A Hybrid Model to Improve Filtering Systems

Kharroubi Sahraoui[1(✉)], Dahmani Youcef[2], and Nouali Omar[3]

[1] National High School of Computer Science E.S.I, and Ibn Khaldoun University Tiaret, Tiaret, Algeria s_kharoubi@esi.dz
[2] Department of Computer Science, Ibn Khaldoun University, Tiaret, Algeria
[3]
dahmani_y@yahoo.fr
[4] Basic Software Laboratory, C.E.R.I.S.T, Ben Aknoun, Algeria
o_nouali@cerist.dz

Abstract. There is a continuous information overload on the Web. The problem treated is how to have relevant information (documents, products, services etc.) at time and without difficulty. Filtering system also called recommender systems have widely used to recommend relevant resources to users by similarity process such as Amazon, MovieLens, Cdnow etc. The trend is to improve the information filtering approaches to better answer the users expectations. In this work, we model a collaborative filtering system by using Friend Of A Friend (FOAF) formalism to represent the users and the Dublin Core (DC) vocabulary to represent the resources "items". In addition, to ensure the interoperability and openness of this model, we adopt the Resource Description Framework (RDF) syntax to describe the various modules of the system. A hybrid function is introduced for the calculation of prediction. Empirical tests on various real data sets (Book-Crossing, FoafPub) showed satisfactory performances in terms of relevance and precision.

Keywords: Recommender systems · Resource description framework · Dublin core · FOAF · Semantic

1 Introduction

The multiplicity of the services offered via the Web excites the Net surfers to expose and communicate an enormous traffic of data of various formats. The gigantic mass of existing information and the speed of its instantaneous production triggers the problem of informational overload. This phenomenon known under the name big data imposes multiple difficulties such as management, storage, the control and the security of circulated data. On the other hand, the access to relevant information in time is a major occupation of the developers and users, in spite of his availability it is lost in the mass. The performances of the existing tools degrade when we handle large volume of data, more precisely the search engines are involved by this phenomenon in terms of recall and precision as well as the process of the indexing. Our work is more particularly listed under filtering information tab, specifically custom filtering in order to submit

A. Amine et al. (Eds.): CIIA 2015, IFIP AICT 456, pp. 303–314, 2015.
DOI: 10.1007/978-3-319-19578-0_25

the useful information to the users. Many commercial and educational sites are based on the filtering algorithms to recommend their products such as the Amazon, Movielens, Netflix, EducationWorld etc [5]. Filtering systems (FS), known as "recommender systems", have become essential with the increasing variety of web resources such as news, games, videos, documents or others [10]. The majority of the recent FS explores semantic information and share the metadata of the resources in order to improve the relevance factor[8]. Additionally, another type of these systems is based on ontology for conceptualizing and valorising the application domain, which makes it possible to increase their performances [1]. However, FS suffer from some common weaknesses, such as cold start, sparsity and scalability. In our study, we adopted the RDF model to represent all elements of the system with an open and interoperable manner. With the formalism Friend Of A Friend (FOAF), we weighted the attributes of the user profiles in order to gather them by degree of similarity. In addition, the items of system are represented by the Dublin Core vocabulary (DC) in RDF model to describe the web resources formally. These two formalisms that are recommended by W3C ensure interoperability and easy integration of the data. This approach allowed us to avoid focusing the approaches on a specific and closed field, and treats all kinds of resource using the URI and namespace clauses. The rest of the paper is organized as follows, we will briefly review the various forms of FS in section 2. The section 3 presents the details of our proposal. The results of experiments followed by discussions were exposed in section 4. In the end, we conclude our work with a conclusion and perspective.

