Quality Web Information Retrieval: Towards Improving Semantic Recommender Systems with Friendsourcing

Alicia Díaz
LIFIA, Fac. Informática, UNLP
La Plata, Argentina
+54 221 4236585
alicia.diaz@lifia.info.unlp.edu.ar

Regina Motz
INCO, Fac. Ingeniería, UdelaR
Montevideo, Uruguay
+59 8 2711 42 44
rmotz@fing.edu.uy

Alejandro Fernández
LIFIA, Fac. Informática, UNLP
La Plata, Argentina
+54 221 4236585
casco@lifia.info.unlp.edu.a

José Valdeni de Lima
UFRGS
Porto Alegre, Brazil
+55 51 33166795
valdeni@inf.ufrgs.br

Diego López
UniCauca
Popayán, Colombia
+57 2 8209800
dmlopez@unicauca.edu.co

ABSTRACT
Web content quality is crucial in any domains, but it is even more critical in the health and e-learning ones. Users need to retrieve information that is precise, believable, and relevant to their problem. With the exponential growth of web contents, Recommender System has become indispensable for discovering quality information that might interest or be needed by web users. Quality-based Recommender Systems take into account quality criteria like credibility, believability, readability. In this paper, we present an approach to conceive Social Semantic Recommender Systems. In this approach a friendsourcing strategy is applied to better adequate recommendations to the user needs. The friendsourcing strategy focuses on the use of social force to assess quality of web content. In this paper we introduce the main research issues of this approach and detail the road-map we are following in the QHIR Project.

Categories and Subject Descriptors
H.4 [**Information Systems Applications**]: Miscellaneous

General Terms
Algorithms, Management, Experimentation, Human Factors

Keywords
Recommender Systems, Collaborative Filtering, Social Networks, Friendsourcing, Ontology

1. INTRODUCTION
Website content quality is crucial in many domains, but it is even more critical in the health and e-learning domain. Users need to retrieve health information or learning object that should be precise, believable, and relevant to their problem. With the exponential growth of web contents, quality-based Recommender System (RS) has become indispensable for discovering new information that might interest to web users. Website RS are supposed to help and guide users in retrieving/finding high quality health-related web sites or learning objects (LO) according to the their needs.

The huge amount of existing web sites makes necessary the support on recommender systems to retrieve information tailored to our needs while maintaining efficiency ratio between the results and time spent in the recovery process. Furthermore this process must be done without neglecting other very important aspect that is the quality of information obtained. In this sense, Quality-based Recommender Systems (RS) take into account quality criteria like credibility, believability and/or readability.

Decentralized, intelligent RS automatically give an evaluation about the quality of the information sources in the web according to the consumers needs. Semantic Recommender Systems uses ontologies to determine semantically similar items following the widely known content-based, context-aware and collaborative recommendation techniques. However, an automatic evaluation of web content quality in the most of cases is very expensive and many times is inapplicable. Therefore, an approach that uses social force might be highly relevant.

Collaborative Filtering (CF) [6, 14, 3, 8] is well known as a technique to improve RS with a social feature. In traditional CF-based RS, entities are recommended to new users based on the stated preferences of other similar users [7]. Breese et al. [5] identify two major classes of prediction algorithms: memory-based and model-based algorithms. The first type of algorithms maintains a database of all users' known references for all items, and, for each prediction, performs some computation across the entire database. On the other hand, model-based algorithms first compile the users' preferences into a descriptive model of users, items, and/or ratings; recommendations are then generated by appealing to the model.

Breese's research has shown than CF can be improved by using content features or hybrid content-collaborative features e.g., [2,

7]. Despite their success in the industry, RS, even those CF-based, suffer from several problems. First, the sparseness of the user-item matrix seriously affects the recommendation quality. Second, RS ignore the connections among users, which lose the opportunity to provide more accurate and personalized recommendations [10]. A tag-based contextual CF model which takes into consideration user' context based upon tagging information such as the type of information available from recently popular social tagging systems, is proposed by Nakamoto et al. [11]. These systems provide a well suited combination of context clues through tags as well as important social connectivity among users.

In our QHIR Project[1] we would want to apply these results to the fact of tagging web resources with quality assessment, that we named quality tagging. We mean by *quality tagging* the activity of tagging web content with quality assessments. Quality assessment corresponds to different quality criteria like credibility, believability, readability, timeliness, accessibility, etc. Then, when an user is quality tagging a web content she/he gives a evaluation of this resource in terms of each quality criteria, for instance readability: good.

