Improving a Recommender System Through Integration of User Profiles: a Semantic Approach

Jonathas Magalhães
Federal University of Campina Grande
Campina Grande - PB - Brazil
jonathas@copin.ufcg.edu.br

Cleyton Souza
Federal University of Campina Grande
Campina Grande - PB - Brazil
cleyton.caetano.souza@gmail.com

Priscylla Silva
Federal University of Alagoas
Maceió - AL - Brazil
pmss@ic.ufal.br

Evandro Costa
Federal University of Alagoas
Maceió - AL - Brazil
evandro@ic.ufal.br

Joseana Fechine
Federal University of Campina Grande
Campina Grande - PB - Brazil
joseana@dsc.ufcg.edu.br

ABSTRACT
The users are present in multiple social networks/virtual communities and each one can be considered as a source of information about this user. In face to this question it is important a mechanism to integrate the user profiles. Through the integration of user profiles it is possible identifier more accurately their interests analyzing other data sources that they are present, possible reducing the *cold-start problem*. In this context, we present a semantic approach to help integrate data from multiple sources, for the construction and maintenance of user profiles that will be used to improve the quality of a recommender system. To integrate data from multiple sources, we defined a heuristic that quantifies the importance of each data source for a given user. To validate our approach, we perform a case study, where the solution was coupled into a recommender system of papers focused in Software Engineering domain. The user profiles were built extracting their information from the Brazilian Curriculum Vitae database named CV-Lattes, an academic platform, and Linkedin, a business network. We compared the quality of the recommendation based on the profiles integrated and non-integrated. The results show the superior quality of the recommendation based on integrated profile.

Author Keywords
User profile, building user profile, integration of user profile, maintaining user profile.

ACM Classification Keywords
H.3.3 Information systems Search and Retrieval: [Information Search and Retrieval; Retrieval models]

CHI 2009, April 4 - 9, 2009, Boston, Massachusetts, USA.

INTRODUCTION
The advent of so-called social web transformed the users, now they are not just consumers of content, but they have an important participation in the creation of content. At the same time that it facilities to its users, offering services that yield information such as blogs, forums, social networks, among others it has helped to accelerate the growth of digital content. However, the excess of available information and the difficulty of find relevant content led these users face the well-known problem of information overload [2].

Personalized recommendation systems [1] have been used to alleviate this problem. In the content-based approach, the recommender system models the interests of user in a profile that is built based on the features of the contents associated to the user and recommends other items with similar features [8]. An interesting question is that the users are present in multiple social networks/virtual communities and each one can be considered as a data source about this user [12]. For example, in Figure 1, the user has profiles in three different data sources (could be, for instance: a social network, a CV online and a blog). Each data source maintains a profile to represent the user (profile description). This profile describes the interests of the user and can be represented in many ways, e.g., a list with all posts that the user has done.

Thus, users can have many profiles representing his interests, in face to this question it is important a mechanism to integrate the user profiles. Through the integration of profiles it is possible identifier more accurately their interests analyzing other data sources that they are present [4]. In this context, we present a semantic approach to help integrate data from multiple sources, for the construction and maintenance of user profiles that will be used to improve the quality of a recommendation system.

The Figure 2 presents the process of profiles integration and where it fits into the recommendation process. So, the integration process is a preliminary step to recommending. The multiple profiles are integrated using our semantic approach resulting in a more complete profile.

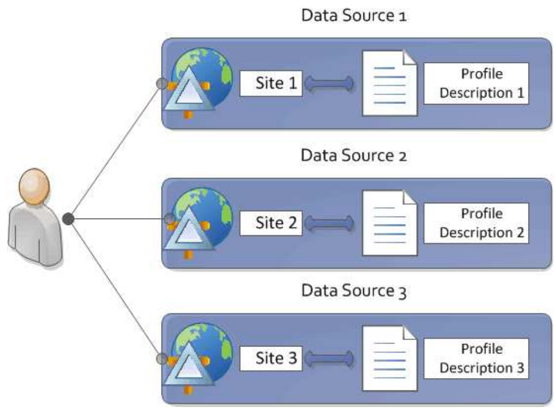

Figure 1. The user in multiple data sources.