2 State of the Art

The number of Internet users has now reached 38.8% of the world population in 2013 against 0.4% in 1995 according to statistics provided by ITU *(http://www.itu.int/en/ITU-D/Statistics/Pages/stat/default.aspx)*. On the other hand, resources called commonly items occur at an incredible speed either by users or companies. Current tools are not consistent with this huge volume of data in order to analyze, control or have relevant information at time. The birth of FS is used to manage information overload by filtering [3,8]. Items can be extremely varied DVDs, books, images, web pages, restaurants ... etc. These systems are now increasingly present on the web and certainly will become essential in the future with the continuous increase of data [12]. According to how to estimate the relevance, researchers classify recommendation algorithms into three main approaches: content-based, collaborative and hybrid [4]. In the first approach, the system will support the content of the thematic items "documents" to compare them with a user profile, itself consists of topics explaining his interests, that is to say, the system compares the document themes with those of the profile and decides if the document is recommended or rejected according to the threshold of satisfaction function [17]. In the second approach, also known as social, the system uses the ratings of certain items or users and in order to recommend them to other users through the application of similarity process and without it being necessary

to analyze the content of items [2], in this approach, there are two main techniques which builds on memory-based algorithms, that operates a portion or all of the ratings to generate a new prediction [12] and which is founded on the model-based algorithms to create a descriptive model of the user so, estimate the prediction. The collaborative approaches are widely adopted in recommender systems such as Tapestry [4] GroupeLens [15], Amazon, Netflix ... etc. The hybrid methods operate to attenuate the insufficiencies of each of the two previous approaches by combining them in various manners. Recently, a new generation of FS boosted by semantic web formalisms or adaptable to contexts that uses a taxonomies or ontologies [13]. Commonly, these systems have shortcomings that prevent the recommendation process and degrade their performances, like the effect of the funnel where the user does not profited from the innovation and diversity of the items recommended in content-based filtering; the scalability where the system handles a large number of users and items online what makes difficult to predict in time; the sparsity problem, where there's a lack of sufficient evaluations to estimate the prediction well as the problem of the cold start to a user and/or item lately integrated into the system [11]. In this paper, we will extend the filtering systems in an open and interoperable specification, each component of the system is formalized by an appropriate RDF vocabulary. The following section explains the basic concepts of this specification.

3 Proposed Approach

Our study focuses on reducing the sparsity problem through the similarity of items via the values of DC properties, as well as the similarity of users through the values of FOAF properties. The values of properties are heterogeneous type nominal, ordinal, qualitative, etc ., so we have defined several functions of encoding and normalization to convert these properties in a numeric scale. i.e. quantitative values in the range [0-1].

3.1 RDF Specification

Resource Description Framework RDF (http://www.w3.org/TR/2004/REC-rdf-syntax-grammar-20040210/) is a data model for the description of various types of resources (person, web page, movie, service, book etc.). It treats the data and its properties and the relationship between them, in other words it is a formal specification by meta-data, originally designed by W3C, whose purpose is to allow a com-munity of users to share the same meta-data for shared resources. However, an RDF document is a set of triplet ¡subject, predicate, object¿ where the subject is the resource to be described, the predicate is the property of this resource and the object it is the value of this property or another resource. One of the great advantages of RDF is its extensibility through the use of RDF schemas that can be integrated and not mutually exclusive with the use of namespace and URI (Uniform Resource Identifier) concepts [7]. It is always possible to present a RDF document by a labelled directed graph. For example, "the book Semantic

Web for the Working Ontologist written by Dean Allemang on July 5, 2011", in RDF/XML Syntax: < ?xml version="1.0"? >

```
<rdf:RDF xmlns:ss="http://workingontologist.org/"
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#">
<rdf:Description rdf:about="http://www.amazon.fr/
Semantic-Web-Working-Ontologist-Effective/dp/0123859654/">
<ss:written_by rdf:resource="http://www.cs.bu.edu/fac/
allemang/"/> </rdf:Description>
<rdf:Description rdf:about="http://www.amazon.fr/
Semantic-Web-Working-Ontologist-Effective/dp/0123859654/">
<ss:hasTitle>SemanticWeb for the WorkingOntologist</ss:hasTitle>
</rdf:Description>
<rdf:Description rdf:about="http://www.amazon.fr/
Semantic-Web-Working-Ontologist-Effective/dp/0123859654/">
<ss:hasDate >July 5, 2011 </ss:hasDate >
</rdf:Description>
</rdf:RDF>
```

Our solution (figure1) based on a modelling in RDF through FOAF and Dublin core standards,describing the set of the users and items.