In terms of social web, CF falls in the category of crowdsourcing techniques, which use the wisdom of crowd's theory [15]. *Crowdsourcing* is a distributed problem-solving approach; in the classic use of the term, problems are broadcast to an unknown group of solvers in the form of an open call for solutions. Social tagging is a manner of crowdsourcing where the collaborative activity is to have a description of web contents. Crowdsourcing can be useful to get quality assessment of web content. According to Bernstein et al. [4], when information is known only to friends in a social network, traditional crowdsourcing mechanisms struggle to motivate a large enough user population and to ensure accuracy of the collected information. On the other hand, *friendsourcing* is a form of crowdsourcing aimed at collecting accurate information available only to a small, socially connected group of individuals. Two key challenges arise in such small-network situations: motivating enough members of a small pool of people to participate, and ensuring the accuracy of the generated information.

This is the reason why we have decided to apply a friendsourcing approach, mainly taking into account application domain like health and e-learning. People of both domains feel better comfortable in small group that contain them. Therefore, in this paper we introduce an novel approach to make recommendations based on the nearest social network, instead of considering the crowd.

The expected novel contribution of our project is to demonstrate that a CF strategy can also be performed using data collected by social networks. Thus, the application of these results to RS for health and e-learning domain can improve the quality of the recommendations according to the nearest social context of the user. Specifically, in this research we propose to extend RS with a friendsourcing CF strategy, this means to calibrate quality-based CF algorithm with the social network quality tagging.

The remain of the paper is organized as follow: Section 2 describes our previous work on the modeling of semantic recommender systems. Then, in Section 3, we discuss the research issues of the QHIR project as an extension of previous work. The Section 4 discusses the proposed evaluation that will be carried out by performing different experiment on the top a provided software system. Finally, we will briefly discuss the conclusions.

2. BACKGROUND RESEARCH

This project is the natural continuation of other two projects: the SALUS/CYTED and SALUS/PROSUL projects. Results of this project were reported in [12, 13, 17].

In this previous work, we were interested in the modelling of semantic recommender systems. The modeling of website recommender systems involve the combination of many features: website domain, metrics of quality, quality criteria, recommendation criteria, user profile, and specific domain features. When specifying these systems, it must be ensured the proper interrelationship of all these features. In order to ensure the proper relationships of all these features, we propose an ontology network, the *Salus* ontology. Currently, the development of ontologies to the Semantic Web is based on the integration of existing ontologies [9]. In this work we have followed this approach to develop the *Salus* ontology used in a Health Website Recommender System as an ontology network. More precisely, the *Salus* ontology is a network of ontology networks; this means that each component of the *Salus* ontology is itself an ontology network and all of them are related among each other.

Salus networked ontologies are interrelated by four different relations: *isAConservativeExtentionOf*, *mappingSimilarTo* and *isTheSchemaFor*, took from the DOOR ontology [1], and the *usesSymbolsOf* relation, defined by us, which describes an extension of a given ontology by importing individuals from another ontology.

The different specific-domain ontology networks correspond to the different knowledge domains conceptualized by the *Salus* ontology. The "Health" ontology network models the health domain: the "Health" ontology conceptualizes any diseases and the "Specific Health" ontology is a more specific disease. Both ontologies are related by the *isAConservativeExtentionOf* relation. Particularly, the *Salus* ontology is for the health domain, but in more recently work it was adapted to the e-learning domain [16].

The "WebSite" ontology network conceptualizes the domain of webpages and describes web resources considered in a quality assessment. The "WebSite Specification ontology" plays the role of a meta-model for the "WebSite ontology" (*isTheShemaFor* relation). Its main concepts are "Web Resource" and "Web Resource Property". A web resource is any resource which is identified by an URL. "Web resource properties" model the properties attached to a web resource.

"Quality Assurance" ontology network conceptualizes metrics, quality specifications and quality assessments, each one in an ontology. The relationship *mappingSimilarTo* exists between the "Quality Assessment" and the "WebSite Specialization" ontologies, in order to define an alignment between "Web Resource" and "Web Content" concepts.

"Context ontology network" describes "user profiles" and "query situation" resources. The "Context Specification" ontology is a meta-model to the "User Profile" and "Query Situation" ontologies.

[1] QHIR: Quality Health Information Retrieval at https://sites.google.com/site/laccirr1210lac007/

"Recommendation" ontology network describes the criteria of recommendation for a particular context and quality dimensions, and the obtained recommendation level. The "Recommendation Specification" ontology describes the criterion to make a recommendation. The "Recommendation" ontology models concrete recommendation assessments. This ontology uses the "Recommendation Specification" ontology.

The *Salus* ontology was designed to be populated automatically, although it is very hard to find out metrics of quality that can be perform over the Web. This ontology can also be used if it is populated by a domain expert.

After the population, the *Salus* ontology is used in combination with recommendation rules to automatically retrieve quality web content according to the user needs. The recommendation rules are also specified by a domain expert.

As was reported in [12], [13] and [17], the *Salus* ontology is very useful in the conceptualization of a content-based RS. Currently, we pretend to use the *Salus* ontology as the backbone of the new approach in order to model a friendsourcing based RS.