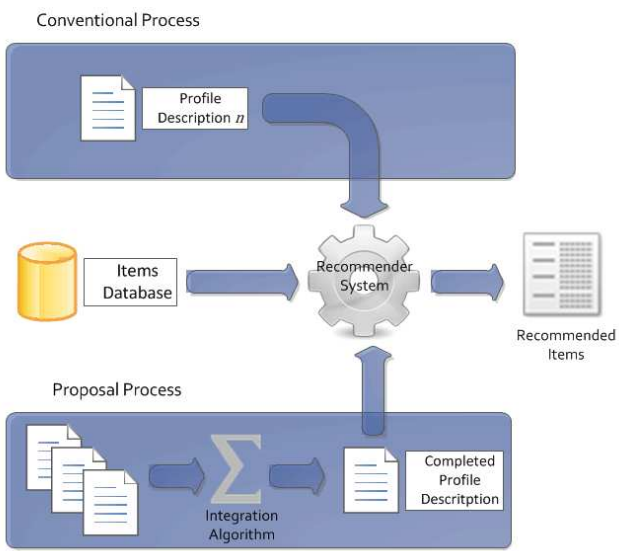

Figure 2. The user in multiple data sources.

Therefore, the main purpose of our work is answer the following research question Q_1: *Will the profiles integration improve the quality of recommendation of a recommender system?* In this context we consider an improve in recommendation when the relevance of the recommended items is maximized and when the most important items are prioritized. So, to answer this question, we define the following hypotheses:

- $H_{a,1}$ – Yes, the profiles integration will improve the recommender system.
- $H_{0,1}$ – No, the profiles integration will not improve the recommender system.

Other question about the integration of profiles is that we expect that our approach may reduce the well-know *cold-start problem* of new users in a data source. So, other investigation of this work is answer the following research question Q_2:*Will the profiles integration reduce the cold-start problem in a recommender system?* For this question we define the following hypotheses:

- $H_{a,2}$ – Yes, the profiles integration will reduce the *cold-start problem*.
- $H_{0,2}$ – No, the profiles integration will not reduce the *cold-start problem*.

To validate our approach and answer the presented research questions, Q_1 and Q_2, we perform a case study where the solution was coupled into a recommender system of papers focused in Software Engineering domain. Ten volunteers participated of the case study, their profiles were built extracting their information from the platform CV-Lattes[1], a Brazilian academic database contained an academic resume, institutions and research groups, and (ii) the Linkedin[2], a network used by professionals where the users maintain a profile about their interests and can interact with other professionals. Where four volunteers were new users in the platform CV-Lattes.

We compared the profiles integrated with non-integrated profiles according to the recommendation quality using the metric *Normalized Discounted Cumulative* (nDCG). This comparison was done with all users to verify the Q_1 and with new users to verify the Q_2. The results show that in both cases (all users and new users) the integrated profile improved the quality of the recommendation, thus the presented alternative hypothesis, $H_{a,1}$ and $H_{a,2}$, were accepted.

DEFINITIONS

In this section, we present the preliminary definitions about the problem and variables used. First, we present the problem definition, so the construction of the domain knowledge, then the construction of the user profile in one data source.

Problem Definition

Let $S = \{s_1, s_2, ..., s_{|S|}\}$ be the set of data sources and $U = \{u_1, u_2, ..., u_{|U|}\}$ be the set of users, where a source $s \in S$ is characterized as being a place that allows to users create and consume content. The set of contents available to the users $u \in U$ is defined by $I = \{i_1, i_2, ..., i_{|I|}\}$, where the content i has the attribute: m_i that represents a text description of the content.