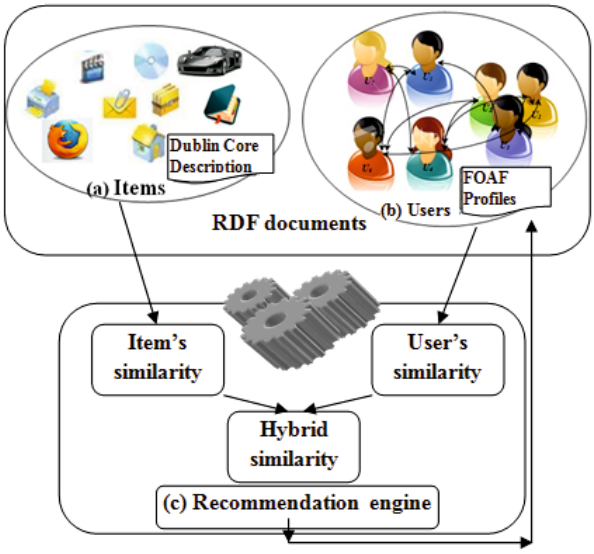

Fig. 1. Overall scheme of the proposal

Thus, in order to keep the collaborative filtering approach we took into account the feedback of the users in the process of computing similarity, moreover we used a hybrid function to define the prediction value. To facilitate the integrity and interoperability, all the documents are represented in RDF/XML notation.

3.2 Item's Representation

A social FS consists of resources items, the users profiles and the histories which memorizes the interactions of the users (ratings) about items recommended. We exploited the meta-data of the Dublin core vocabulary as being a standardization description of items, the attributes values of the vocabulary allowed us to calculate the degree of similarity between items and group them into communities.

Dublin Core vocabulary. Dublin Core DC (*http://dublincore.org*) is a set of simple and effective elements to describe a wide variety of web resources, the standard version of this format includes 15 elements of which semantics has been established by an international consensus coming from various disciplines recommended by W3C. These elements are gathered in three categories those which describe the contents (*Cover, Description, Type, Relation, Source, Subject*) and those which describe the individual properties (*Collaborator, Creator, Editor, Rights*) and others for instantiations (*Date, Format, Identifier, Language*), the current version is known as 1.1, validated in 2007 and revised in 2012 by DCMI (Dublin Core Metadata Initiative, (*http://dublincore.org/documents/dces/*).

Description of items. The core of FS is to form properly the communities, according to well determined criteria, in our research we propose to form the items by taking of account the qualifier DC meta-data QDCMI. We define the set of items as follows:
$I = \{(i_1^1, i_1^2, ...i_1^p), (i_2^1, i_2^2, ...i_2^p), ...(i_m^1, i_m^2, ...i_m^p)\}$ where i_k^j represent the j^{th} property for item k which is identified by its URI and is specified by its qualifiers. We group items by degree of similarity, so I_1 the set of properties assigned to the i_k item and I_2 is the set of properties assigned to the i_l item, then the degree of similarity between i_k and i_l by cosine measurement is given by:

$$sim(i_k, i_l) = \frac{\sum_{j \in I_1 \cap I_2} i_k^j . i_l^j}{\sqrt{\sum_{j \in I_1} \left(i_k^j\right)^2} \cdot \sqrt{\sum_{j \in I_2} \left(i_l^j\right)^2}} \quad (1)$$

This similarity value, allows to group items based on their associated DC properties.

3.3 User's Representation

The objective of FS is to deliver the relevant items to the user, because the formation of the communities depends on the attributes values defined in the user profile. Among the most common current practices we adopted the FOAF vocabulary to represent our profiles.