3. RESEARCH ISSUES

In order to extend our previous approach by considering a friendsourcing CF strategy, we need to face new research issues that can be enunciate as:

1. How the *Salus* ontology can be extended to conceptualize a friendsourcing-based RS?
2. What are the new features a CF algorithm needs to take into account to aim a friendsourcing-based CF approach?
3. How a content-based RS can use a friendsourcing-based CF results to calibrate recommendations?

Following subsections discuss these issues in more details.

3.1 A Semantic Friendsourcing-based RS

As was described in the previous section, the *Salus* ontology is the backbone of a semantic RS for the health domain. However this ontology does not consider the concept of friendsourcing. Particularly, in this new approach the semantic RS should also be able to record quality assessment coming from social quality tagging.

In order to face this issue, we need to extend the *Salus* ontology to conceptualize social networks and social quality assessments from quality tagging. Social networks, from an individual point of view, are considered as a set of friends, where she arranges her friends by interests, for instance, from school/job, neighbors, sport, diabetes, etc. The grouping of friends is the cue to define nearest context of a user. It is similar to the group functionality of MSN Microsoft Messenger) for organizing contacts.

According to previous remarks, it is needed to extend the *Salus* ontology to conceptualize social networks with grouping of friends and social quality assessment of web content. This extension will affect the "Quality Assurance" and "Context" ontology networks.

3.2 Friendsourcing-based CF Strategy

Regarding the second research issue, a CF algorithm that supports a friendsourcing strategy based on web content quality assessment should be proposed. The importance of social network is that they are a means of capturing quality criteria of the web information. People belonging to this social network will share quality criteria of shared resources. Applying friendsourcing CF algorithm, the RS will be able to adapt its recommendation to the user context, this means to profit from the fact that nearest friends' quality tagging will be more relevant for the user. However, according to the current state of the art of CF, there are no CF algorithm that considers the social network organization. In this work, we aim at defining a CF algorithm that takes into account how the user has organized her friends in order to adapt the recommendation to the opinion (quality assessment) of those friends that better fits to the intended use of demanded information.

3.3 Combining content-based and friendsourcing based RS

This issue involves the analysis and development of different strategies to combine a content-base RS approach and a friendsourcing-base one. There are many possibilities to do this. One is to consider both kind of RS independently; this means that both can be performed separately, recovering a set of recommendations. Then both set of recommendation can be integrated following different criteria, for instance by getting get the intersection of both sets. Other strategy could be to take the set of recommendation of one RS as input of the other one. There are two possibility of combine them: first to execute the content-base RS and then the friendsourcing-based and the opposite one. These strategies are very easy to implement because the different kind of systems are user as complement one of the other. There could be even more complex combinations, but in the first stage of this project we will only experiment with the aforementioned ones.

4. EVALUATION APPROACH

This research proposal also involves the evaluation of the aforementioned theoretical issues. In order to hold this evaluation, it will be develop a software system that will work as the interface of RS whose main requirements are:

- to support the quality tagging activity. This system will be the means to capture users' quality assessments of health web pages. User should be able to post quality tags to the current web content. The posted quality tags will feed the friendsourcing CF algorithm
- to be the interface to friendsourcing-based RS that recommends to suitable web contents to a user.

This interface will be designed as an add-on to a web browser. This add-on will be able to import the user's social network from other application as Facebook and MySpace. Then, a user can classify her friends according to group of interest.

After developing this system, we will perform an assessment of our approach. This assessment is made up of two phases. The first one is in charge of getting information to feed the friendsourcing CF algorithm, and the second phase is in charge of evaluating the usefulness of the algorithm.

- *1st evaluation phase - getting information to feed the CF algorithm.* It involves: to select a group of representative users to assess the quality of health web pages; carrying out the experiment: ask to the selected

user to do the quality assessment; and the population of the *Salus* ontology and seed the CF algorithm.

– *2nd evaluation phase -evaluation of the usefulness of the CF algorithm.* This phase involves two groups of users with similar characteristics who will be invited to use the RS. One group will use only a content-based RS and the other will use the friendsourcing-based RS. Finally, the users will be asked to respond about the usefulness of the recommendations in order to conclude about the usefulness of our approach.

5. CONCLUSIONS

In this paper we introduce an innovative approach to quality web information retrieval. The approach intent is to improve semantic recommender systems with a friendsourcing strategy. Particularly, our project is focused on the health and e-learning domain, due to the importance of getting high quality information in these areas. In this paper we have introduced the main research issues of this project and discuss the scope of them. We also have remarks the main theoretical issues to tackle during the research project and described the road-map of the project.

It should be noted that this project is an Ibero-American research project that involves 4 research groups coming from four different Ibero-American countries (as it is demonstrated by the collage of authors).