The set $I_{u,s} = \{(i,t)_1, (i,t)_2, ..., (i,t)_{|I_{u,s}|}\}$ represents the contents $i \in I$ that the user u consumed in the data source s, the label t represents the time when the user created the content. In this model, the content can be videos, music, papers, posts, etc, since they have a text description to represent them. The user u has his preferences in the data source s represented by a profile $\vec{p_u^s}$ that is content-based built using the set $I_{u,s}$ [1]. Our goal in this work is define and validate a function $integration$ that integrates the user profiles and returns an unique profile that represents the interests of the

[1] `http://lattes.cnpq.br/`
[2] `http://www.linkedin.com`

user in all data sources, thus:

$$\vec{p_u} = integration(\vec{p_u^{s_1}}, \vec{p_u^{s_2}}, ..., \vec{p_u^{s_{|S|}}}). \quad (1)$$

Determining the Domain Knowledge

We used an ontology-based approach to represent the domain, the use of ontology aggregates semantic to the profile. We follow the approach proposed by Loh et al. [7]. The ontology O is defined as a tuple $O = (C, E, K)$. Where $C = \{c_1, c_2, ..., c_{|C|}\}$ is the set of concepts associated to the domain, each concept c is a node in the ontology O. The set $E = \{e_1, e_2, ..., e_{|E|}\}$ represents the taxonomy among concepts, where $e = (c_y, c_j)$ means that the concept c_l is parent of c_j. The concepts are disposed in a tree structure, i.e., each concept can have more than one child, but can have only one parent. The set $K = \{k_1, k_2, ..., k_{|K|}\}$ represents the terms (words) that are associated with the concepts $c \in C$.

Each concept $c \in C$ is represented by a vector of weights, thus $\vec{c} = (w_{1,c,O}, w_{2,c,O}, ..., w_{|K|,c,O})$, where the weight $w_{k,c,O}$ is associated with the tuple (k, c, O). The weight $w_{k,c,O}$ can be considered the probability of the term k be related to the concept c, i.e., $w \propto P(c|k)$.

The weights $w_{k,c,O}$ are calculated statistically based in a training set of documents $D = \{d_1, d_2, ..., d_{|D|}\}$, each concept has a training set $D_c \subset D$. A document $d \in D$ contains a text description, the stop-words are disregarded, then the weight $w_{k,c,O}$ is defined by the TF-IDF [10]:

$$w_{k,c,O} = tf(k, D_c) * idf(k, D), \quad (2)$$

where $tf(k, D_c)$ is the *term frequency* of k in D_c and $idf(k, D)$ is defined by:

$$idf(k, D) = \log\left(\frac{|D|}{|D^k|}\right), \quad (3)$$

where D^k is the set of documents that the term k occurs, thus $D^k \subset D$.

The Construction of the User Profile

We used the Space Vector Model to represent the user profile [2], so the profile $\vec{p_u^s}$ is a vector of terms:

$$\vec{p_u^s} = (w_{1,s,u}, w_{2,s,u}, ..., w_{|K|,s,u}), \quad (4)$$

where the weight $w_{k,s,u}$ represents the importance that the term k has to the user u in the data source s. The terms $k \in K$ were learned from ontology.

The weights $w_{k,s,u}$ are calculated based in the set of contents $I_{u,s}$ that the user u consumed in the data source s. The weight $w_{k,s,u}$ is defined by:

$$w_{k,s,u} = \sum_{(i,t) \in I_{u,s}} \frac{q(k,i)}{q(K,i)} * \lambda_t, \quad (5)$$

where $q(k, i)$ is the quantity of the terms k in the i, $q(K, i)$ is the quantity of all terms $k \in K$ in i. The Equation 5 is very related to the tf but difference is the temporal factor λ that gives more importance to the newer contents than the older in the set $I_{u,s}$. The factor λ is defined according to Lopes et al. [9]:

$$\lambda_t = \frac{v - \Delta t + 1}{v}, \quad (6)$$

where $\lambda_t \in [\frac{1}{v}, 1]$, v is the interval of years considered of a content in the user profile and Δt is the interval between the present year t_{now} and the year t of the content.

THE USER PROFILES INTEGRATION

In this section, we present the proposed solution to integration of user profiles. The profiles are composed by the weights related to the same terms of the ontology, but what should be the operation involving the profiles to compose the unique profile? To answer this question, we define the importance of a data source to an user based in his activity there. We define the activity $a_u^{s_j}$ of the user u in the source data s_j using the equation proposed by Souza et al. [3]:

$$a_u^s = \frac{t_{now} - t_{|I_{u,s}|} + \sum_{j=1}^{|I_{u,s}|-1}(t_{j+1} - t_j)}{|I_{u,s}| + 1}, \quad (7)$$

where t_{now} is the present time. In Equation 7, as lower is the value of a_u^s more active is the user, therefore we normalize the values of a_u so that $\sum_{s \in S} a_u^s = 1$ and as higher is the value of a_u^s more active the user will be in the data source s.