FOAF vocabulary. FOAF (Friend Of A Friend), is an RDF vocabulary for describing in structured manner a person and his relationships (*http://www.foaf-project.org*). However, it can be used to search for individuals and communities: CV, social networks and management of the online communities, online identification and management of participation in projects etc. A file FOAF can contain various information (*name, family_name, dateOfBirth, gender, mbox, Home Page, weblog, interest, accountName, Knows,etc.*). The major advantage of this representation is the ability to integrate other vocabularies as *DC* (describing a resource), *BIO* (to reveal biographical information), *MeNow* (describing the current status of a person), relationship (to see the type of relation maintained with a person).

Modelling of the user profile. Following the very high number of the users in interaction, it is very important to well form the community as a building block in the FS and assuming one for all and all for one. In order to formulate knowledge, we organized the user profile with categories of FOAF properties and each category c_i associated with a weight w_i, thus we defined the FOAF similarity according to n categories registered in profile by:

$$sim_f = w_1 sim_{c1} + w_2 sim_{c2} + ...w_n sim_{cn} \quad \begin{cases} \sum_i w_i = 1 \\ 0 \leq w_i \leq 1 \end{cases} \tag{2}$$

For our study, we retained three principal categories according to the evolution on the time axis, the first category $c1$, as no evolutionary, includes the non-changeable foaf properties such as: *name, birth_day, gender, mbox,etc.*, the second category in the medium and long term $c2$ contains the foaf changeable properties such as: *account, focus, homepage, phone, skypeID, status, depiction* etc., and the third category $c3$ is defined as category of the preferences includes the foaf properties which interest and preferred by the user like *know, interest, logo, topic_interest, weblog, workplace, based_near, membership* etc. so each class is properly associated with a weight w_i . However, the similarity by foaf properties based on the three categories mentioned above becomes:

$$sim_f = w_1 sim_{c1} + w_2 sim_{c2} + w_3 sim_{c3} \tag{3}$$

Let $u_{f1} = f_1^1, f_1^2, ..., f_1^k$ and $u_{f2} = f_2^1, f_2^2, ..., f_2^k$ the set of the foaf properties of the user u_{f1} and u_{f2} user in a given c_i class, then the value of similarity between these two users by the measurement of cosine that given by the following relation:

$$sim_{ci}(u_{f1}, u_{f2}) = \frac{\sum_{j=1}^{k} f_1^j . f_2^j}{\sqrt{\sum_{j=1}^{k} (f_1^j)^2} . \sqrt{\sum_{j=1}^{k} (f_2^j)^2}} \tag{4}$$

If the value of similarity of two users is close to 1 meant that they belong to the same community.

3.4 Recommendation Engine

The purpose of a FS is to distribute relevant items to users, and avoid a hard task of search in a "big data", the current recommender systems lean on the hybrid approaches which our research is belongs. We have proposed a hybrid similarity based on three types of relationships.

Hybrid similarity. In order to adjust the values of predictions, we conceived a formula to calculate the hybrid similarity, definite as follows:

$$sim_h = \alpha sim_{dc} + \beta sim_f + \gamma sim_r \tag{5}$$

The parameters $\alpha, \beta, \gamma \in [0, 1]$ adjusted by the system administrator according to the efficiency and availability of data.

- sim_{dc}, similarity that using the Dublin Core vocabulary for describing items. By the use of the URI, while identifying item and by exploiting its own metadata allowing reduce the sparsity problem.
- sim_f, similarity which depends on the representation of the profiles by the means of FOAF formalism, in favour of the variety of the fields and the availability of the data in profile, thus, we can overcome the problem of cold start of a new user and to still better forming the communities.
- sim_r, concretize the principal of collaboration through the ratings histories of users to estimate the prediction and to establish the recommendation, so consider their implicit tastes that are often difficult to value by attributes depicted in profile.

