Wiki-Rec: A Semantic-Based Recommendation System Using Wikipedia as an Ontology

Ahmed Elgohary, Hussein Nomir, Ibrahim Sabek, Mohamed Samir, Moustafa Badawy, Noha A. Yousri
Computer and Systems Engineering Department
Faculty of Engineering, Alexandria University
Alexandria, Egypt
{algoharyalex, hussein.nomier, ibrahim.sabek, m.samir.galal, moustafa.badawym}@gmail.com, nyousri@alexeng.edu.eg

Abstract— **Nowadays, satisfying user needs has become the main challenge in a variety of web applications. Recommender systems play a major role in that direction. However, as most of the information is present in a textual form, recommender systems face the challenge of efficiently analyzing huge amounts of text. The usage of semantic-based analysis has gained much interest in recent years. The emergence of ontologies has yet facilitated semantic interpretation of text. However, relying on an ontology for performing the semantic analysis requires too much effort to construct and maintain the used ontologies. Besides, the currently known ontologies cover a small number of the world's concepts especially when a non-domain-specific concepts are needed.**

This paper proposes the use of Wikipedia as ontology to solve the problems of using traditional ontologies for the text analysis in text-based recommendation systems. A full system model that unifies semantic-based analysis with a collaborative via content recommendation system is presented.

Recommendation Systems; Semantic analysis; ontology-based analysis; Wikipedia ontology

I. INTRODUCTION

The web has become the dominating source of information in people's life. Nowadays, internet users are able to update web content in a variety of web applications. Although this enriches the amount of information, on the Web, an information overload problem arises. This problem occurs as a result of the enormous amount of information a user has to go through tediously in order to find those pieces relevant to his interest. Here comes the need for recommender systems to perform such tasks on behalf of the users. Recommender systems are those software systems that, relying on the prior knowledge of the users' interests, filter a large amount of information and provide the users only with the pieces relevant to them.

In order to provide the user with the relevant information, the system needs to learn about the user's interests to construct a user profile. A user profile can be constructed by explicitly asking the users about the topics they are interested in. This approach will not be suitable if there are too many topics. Another approach is to gradually and dynamically learn the user profile while the user is using the system.

Several models for recommendation exist. The most notable are content-based and collaborative approaches. Hybrid models try to overcome the limitations of each in order to generate better quality recommendations.

Since most of the information on the web is present in a textual form, recommendation systems have to deal with huge amounts of unstructured text. Efficient text mining techniques are, therefore, needed to understand documents in order to extract important information. Traditional term-based or lexical-based analysis cannot capture the underlying semantics when used on their own. That is why semantic-based analysis approaches have been introduced [1] [2] to overcome such a limitation. The use of ontologies have also aided in enhancing semantic-based analysis [3] where hierarchies of concepts are built to be able to capture conceptual relations between terms. Examples of commonly used ontologies are WordNet [4], OpenCyc [5], SNOMED[1], Gene Ontology[2].

Although traditional ontologies enhance the performance of semantic-based text analysis, yet they introduced other problems and challenges, for example, 1) Ontologies need to be maintained and updated periodically to cope with the dynamic nature of information, 2) Supporting different languages other than English is needed to meet the needs of different users, 3) Current ontologies cover a relatively small number of the world's concepts [6].

The recent work in [6], [7], [8], [9] proposed using Wikipedia as a knowledge base/ontology for the semantic analysis of text. Relying on Wikipedia instead of traditional ontologies solves the previously stated problems and achieves more accurate results as shown in [7].

Stemming from the efficiency of using Wikipedia as an ontology, the work presented exploits that fact for proposing a semantic-based text recommendation system model. The proposed semantic analysis modifies part of the work done in [7] using spreading activation and the concepts hierarchy for extracting concepts from the ontology. A hybrid collaborative via content recommender model is used for recommendation as it proved to be more promising compared to traditional models [3].

The paper is organized as follows: section II reviews some previous work on both semantic-based text analysis and recommendation system models. In section III, the proposed work is presented. In section IV, evaluation results are demonstrated, and finally section V concludes the paper.