So, with the activity of each user defined, we define the function $\vec{p_u} = integration(\vec{p_u^{s_1}}, \vec{p_u^{s_2}}, ..., \vec{p_u^{|S|}})$ as a linear combination:

$$\vec{p_u} = integration(\vec{p_u^1}, \vec{p_u^{s_2}}, ..., \vec{p_u^{|S|}}) = \sum_{s \in S} \vec{p_u^s} \cdot a_u^s, \quad (8)$$

EVALUATION

This section presents the methodology to evaluate our approach and answer our research question, we elaborated a study case. The integration solution was coupled in a recommender system of papers in the Software Engineering domain.

The domain knowledge is based in ontology, in this study, we adapted the ontology proposed by Wong et al. [13]. The ontology has a total of 27 concepts. To learn the terms and their weights, for each concept in ontology we established a training set of 100 papers. The papers were obtained through the Mendeley API[3], for each concept we perform a search using as query the concept description and search results were manually verified to define the training set. We used the title and the abstract of the papers to build the vectors for each concept.

The user profile was constructed using two data sources: (i) the platform CV-Lattes, a Brazilian academic database contained academic resume, institutions and research groups,

[3] `http://dev.mendeley.com/`

and (ii) the Linkedin, a network used by professionals where the users maintain a profile about their interests and can interact with other professionals. The profile from CV-Lattes is built using the publications of papers of the users and the profile from Linkedin is built using the field *Expertise and Skills* where the users can determine their professional skills and the topics where are experts.

The recommender system used to generate recommendation to the users is content-based [1]. Let the set $I^{rec} \subset I = \{i_1, i_2, ..., i_{|I^{rec}|}\}$ be the papers available to recommendation, where each paper is represented by a vector of weights $\vec{p_i}$ calculated using the TF-IDF, as showed in Equation 3. For an user u is recommended a set $I_u^{rec} \subset I^{rec}$ with the n papers more similar to his profile $\vec{p}_u$ according the similarity measure sim:

$$I_u^{rec} = \underset{i \in I^{rec}}{\overset{n}{\operatorname{argmax}}}\, sim(\vec{p_u}, \vec{p_i}), \tag{9}$$

where the function sim calculates the similarity between two vectors, in this work we utilized the Cosine similarity [1].

$$sim(\vec{x}, \vec{y}) = \cos(\vec{x}, \vec{y}) = \frac{\vec{x} \cdot \vec{y}}{\parallel \vec{x} \parallel \times \parallel \vec{y} \parallel}. \tag{10}$$

In this case study participated ten researches that have profile in CV-Lattes and Linkedin, of which four of them were *cold-start* of CV-Lattes. We define as cold-start the users that have less than five publications in the CV-Lattes [5][4]. So, it was suggested to the volunteers recommendation using three different strategies of user profile: i) using the profile built from platform CV-Lattes (lattes); ii) using the profile built from Linkedin (linkedin) and iii) using the integrated profile (integrated). For each subject, we recommended a list with 15 papers, among which five for each type of recommendation, they were not informed how the recommendation were done.

The system recommended papers with information loaded from the digital library CiteerSeerX[5], with 40,855 papers (I^{rec}). The volunteers evaluated the quality of papers recommendation according five degrees of quality, then we mapped the degrees in a scale of relevance from 0 to 4, the Table 1 shows the degrees of relevance and the corresponding values of relevance.

For each user, we compare the profiles strategies using the metric (nDCG). This metric computes a comparison between a vector of relevance returned by the recommender system and an optimal vector, so if the most relevant documents are in the top of the recommendation list, higher will be the score. The nDCG is calculated by [6]:

$$nDCG = \frac{DCG}{DCG_{ideal}}, \tag{11}$$

[4]The work in [5] considers as *cold-start* users the users who have expressed less than five ratings.