Prediction function. Before proceeding to the recommendation task, the system calculates the predicted value of an i item for the active user a, for that, we must select the S most similar items to i_l, then we retain the rating feedback of this user for these S similar items according to the relation:

$$p_{a,l} = \frac{\sum_{m=1}^{s} r_{a,m}.sim_h(i_l, i_m)}{\sum_{m=1}^{s} sim(i_l, i_m)} \tag{6}$$

Where $r_(a, m)$: is the rating value of the current user on the m^{th} similar item. S: size of the most similar items.

Recommendation process . The recommendation process is purely automatic and directly related to the prediction value, so a given item is deemed relevant and deserves to be sent to the user if and only if its predictive value is greater than a given threshold.

$$R_{a,l} = \begin{cases} i_l & recommended\ to\ u_a \quad if\ p_{a,l} \geq \rho \\ i_l & not\ recommended\ to\ u_a \quad otherwise \end{cases} \tag{7}$$

4 Experimentation

This section is devoted to the experimental results of our hybrid solution on real data sets. For evaluation and comparison, we implemented item-CF (item based collaborative filtering) approach widely referenced in Collaborative filtering search [6].

4.1 Datasets

For experimental tests we exploited two sets of data:

- *Book − Crossing* dataset (http://www.informatik.unifreiburg.de/ cziegler/ BX/), a free download dataset for ends of research collected by Cai-Nicolas Zeigler in 2004 from the famous Amazone.com site. The dataset constitutes of 278858 users producing 1149780 votes for 271379 books.
- $foafPub$ dataset (http://ebiquity.umbc.edu/resource/), is a set of data extracted from FOAF files collected during the year 2004, includes 7118 FOAF documents collected from 2044 sites and distributed under the Creative Commons license (v2.0). This set has allowed us to import FOAF properties by SPARQL queries to determine the similarity sim_f.

Our empirical tests require the deployment of a parser to extract FOAF and DC properties through the SPARQL engine of the framework jena 2-6-4 (*https://jena.apache.org/*). Several functions have been defined to aggregate and standardize heterogeneous properties. 80% of the data sets allocated to the training phase and 20% for testing phase.

4.2 Relevance Metrics

To evaluate the method presented in this article, we held a special metric and widely used in the FS, it is MAE, and two other metrics, recall and precision of information retrieval field [16,9].

- MAE: Mean Absolute Error, calculating the mean absolute difference between predictions p_i retained by the system and the real evaluations e_i given by users. This measure is simple to implement and directly interpretable.

$$MAE = \frac{\sum_{i=1}^{N} |p_i - e_i|}{N}$$

- *Precision*: it is the ratio between the number of relevant items returned by the system and the total number of items returned.

$$P = \frac{N_{pr}}{N_r}$$

- *Recall*: it is the ratio between the number of relevant items returned by the system and the total number of existing relevant items in the database.

$$R = \frac{N_{pr}}{N_p}$$

These metrics respectively measures the error, the effectiveness and the quality of FS.

4.3 Results and Discussion

In this section, we discuss the experimental results obtained, for that, we divide the dataset size in two parts, one having a proportion of 80% has dedicated for training phase and the other of proportion of a 20% has dedicated for test phase. From Figure 2, the curves show that the MAE error is minimal in the neigh-