[1] http://www.snomed.org
[2] http://www.geneontology.org/

II. BACKGROUND

This section reviews some previous work in the field of text recommendation systems and semantic-based text analysis then recent similar systems are reviewed.

A. Recommendation Systems

Recommendation techniques can be classified into Content-Based and Collaborative filtering techniques each has inherent limitations in its method of operation. Several Hybrid recommendation techniques were proposed to alleviate the limitations of each [10]. In this subsection, each of the techniques mentioned above is briefly demonstrated .

A content-based filtering system recommends items based on the correlation between the content of the items and the user's preferences. Content-based recommenders, firstly capture the target user's preferences, build his personal profile. Afterwards, the preferences stored in this profile are compared against the features of the items, recommending the most similar to the user's profile. Limitations of Content based systems have been identified in [10], [11], [12]. The most notable ones are restricted content analysis where recommendations are restricted only to textual content and Portfolio effect where the recommended topics get stuck only to those topics in the profile without recommending new stuff "out of the box" that the user might be interested in.

Collaborative Filtering provides recommendations based on the suggestions of users who have similar preferences. Since collaborative filtering is able to capture the particular preferences of a user, it has become one of the most popular methods in recommender systems. Collaborative filtering is classified as user-based and item-based [13]. Limitations of collaborative filtering as discussed in [10], [11], [12] are Sparse rating problem where users don't have enough common ratings, Grey sheep in which a user with unique tastes suffers from low quality of recommendations and cold start where a new item is never recommended until it's rated.

Hybrid recommender systems combine two or more recommendation techniques to gain better performance with fewer of the drawbacks of any individual one. Most commonly, collaborative filtering is combined with some other techniques in an attempt to avoid the cold-start problem. [10] surveyed various combination methods that have been employed.

Possible ways for the hybridization of recommendation techniques are by **weighting**, **switching** or **mixing** recommendations from more than one technique. Another way to achieve the content/collaborative merging is to treat collaborative information as simply additional feature data associated with each example and use content-based techniques over this augmented data set. This is referred to as **Feature Combination**. In **Feature Augmentation** on the other hand, one technique is employed to produce a rating or classification of an item and that information is then incorporated into the processing of the next recommendation technique. Another way that two recommendation techniques can be combined is by using the model generated by one as the input for another. This is referred to as a **Meta-Level** hybrid technique.

In hybrid recommendation systems, inter-user similarities significantly impact the collaborative recommendation quality. Several semantic similarity approaches were developed for that purpose. They are demonstrated in the next subsection.

B. Semantic Analysis

Semantic text analysis approaches need to be used to provide a conceptual understanding of the documents. Several approaches were proposed in the literature. In this subsection, some of the recent work on the ontology-based text analysis is reviewed. Afterwards, recent approaches that used Wikipedia for the text analysis purposes are described.

With the development of semantic retrieval, ontologies have become one of the hotspot approaches used in semantic annotation and in semantic similarity computation. The main role of ontologies in semantic analysis is to map terms to semantic concepts; ontology concepts are linked together to provide useful semantic relations can be used.

In 2005, [14], [15] implemented text-to-text semantic similarity methods using WordNet. In 2006, [16] presented a new approach for similarity computation between two documents by building a concepts graph for each document and then measuring the intersection between them. In 2008, [17] worked on blogs similarity and used the same idea of graph similarity after extracting the significant keywords. In 2009, [18] improved the ontology-based semantic similarity by adding the ontology instances to the general model of semantic similarity computation. For the semantic annotation purpose, [19] presented an approach for annotating document segments using a taxonomic ontology AGROVOC and tried to address the problems of extending ontology with the Arabic language.

Wikipedia started to be used as a knowledge base for the information retrieval purposes as it is the largest knowledge repository on the Web. Wikipedia is available in dozens of languages [7]. Also, it provides entries on a vast number of named entities and very specialized concepts [8]. For these reasons, In 2006, Strube and Ponzetto used it to measure the semantic relatedness of words [8].