[5]`http://citeseer.ist.psu.edu/index`

Table 1. The degrees of relevance used by the users in the paper evaluation.

Degree	Value of Relevance
Inadequate	0
Bad	1
Average	2
Good	3
Excellent	4

where DCG is defined by:

$$DCG = \sum_{j=1}^{n} \frac{2^{r_j} - 1}{\log_2(1+j)}, \tag{12}$$

where r_j is the relevance gave to the paper in the j-th position in the recommended list and DCG_{ideal} is the DCG when all returned papers are *excellent*.

RESULTS AND DISCUSSION

In this section we present and discuss the results obtained in the case study. The Figure 3 presents the distribution of the degrees of relevance among the three strategies of profile construction, we retired the *excellent* degree because there were not papers with this classification.

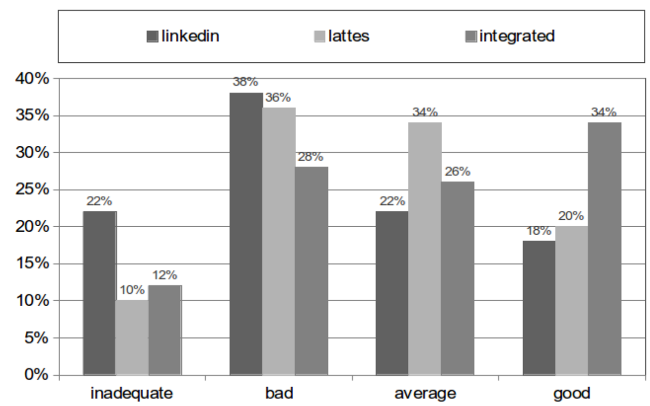

Figure 3. The evaluation of the recommendation using the different strategies to build the user profile.

Analyzing the Figure 3 is possible note that the integrated profile in comparison with the other strategies obtained a greater percentage of *good* feedback. This fact is an indication of the advantage of the integrated profile over the other profiles.

To give a better explanation about the results, we compare the profiles using the nDCG. The Table 2 presents the results of the nDCG for all users using the three strategies. Analyzing the Table 2 is possible see that the integrated profile obtained better results than the other strategies. Comparing user by user, in 30% of cases (users 4, 6 and 8) the nDCG of the integrated was worse than the other strategies. The advantage of the integrated profile is more clearly noted in the Figure 4, where a graphic representation is given by boxplot.

Table 2. The results of the nDCG for all users using the three strategies. Observation: the *cold-start* users are assignment with the field *Cold* checked and the bold numbers indicate the better result for that user.

User	Cold	Linkedin	Lattes	Integrated
1	✓	**0.1050**	0.0087	**0.1050**
2		0.1861	0.2444	**0.2785**
3		0.2036	0.1118	**0.2402**
4		**0.3706**	0.2739	0.3412
5	✓	0.1929	0.1762	**0.2499**
6		0.1717	**0.2375**	0.2023
7	✓	0.0440	0.0951	**0.3007**
8	✓	0.0718	**0.1178**	0.0718
9		0.2503	**0.2954**	**0.2954**
10		0.0000	**0.3475**	**0.3475**
Mean		0.1596	0.1908	**0.2433**
Mean (cold)		0.1034	0.0994	**0.1819**

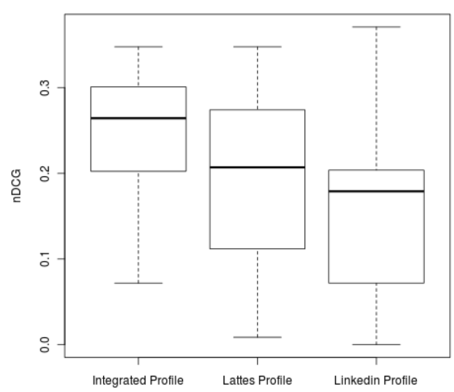

Figure 4. The boxplot comparing the three different strategies to build the user profile.