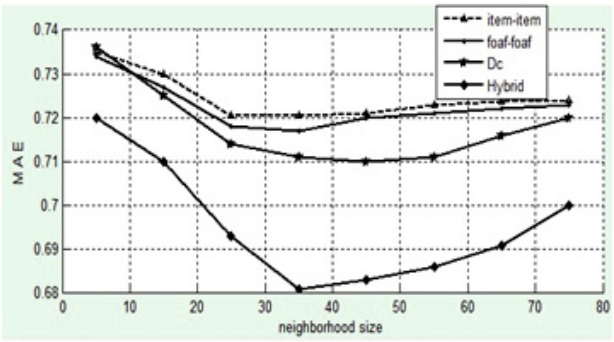

Fig. 2. Comparison of MAE

bourhood range [25-45] and important in outside of this range, it means that as the number of neighbours is less than 25 so there are not enough neighbours to calculate the similarity which lowers the prediction quality, unlike the other side, or the number of neighbours exceeds 45, there are sufficient neighbours, but less similar which degrades prediction quality , this explains that between 25 and 45 there are enough better similar neighbours. Also we observe that the DC curve illustrates a slightly favourable result compared to the FOAF curve, as the items are identified and enriched by descriptions and meta-data with certain stability better than valorising links and subjective opinions between a user′s networks. The best result is obtained in Hybrid curve, or the error is reduced to 0.68 for a neighbourhood size of 35, this favourable result is argued by exploiting items implicit information′s and estimating attributes of user profiles and links between them such as *see also* or *know* properties, which form a social network on the web and therefore a rich database that reduces the MAE, in addition, taking into account the opinions of users through their notes with respect to the items recommended what leads to a profitable collaboration. Two conclusions can be drawn the benefit of this additional data mass reduces the effect of sparsity as a problem moderating filtering systems, and adequately addresses the cold start problem for a new item. Moreover, the URI clause for the unique

resource identification in rdf documents lowers the effect of scalability. In the experiment below, we study the behaviour of our algorithms via the precision and recall metrics. Figure 3 shows a better accuracy rate (up to 73%) for the Hybrid solution, indicates the ability of the system to reject irrelevant items with minimal attribute values.

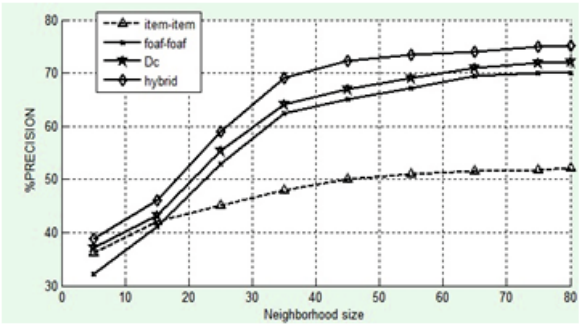

Fig. 3. Precision rate

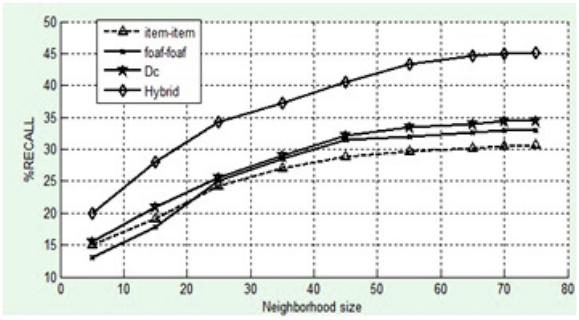

Fig. 4. Recall rate

We also observe that the recall rate (figure 4) which reaches a maximum rate of 45% for the optimal Hybrid solution involves the role of property values of adopted vocabularies to filter only the relevant items.

5 Conclusion and Future Work

Filtering systems are powerful and widely used systems on the web, especially for e-commerce or custom search. Our idea is not to hold closed applications that hide behind a particular data warehouse, but go further, and exploit all

kinds of information and to highlight it for integrity, dissemination and interoperability. In order to alleviate the limitations of collaborative filtering systems, we have presented in this paper, a hybrid model based on the FOAF formalism to better appreciate and enrich user profiles via social networks and information networks. The weighted classification that we have defined for the representation yield more adaptable and flexible profiles and still better adjustable, which alleviate the sparsity problem. On the other hand, the use of DC elements to describe items in a standard way leads to the good development of communities and overcome the problem of cold start for a new resource. The notable progress in the results founded by the formal use of meta-data to describe the valued resources and links with a standard and unified structure. Moreover, the union of similarities adopted for the recommendation is considered a balance between using different data sources and therefore increased the quality of prediction. In our opinion, the system model seems to a network of resources in collaboration with a network of properties describing these resources. The adoption of RDF syntax to the representation and implementation ensures openness, sharing and interoperability of all kind of data on the web, thus allows concretizing and developing semantics via these new practices, we think it is important to study the problem of scalability and reduce the computation time through the reduction techniques of the vector space, thus we also plan to still improve the rate of recall by the semantic disambiguation techniques of the users profiles.