Also in 2007, [7] proposed a model that utilizes Wikipedia for measuring the semantic relatedness between texts. In their model, each Wikipedia article is considered as a Concept. In their model (ESA), they built a weighted inverted index of Wikipedia where each word is represented by a weighted vector of the concepts it appeared in. For a text fragment, the vector of each word is retrieved from the index. All these vectors are merged together and the resulting vector is the interpretation vector of the given text fragment. Their evaluation results showed the effectiveness of the model. In 2009, the ESA model was refined by [9]. They observed the existence of noisy concepts in the interpretation vector of each document. They utilized the hyper-links between Wikipedia articles to cluster concepts (Wikipedia

articles) of each document in order to eliminate the noisy ones.

In 2008, [6] supported the idea of using Wikipedia concepts. But they extended the idea to formalize Wikipedia as an ontology. They utilized Wikipedia category graph structure where each category is categorized under some other categories. The aim of their work was to describe a document with a set of Wikipedia concepts. The concepts here turned to be Wikipedia categories. Similar to the ESA, they built an index of Wikipedia articles. To annotate a document, the document's text is submitted as a query to the index. The categories of top matching articles are used initially for the annotation. Then, they applied the spreading activation [20] technique through Wikipedia Category graph to extract a more generalization concepts.

In the rest of this part, we describe the overall model of recent two recommendation systems for textual data based on using ontologies for performing the semantic text analysis. In 2007, Degemmis [21] proposed a hybrid recommender model which computes similarities between users based on their content-based profiles. A distinctive feature of their work is the representation of semantic user profiles by integrating machine learning algorithms for text categorization with a word sense disambiguation strategy based exclusively on WordNet.

In 2008, I. Cantador [22] developed News@Hand, a multi-layer ontology-based hybrid recommendation model for recommending news articles. Exploiting IPTC news codes ontology, concept-based user profiles and item annotations are built. A personalized hybrid recommendation model based on the Collaboration-via-Content [23] is then established which allows the incorporation of semantic context.

III. System Model

In this section, the overall system model is described and then the details of the semantic analyzer component and the recommendation technique are given.

Figure (1) illustrates the main components of the system and their intercalation. Wikipedia annotator is used as the semantic analyzer of the system. All the text documents get annotated with Wikipedia concepts and stored into a repository. A profile needs to be maintained for each user representing his topics of interest. When a user rates documents, the semantic annotations of these documents are retrieved from the documents repository and used for refining his profile. The recommender component utilizes the user profile to find users with similar interests in order to run the recommendation algorithm. The result of the recommendation is a list of documents that are highly expected to be interesting to that user.

Following are more details of the system components. The proposed Wikipedia-Based Semantic analyzer is described in subsection A. Then the User Profile component and the Recommendation Technique are described in subsections B and C respectively.