To validate this conclusion, we analyzed the data statistically. So, to choose the appropriate test, we first tested the normality of the data of the three strategies with the *Shapiro-Wilk* test ($\alpha = 0.05$), we obtained the following results: *integrated* (*p-value* = 0.2189), *lattes* (*p-value* = 0.8378) and *linkedin* (*p-value* = 0.8432). Thus, we can conclude that the three data are normally distributed. With the results of the normality test was possible choose the appropriate statistical test. So, we used the *Student's t-test*($\alpha = 0.05$) to perform a paired comparison between the strategies. The intention here was prove that the *integrated* was greater (greater values of nDCG) than *linkedin* and *lattes*, so we perform two *Student's t-test*: t_1 with $H_{a,1,1}$: *integrated is greater than lattes* and t_2 with $H_{a,1,2}$: *integrated is greater than linkedin*. Then, the Table 3 presents the results of the statistical comparison, thus we can concluded that the integrated profile improve the quality of the recommender system, i.e. the $H_{a,1}$ was accepted.

Concerning the *cold-start* users, in the Table 2 is showed that

Table 3. Comparison between the three strategies using the *Student's t-test* ($\alpha = 0.05$) with all users.

T-test	Alternative Hypothesis	p-value	Meaning
t_1	$H_{a,1,1}$	0.03152	$H_{a,1,1}$ accepted
t_2	$H_{a,1,2}$	0.02903	$H_{a,1,2}$ accepted

the integrated obtained better results than the other strategies. To confirm this, we perform a similar analysis to the presented before, first we verified the normality the data with the *Shapiro-Wilk* test ($\alpha = 0.05$). We obtained the following results: *integrated* (*p-value* = 0.4067), *lattes* (*p-value* = 0.8774) and *linkedin* (*p-value* = 0.5809). So, the three data are normality distributed. So, we used the *Student's t-test*($\alpha = 0.05$) to perform a paired comparison between the strategies, t_3 with $H_{a,2,1}$: *integrated is greater than lattes* and t_4 with $H_{a,2,2}$: *integrated is greater than linkedin*. The Table 4 presents the results, thus we can concluded that the integrated profile reduced the *cold-start problem*, i.e., the $H_{a,2}$ was accepted.

Table 4. Comparison between the three strategies using the *Student's t-test* ($\alpha = 0.05$) with the cold users.

T-test	Alternative Hypothesis	p-value	Meaning
t_1	$H_{a,2,1}$	0.03152	$H_{a,1,1}$ accepted
t_2	$H_{a,2,2}$	0.02903	$H_{a,1,2}$ accepted

Treats to Validity

Although we have achieved good results with our approach, we verified three treats to validity of our work: i) *The small number of volunteers* – the number of volunteers (10) is due to the computational effort to construct the integrated profiles and analyze the papers (40,855) to be recommended. But we plan experiments with a greater number of volunteers to increase the significance of our findings. However, even with just ten volunteers was possible to confirm the superiority of recommendation based on the integrated profile; ii) *The limited domain* – we performed this study in the Software Engineering domain, however we pretend to perform experiments in other domains expanding the used ontology; iii) *The limited number of data sources* – in our study we used only two data sources: the CV-Lattes and Linkedin. These data sources were chosen because they were more related with the type of content that we recommended (papers). Using other data sources, e.g., Facebook[6], Twitter[7], possibly the model will achieve different results. So, we pretend to study what kind of information in those data sources that are relevant to type of content that we want to recommend.

RELATED WORK

Some proposal for integrate user profiles have been considered in literature (e.g., [11, 12, 4]), but they consider different aspects about integration. In [11] is presented a solution

[6] `http://www.facebook.com/`
[7] `https://twitter.com/`

about extraction and integration of user profiles, it contains attributes about the user as affiliation, address, birth date, etc. Our work is different from [11] in sense that we want to integrate the interests of the users, not his attributes. The work of Wang et al. [12] presents the SocConnect, a system that permits users to management their own information from multiple social networks. The SocConnect provides recommendation of content based on the feedback of the users, however the system does not analyze text descriptions of the contents.

