konstan J, Riedl J (2000) Explaining collaborative filtering recommendations. In: Proceedings of CSCW, pp 241–250
4. Hofmann T (2004) Latent semantic models for collaborative filtering. ACM Trans Inf Syst 22:89–115
5. Basu C, Hirsh H, Cohen WW (1998) Recommendation as classification: using social and Content-based information in recommendation. AAAI/IAAI, Menlo Park, pp 714–720
6. Garden M, Dudek G (2006) Mixed collaborative and Content-based filtering with User-contributed semantic features. AAAI, Menlo Park
7. Konstas I, Stathopoulos V, Jose JM (2009) On social networks and collaborative recommendation. In: Proceedings of the 32nd international ACM SIGIR conference on research and development in information retrieval, SIGIR '09, New York, USA, pp 195–202
8. Good N, Schafer JB, Konstan JA, Borchers A, Sarwar B, Herlocker J, Riedl J (1999) Combining collaborative filtering with personal agents for better recommendations. In: Proceedings of the sixteenth national conference on artificial intelligence and the eleventh innovative applications of artificial intelligence conference innovative applications of artificial intelligence, AAAI '99/IAAI '99, Menlo Park, CA, USA, pp 439–446
9. Renda ME, Straccia U (2002) A personalized collaborative digital library environment. In: Proceedings of the 5th international conference on Asian digital libraries: digital libraries: people, knowledge, and technology, ICADL '02, London, UK, pp 262–274
10. Perugini S, Goncalves MA, Fox EA (2004) A connection centric survey of recommender system research. J Intell Inf Syst 23(1)
11. Lueg C (1997) Social filtering and social reality. In: Proceedings of the 5th DELOS workshop on filtering and collaborative filtering. ERCIM Press, Biot, pp 10–12
12. Bedi P, Kaur H, Marwaha S (2007) Trust based recommender system for the semantic web. In: Proceedings of the 20th international joint conference on artificial intelligence, San Francisco, CA, USA, pp 2677–2682
13. Yoo I, Hu X, Song I-Y (2006) Integration of semantic-based bipartite graph representation and mutual refinement strategy for biomedical literature clustering. In: Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining, KDD '06, New York, USA, pp 791–796
14. Zhang X, Jing L, Hu X, Ng M, Zhou X (2007) A comparative study of ontology based term similarity measures on pubmed document clustering. In: Proceedings of the 12th international conference on database systems for advanced applications, DASFAA'07, Berlin, Heidelberg, pp 115–126
15. Baeza-Yates RA, Ribeiro-Neto B (1999) Modern information retrieval. Addison-Wesley Longman, Boston
16. Hotho A, Maedche A, Staab S (2001) Text clustering based on good aggregations. In: Proceedings of the 2001 IEEE international conference on data mining, ICDM '01, Washington, DC, USA, pp 607–608
17. Hotho A, Staab S, Stumme G (2003) WordNet improves text document clustering. In: Ding Y, van Rijsbergen K, Ounis I, Jose J (eds) Proceedings of the semantic web workshop of the 26th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR 2003), Toronto, Canada
18. Sinha RR, Swearingen K (2001) Comparing recommendations made by online systems and friends. In: DELOS workshop: personalisation and recommender systems in digital libraries
19. Wang P, Hu J, Zeng H-J, Chen L, Chen Z (2007) Improving text classification by using encyclopedia knowledge. In: Proceedings of the 2007 seventh IEEE international conference on data mining, Washington, DC, USA, pp 332–341
20. Eyharabide V, Amandi A (2012) Ontology-based user profile learning. Appl Intell 36(4):857–869
21. Lee LH, Wan CH, Rajkumar R, Isa D (2012) An enhanced support vector machine classification framework by using Euclidean distance function for text document categorization. Appl Intell 37(1):80–99
22. Daud A, Muhammad F (2012) Group topic modeling for academic knowledge discovery. Appl Intell 36(4):870–886