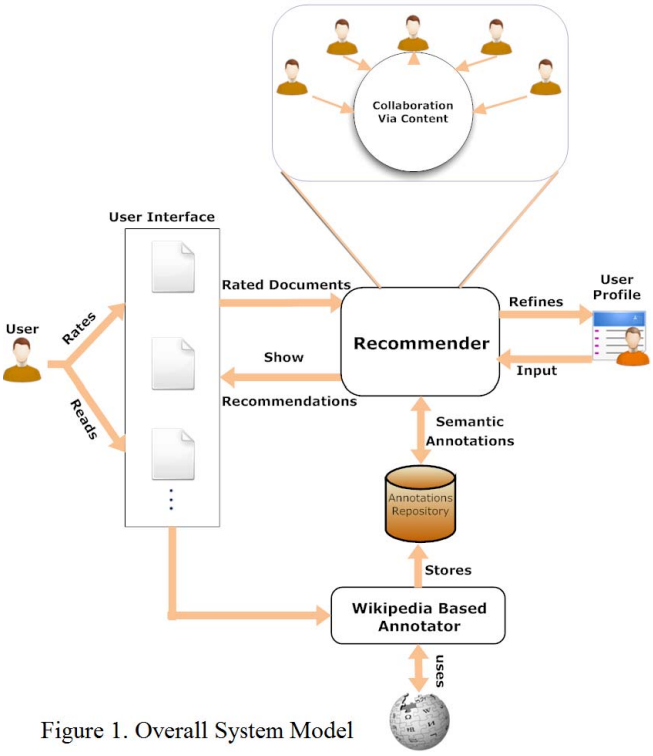

Figure 1. Overall System Model

A. Wikipedia-Based Semantic Analysis

Referring to the ESA model explained in the background section, documents can be annotated with Wikipedia articles as concepts. The benefits of this approach compared to using traditional ontologies are: 1) Wide coverage for many concepts which make Wikipedia a suitable general-domain ontology. 2) Wikipedia is continuously updated with articles about the recent topics (Concepts). Besides, the new relationships between the different concepts are defined implicitly by just stating the categories of each new article. For these two reasons, Wikipedia is currently the most up-to-date ontology which provides more accurate results specially when using it in an online text mining application. 3) Wikipedia is available in dozens of languages which makes it a multilingual ontology. 4) Wikipedia concepts are well described with a fairly large text fragment which eliminates any semantic ambiguity in the concepts.

Basically, the proposed semantic analysis model relies on the ESA model. A document gets annotated with a vector of weighted Wikipedia concepts (Articles). That weight represents the lexical matching between the document and the concept.

As pointed out in [9], the main problem of the ESA model is how to select the number of the concepts (N) each document should be annotated with. That is why they proposed annotating each documents with relatively large number of concepts (which adds some noisy concepts to the interpretation vector of the document) then, clustering those concepts using the hyper-link relationships between the resulting concepts (which are Wikipedia articles) to detect and eliminate of these noisy concepts. A different approach

is proposed in the work presented here. A document shall be annotated with a fairly small number of concepts (the top lexically matching ones) which minimizes including noisy concepts. Then, starting with those concepts, more relevant concepts are retrieved and added to the concept vector in a way that never introduces further noisy concepts. For that purpose, we applied the spreading activation technique.

The spreading activation model is one of the associative retrieval techniques. It is made up of a conceptually simple processing technique on a network data structure. It depends on the value dissemination in a network (e.g. a semantic network). The spreading activation on a network is performed according to some inference rules. Essential spreading activation functions are implemented to determine the activation flow and how its constraints are handled. The most important spreading activation factor is number of pulses, each pulse means the transition from the activated nodes to their parents. The processing technique is defined by a sequence of iterations (each iteration is called a pulse) that runs until halted by a termination condition or after a certain number of pulses [20].

To apply the spreading activation the following approach is proposed:

The top N articles used to annotate a document resulting from the ESA model represent the top N concepts to start with, where N is a fairly small number. The corresponding Wikipedia categories of the retrieved articles are extracted. Each concept is then weighted by adding up the scores of the articles belonging to it. The top N Wikipedia categories are used then as the initial set of nodes in Wikipedia category graph to start applying spreading activation. At the spreading activation termination, categories are ranked by their final activation score.

Spreading activation provides more accurate semantic representation for the document, because it begins dissemination from the top N concepts and decays their weights with each pulse which helps in reducing the noisy concepts. Thus, it is better than relying on the initial concepts as used in ESA model. Also, it enriches the annotation by using new related concepts.

To enhance the quality of semantic annotation, another approach is applied and we refer it in this work as concepts hierarchy-based approach. The Wikipedia categories graph structure is used in spreading activation to determine the parent-child relationships. On the other hand, it can be used to re-weight the concepts according to the hierarchical structure. The lower level concepts (the more specific ones) are assigned more weight than those in the upper level (more general) [16].

B. User Profile Learning

We adopted the Vector Space Model to represent User Profiles. Profiles are represented as vectors of weighted concepts obtained by the model described before. Each user profile contains two concept vectors, POSITIVE concept vector, which models the concepts that attract the user along with their weights (degree of attractiveness) and NEGATIVE concept vector, which models the concepts that the user dislikes along with their weights (degree of aversion). In order to learn a user's profile, these vectors are continuously tuned upon user feedback on the relevancy of recommended items, as shown next.