The work of Guy et al. [4] is closely related to the present one. They present the SONAR, an API for sharing social network data and aggregating it across applications. They investigate how integrate the connections among the users in different data sources comparing different linear combinations. Our work present as main difference the objective of integration, we study how to integrate the content produced by the user. Another difference is that we use the activity of an user in a data source to compute the weight of the relation between the user and the data source.

Concerning the use of ontologies to represent domain concepts and learn related terms, we follow a similar approach of the purpose by Loh et al. [7], however they do not use the ontologies with the objective of integrate user profiles.

In relation of recommendation of papers we can cite the work of Lopes et al. [9], that presents a recommender system of papers. The user profile is built using his academic resume from Lattes, but they do not use semantic to define the user profile and not use information of other data source.

CONCLUSION

The users are increasingly present in different data sources, thus is important a mechanism to integrated the information generated by them. Other problem well-know in personalized applications, specially recommender systems, is the cold-start that occurs when the user is new in a data source. A possible solution to this problem is analyze other data sources that the user is present to improve his profile.

In this paper we presented a semantic based approach to integrate the user profiles from multiple data sources. To validate our approach we performed a case study, the results obtained confirm the effectiveness and applicability of our approach in improve the quality of the recommendation. The main contribution of our work is the mechanism of integrate user profiles, that is easily adaptable to other contexts and other personalization systems. For future work, we will conduct experiments to compare our approach with other recommender systems of papers present in literature.

ADDITIONAL AUTHORS

REFERENCES

1. 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*, 17(6):734–749, 2005.
2. R. A. Baeza-Yates and B. A. Ribeiro-Neto. *Modern Information Retrieval - The Concepts and Technology behind search, Second edition*. Pearson Education Ltd., Harlow, England, 2011.
3. C. C. de Souza, J. Magalhães, and E. d. B. Costa. A Formal Model To The Routing Questions Problem In The Ccontext Of Twitter. In *IADIS Internaitional Conference WWW/Internet (ICWI 2011)*, 2011.
4. I. Guy, M. Jacovi, E. Shahar, N. Meshulam, V. Soroka, and S. Farrell. Harvesting with sonar: the value of aggregating social network information. In *Proceedings of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, CHI '08, pages 1017–1026, New York, NY, USA, 2008. ACM.
5. M. Jamali and M. Ester. A matrix factorization technique with trust propagation for recommendation in social networks. In *Proceedings of the fourth ACM conference on Recommender systems*, pages 135–142. ACM, 2010.
6. K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446, Oct. 2002.
7. S. Loh, D. Lichtnow, T. Borges, G. Piltcher, and M. Freitas. Constructing Domain Ontologies for Indexing Texts and Creating Users' Profiles. In *Work. on Ontologies and Metamodeling in Software and Data Engineering, Brazilian Symp. on Databases, UFSC, Florianópolis*, number 2003, pages 72–82, 2006.
8. S. Loh, F. Lorenzi, G. Simões, L. K. Wives, and J. P. M. de Oliveira. Comparing keywords and taxonomies in the representation of users profiles in a content-based recommender system. In *Proceedings of the 2008 ACM symposium on Applied computing*, SAC '08, pages 2030–2034, New York, NY, USA, 2008. ACM.
9. G. Lopes, M. Souto, and L. Wives. Personalizing bibliographic recommendation under semantic web perspective. In *International Workshop on Web Information Systems Modeling, WISM'07; CAISE'07*, pages 779–790, Trondheim, Norway, 2007.
10. G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Inf. Process. Manage.*, 24(5):513–523, 1988.
11. J. I. E. Tang, L. Yao, D. U. O. Zhang, and J. Zhang. A Combination Approach to Web User Profiling. In *Discovery*, volume 5, 2010.
12. Y. Wang, J. Zhang, and J. Vassileva. Personalized Recommendation of Integrated Social Data across Social Networking Sites. In *Workshop on Adaptation in the Social Semantic Web*, Hawaii, USA, 2010.
13. P. Wongthongtham, E. Chang, T. Dillon, and I. Sommerville. Development of a software engineering ontology for multisite software development. *Knowledge and Data Engineering, IEEE Transactions on*, 21(8):1205 –1217, aug. 2009.