References

1. Sieg, A., Moba, B., Burke, R.: Improving the effectiveness of collaborative recommendation with ontology-based user profiles. In: Proceedings of the 1st Interna Workshop on Information Heterogeneity and Fusion in Recommender Systems, HetRec 2010 (2010)
2. Sarwar, B.M., Karypis, G., Konstan, J.A., Riedl, J.: Item based collaborative filtering recommendation algorithms. In: Proceedings of the 10th International Conference on World Wide Web (WWW 2001), pp. 285–295 (May 2001)
3. Goldberg, D., Nichols, D., Oki, B.M., Terry, D.: Using collaborative filtering to weave an information tapestry. Communications of ACM 35(12), 61–70 (1992)
4. Adomavicius, G., Tuzhilin, A.: Toward the next generation of recommender systems: a survey of the state-of-theart and possible extensions. IEEE Transactions on Knowledge and Data Engineering 17(6), 734–749 (2005)
5. Adomavicius, G., Jingjing, Z.: Stability of Collaborative Filtering Recommendation Algorithms. Citeseer (2012), doi:10.1.1.221.7584
6. Hassanzadeh, H., Keyvanpour, M.R.: Semantic Web Requirements through Web Mining Techniques. International Journal of Computer Theory and Engineering 4(4) (August 2012)
7. Konstan, J.A., Riedl, J., Borchers, A., Herlocker, J.L.: Recommender systems: a GroupLens perspective. In: Recommender Systems, Papers from 1998 Workshop. Technical Report WS98-08. AAAI Press (1998)
8. Herlocker, J.L., Konstan, J.A., Terveen, L.G., Riedl, J.: Evaluating collaborative filtering recommender systems. ACM Transactions on Information Systems 22(1) (2004)

9. Abrouk, L., Gross-Amblard, D., Cullot, N.: Community Detection In The Collaborative Web. International Journal of Managing Information Technology 2(4) (2010)
10. Albanese, M., dAcierno, A., Moscato, V.F., Persia, A.: A multimedia recommender system. ACM Transactions on Internet Technology (TOIT) 13(1) (2013)
11. Cuong Pham, M., Cao, Y., Klamma, R., Jarke, M.: A Clustering Approach for Collaborative Filtering Recommendation Using Social Network Analysis. Journal of Universal Computer Science 17(4) (2011)
12. Beam, M.A., Michael, A., Kosicki, G.M.: Personalized News Portals: Filtering Systems and Increased News Exposure. Journalism & Mass Communication Quarterly 91(1), 59–77 (2014)
13. Mohammadnezhad, N., Mahdavi, M.: An effective model for improving the quality of recommender systems in mobile e-tourism. International Journal of Computer Science & Information Technology 4(1) (February 2012)
14. Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., Riedl, J.: Grouplens: an open architecture for collaborative filtering of netnews. Proceedings ACM (1994)
15. Bahrehmand, A., Rafeh, R.: Proposing a New Metric for Collaborative Filtering. Journal of Software Engineering and Applications 4, 411–416 (2011)
16. Burke, R.: Hybrid recommender systems: survey and experiments. UserModelling and User-Adapted Interaction 12(4), 331–370 (2002)
17. Meyffret, S., Médini, L., Laforest, F.: Confidence on Collaborative Filtering and Trust-Based Recommendations. In: Huemer, C., Lops, P. (eds.) EC-Web 2013. LNBIP, vol. 152, pp. 162–173. Springer, Heidelberg (2013)