For learning user profile, Rocchio's algorithm for relevance feedback is used [23]. Rocchio algorithm is adopted from Information Retrieval research. Originally, relevance feedback is used to improve search results using user's feedback on the relevancy of the retrieved documents. Each user profiles is modeled as a classifier for documents that has two classes (POSITIVE and NEGATIVE). A user profile is learned from concept vectors of the user's rated documents (training examples). Learning is achieved by combining document vectors into a prototype vector $\mathbf{c}_j$ for each class Cj. First, both the normalized document vectors of the positive examples for a class as well as those of the negative examples for a class are summed up. The prototype vector is then calculated as a weighted difference as shown equation (1).

$$\vec{C_J} = \alpha \frac{1}{|C_j|} \sum_{\vec{d} \in \vec{c_j}} \frac{\vec{d}}{\| \vec{d} \|} - \beta \frac{1}{|D - C_j|} \sum_{\vec{d} \in D - C_j} \frac{\vec{d}}{\| \vec{d} \|} \quad (1)$$

α and β are parameters that adjust the relative impact of positive and negative training examples. Cj is the set of training documents assigned to class j and ||d|| denotes the Euclidian length of a vector d.

C. *Recommendation Model*

As user profile model is established, a recommendation model is to be defined. As shown in the background, collaborative recommendation models look for similarities between users in order to generate recommendations. Typically, the pattern of ratings of individual users is used to determine similarities between users. Such a correlation is most meaningful when there are many items rated in common among users. In some real situations, we'd expect there to be a smaller number of item ratings in common between users. For example, for someone visiting a city for the first time, there may not be any users with a rating in common. In such situations, collaborative methods might be expected to fail.

MJ Pazzani [22] proposed a Collaborative-via-Content hybrid recommendation technique taking advantage of both content based and hybrid recommendation. From the hybrid recommendation categories discussed in [10], Collaborative-via-Content is classified as a Meta-level hybrid recommendation. It exploits the content-based model of the user profiles to look for similarities between users.

These similarities are used as the weighting factor of the collaborative advices provided by neighbor users in a collaborative recommendation framework as shown in equation 2.

$$g(u_m, i_n) = \sum_{u_j \in \hat{U}_m} sim(u_m, u_j) \times r_{j,n} \quad (2)$$

In Equation (2) $g(u_m, i_n)$ is the utility of item i_n to user u_m. $r_{j,n}$ is the rating provided by user u_j to item i_n. $Sim(u_m, u_j)$ is the cosine similarity between user profiles u_m, u_j. And that's where Collaboration-via-Content enhancements take place.

Recall that the user's content-based profile contains weights for the concepts that indicate that a user will like/dislike an object. When computing Cosine similarity between two profiles, any concept in one profile but not in the other is treated as having a weight of 0 in the other profile. As in collaborative filtering, the prediction made for an item is determined by a weighted average of all users' predictions for that item, using the similarity between profiles as the weight. This is demonstrated in Figure (2).

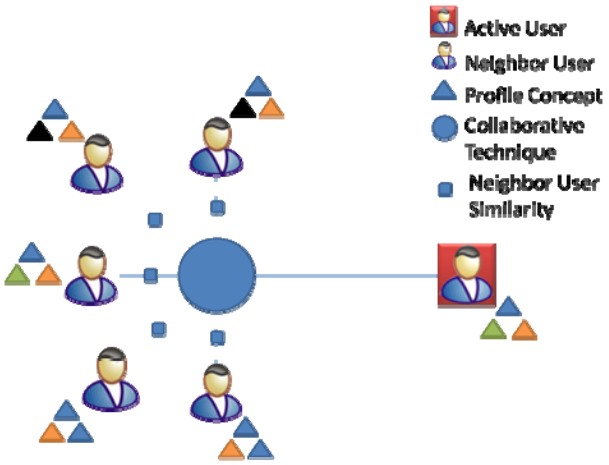

Figure 2. Collaborative via Content Recommendation

IV. Evaluation

The evaluation process is divided into two parts; evaluating the semantic annotation of the documents and evaluating the recommendation technique.