6. ACKNOWLEDGMENTS

This work was partially funded by: the Laccir R1210LAC007 project, the SALUS/CYTED and SALUS/PROSUL projects which are sponsored by the CnPq, Brazil and the CyTED, Spain; and finally by the PAE 37279-PICT 02203 which is sponsored by the ANPCyT, Argentina .

7. REFERENCES

[1] C. Allocca, M. d'Aquin, and E. Motta. Door – towards a formalization of ontology relations. In J. L. G. Dietz, editor, KEOD, pages 13-20. INSTICC Press, 2009.

[2] C. Basu, H. Hirsh, and W. Cohen. Recommendation as classification: using social and content-based information in recommendation. In Proc. of the Fifteenth national/tenth conference on Artificial intelligence/Innovative applications of artificial intelligence, AAAI '98/IAAI '98, pages 714-720, Menlo Park, CA, USA, 1998. American Association for Artificial Intelligence.

[3] L. Bergman, A. Tuzhilin, R. Burke, A. Felfernig, and L. Schmidt-Thieme, editors. RecSys '09: Proceedings of the third ACM conference on Recommender systems, New York, NY, USA, 2009. ACM. 609096.

[4] M. S. Bernstein, D. Tan, G. Smith, M. Czerwinski, and E. Horvitz. Personalization via friendsourcing. ACM Trans. Comput.-Hum. Interact., 17:6:1-6:28, May 2008.

[5] J. S. Breese, D. Heckerman, and C. M. Kadie. Empirical analysis of predictive algorithms for collaborative filtering. In G. F. Cooper and S. Moral, editors, Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence, pages 43-52, 1998.

[6] P. Brusilovsky, A. Kobsa, and W. Nejdl, editors. The Adaptive Web: Methods and Strategies of Web Personalization. Lecture Notes in Computer Science. Springer, Berlin, 1 edition, June 2007.

[7] N. Good, J. B. Schafer, J. A. Konstan, A. Borchers, B. Sarwar, J. Herlocker, and J. Riedl. Combining collaborative filtering with personal agents for better recommendations. In In Proc. of the Sixteenth National Conference on Artificial Intelligence, pages 439-446, 1999.

[8] D. Jannach, W. Geyer, C. Dugan, J. Freyne, S. Singh Anand, B. Mobasher, and A. Kobsa. Workshop on recommender systems and the social web. In Proceedings of the third ACM conference on Recommender systems, RecSys '09, pages 421-422, New York, NY, USA, 2009. ACM.

[9] A. Kleshchev and I. Artemjeva. An analysis of some relations among domain ontologies. Int. Journal on Information Theories and Applications, 12:85-93, 2005.

[10] H. Ma, M. R. Lyu, and I. King. Learning to recommend with trust and distrust relationships. In L. D. Bergman, A. Tuzhilin, R. D. Burke, A. Felfernig, and L. Schmidt-Thieme, editors, RecSys, pages 189-196. ACM, 2009.

[11] R. Y. Nakamoto, S. Nakajima, J. Miyazaki, S. Uemura, H. Kato, and Y. Inagaki. Reasonable tag-based collaborative filtering for social tagging systems. In Proceeding of the 2nd ACM workshop on Information credibility on the web, WICOW '08, pages 11-18, New York, NY, USA, 2008. ACM.

[12] E. Rohrer, A. Díaz, and R. Motz. Modeling a web site quality-based recommendation system. In The 12th International Conference on Information Integration and Web Based Applications and Services (iiWAS2010), pages 147-180. ACM Press, 2010.

[13] E. Rohrer, R. Motz, and A. Díaz. Ontology-based process for recommending health websites. In W. Cellary and E. Estevez, editors, I3E, volume 341 of IFIP, pages 205-214. Springer, 2010.

[14] B. Sarwar, G. Karypis, J. Konstan, and J. Reidl. Item-based collaborative filtering recommendation algorithms. In Proceedings of the 10th international conference on World Wide Web, WWW '01, pages 285-295, New York, NY, USA, 2001. ACM.

[15] J. Surowiecki. The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations. Doubleday, May 2004.

[16] A. Díaz, R. Motz, E. Rohrer, L. Tansini. An Ontology Network for Educational Recommender Systems. In Educational Recommender Systems and Technologies: Practices and Challenges, Eds. Olga C. Santos, Jesús G. Boticario. IGI Global, In Press 2011.

[17] Rohrer E., R. Motz, A. Díaz. 2010. Web Site Recommendation Modelling Assisted by Ontologies Networks. Anais dos Workshops SALUS/CYTED-CNPq, PROSUL-CNPq AvalSaúde e SticAmSUD-CAPES ALAP. Cadernos de Informática. 5(1), 49-68.