To evaluate the semantic annotation part, a benchmark of 50 documents with "human-judged" inter-document similarity matrix is used [24]. That benchmark was also used in the evaluation of the ESA model. The accuracy of the semantic annotation is reflected in the semantic similarity between documents. We used the cosine similarity between the interpretation vectors of the documents to evaluate their similarities. The correlation between the "human-judged" similarity matrix and the resulting similarity matrix of the proposed model indicates the accuracy of the model.

We selected to fix the number of articles N at 200. This is done to avoid the drawback of the ESA model by

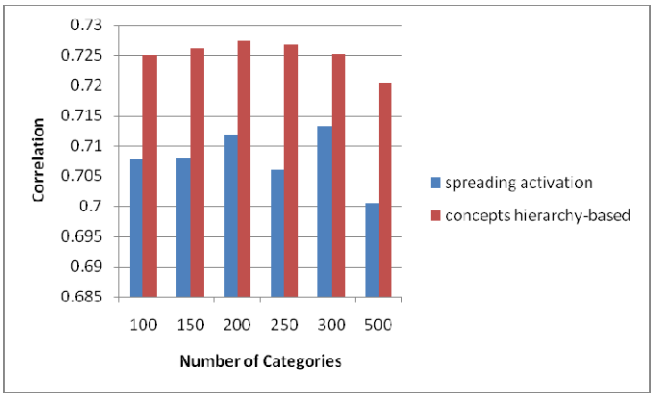

Figure 3. Number of Categories vs. Correlation

eliminating the effect of the used number of articles (N) on the quality of the annotation.

From figure 3 it can be observed that, with different values for N (categories), the proposed model achieves higher correlation than the maximum achieved by the ESA model (0.72) .

As proposed the documents get annotated with categories so, we need also to eliminate the effect of the number of the used categories. In figure 3, we have measured the correlation for different numbers of categories.

From the results in figure 3, the following can be observed: 1) Weighting the concepts based on their level in the category graph, as proposed, achieves better results than the spreading activation without re-weighting. 2) The correlation achieved by the proposed model (concepts hierarchy-based) does not change significantly with the number of categories. This shows that the concept-hierarchy based approach is more robust compared to previous approaches.

B. Overall System Evaluation

In this subsection two experiments were performed to make sure that the system is working correctly and to show the effect of system analysis part in solving recommendation problems.

We have collected 70 blog posts under different categories; technology, politics, life style and sports. 20 users were involved in the experiments.

Table I. Recommender system accuracy results

Accuracy measure	Normal case	Cold start case
RMSE	1.93028	1.865
Precision	0.8181	0.842
Recall	0.9	0.842
F-measure	0.85714	0.842

Table I shows the accuracy results for number of categories = 200 and number of pulses = 2.

In this experiment two cases are compared. These are the normal and the cold start cases. In the normal case, each user has an average of 10 rated documents. While in the cold start case, one user has an average of 2 ratings only for all documents and the other users have an average of 10 ratings. The values of F-measure, precision and recall are given in table I to indicate the accuracy of the results obtained.

The table shows that the accuracy results under cold-start conditions are sufficiently close to normal conditions. This signifies the benefit of enriching user profile concepts using our enhanced semantic annotation model.

V. Conclusion and Future Work

We have presented an enhanced semantic annotation model which achieves more accurate analysis than previous models. This model is integrated in a hybrid text-based recommendation system. Applying the enhanced semantic analysis model on a benchmark data set was shown to alleviate some of the recommendation systems' limitations. The recommendation accuracy is also given, and some previous limitations are solved.

It is important to evaluate the overall system using larger datasets and comparing the results to different text analysis techniques, as will be considered as a future work.

References

[1] S. Deerwester, S. Dumais, G. Furnas,T. Landauer, and R. Harshman, "Indexing by latent semantic analysis," JASIS, 41(6):391–407, 1990.

[2] R. Baeza-Yates and B. Ribeiro-Neto, "Modern Information Retrieval". Addison Wesley, New York, NY, 1999.

[3] I. Cantador, A. Bellogin, P. Castells, "A multi-layer ontology based hybrid recommendation system," 2008

[4] C. Fellbaum, editor, "WordNet: An Electronic Lexical Database," MIT Press, 1998.

[5] C. Matuszek, J. Cabral, M. Witbrock, and J. DeOliveira, "An introduction to the syntax and content of Cyc. In AAAI Spring Symposium," 2006.

[6] Z. S. Syed, T. Finin and A. Joshi," Wikipedia as an Ontology for Describing Documents". Association for the Advancement of Artificial Intelligence, 2007.

[7] E. Gabrilovich and S. Markovitch," Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis". Proc. of the 20th International Joint Conference on Artificial Intelligence (IJCAI'07), 6-12, 2007.

[8] M. Strube and S. Paolo Ponzetto," WikiRelate! Computing Semantic Relatedness Using Wikipedia,"In AAAI'06, Boston, MA, 2006.

[9] A. Prato and M. Ronchetti, "Using Wikipedia as a reference for extracting semantic information from a text".Third International Conference on Advances in Semantic Processing, 2009

[10] R. Burke, "Hybrid recommender systems: Survey and Experiments," User Modeling and User-Adapted Interaction, pp. 331–370. Springer, 2002.

[11] M. Balabanovic and Y. Shoham, Fab:" Content-Based Collaborative Recommendation.," Communications of the ACM archive, 1997.

[12] G. Adomavicius and A. Tuzhilin, "Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions," IEEE Transactions on Knowledge and Data Engineering, 2005.

[13] Z. Huang, D. Zeng and H. Chen," A Comparison of Collaborative-Filtering Recommendation Algorithms for E-commerce," IEEE Intelligent Systems, v.22 n.5, p.68-78, 2007

[14] G. Varelas, E. Voutsakis and P. Raftopoulou, "Semantic Similarity Methods in WordNet and their Application to Information Retrieval on the Web," WIDM'05, 2005.

[15] C. Corley and R. Mihalcea, "Measuring the Semantic Similarity of Texts,"Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment, 2005.

[16] Pei-Ling Hsu, Po-Ching Liu and Yi-Shin Chen, "Using Ontology to Map Categories in Blog," proceeding of the international workshop on integrating AI and Data Mining (AIDM'06), 2006

[17] S. Yan, Z. Lu and J. Gu, "Research on Blog Similarity based on Ontology ," Second International Conference on Future Generation Communication and Networking Symposia, 2008.

[18] X. Wang, J. Zhou," An Improvement on the Model of Ontology-Based Semantic Similarity Computation," first international Workshop on Database Technology and Applications, 2009.

[19] S. El-Beltagy, Maryam Hazman and Ahmed Rafea, "Ontology Based Annotation of Text Segments," SAC'07, 2007.

[20] F. Crestani," Application of Spreading Activation Techniques in Information Retrieval". Artificial Intelligence Review 11: 453–482, 1997.

[21] M. Degemmis , P. Lops , G. Semeraro," A content-collaborative recommender that exploits WordNet-based user profiles for neighborhood formation," User Modeling and User-Adapted Interaction, pp. 217-255 Springer ,2007

[22] MJ. Pazzani, "A Framework for Collaborative, Content-Based, and Demographic Filtering," Artificial Intelligence Rev., pp. 393-408, Springer, 1999.

[23] J. Rocchio, "Relevance feedback in information retrieval. The SMART Retrieval System," Experiments in Automatic Document Processing, pp. 313-323. Prentice-Hall Inc., 1971

[24] M. Lee, B. Pincombe, and M. Welsh, "An emprical evaluation of models of text document similarity," CogSci2005, pages 1254-1295, 2005.