A Semantic Similarity Measure for Recommender Systems

Roza Lémdani, Géraldine Polaillon, Nacéra Bennacer, and Yolaine Bourda
SUPELEC Systems Sciences (E3S) - Computer Science Department
Gif-Sur-Yvette, France
{Roza.Lemdani,Geraldine.Polaillon,Nacera.Bennacer,Yolaine.Bourda}@supelec.fr

ABSTRACT

In the past few years, recommender systems and semantic web technologies have become main subjects of interest in the research community. In this paper, we present a domain independent semantic similarity measure that can be used in the recommendation process. This semantic similarity is based on the relations between the individuals of an ontology. The assessment can be done offline which allows time to be saved and then, get real-time recommendations. The measure has been experimented on two different domains: movies and research papers. Moreover, the generated recommendations by the semantic similarity have been evaluated by a set of volunteers and the results have been promising.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Information filtering

General Terms

Algorithms, Experimentation

Keywords

Recommender systems, ontology, semantic similarity

1. INTRODUCTION

Nowadays, data available on the Internet are beyond imagination, and looking for information is more and more difficult unless you know exactly what you are looking for. This is why, recommender systems have been introduced in the mid-1990s. Even if the recommendation task is perceived pejoratively, it has a real advantage for users and system owners. Apart from the significant increase in sales on the e-commerce websites, recommender systems are designed to help users by reducing their time and research effort. Such a mechanism requires the usage of the previous items (called the history) viewed by the user as an inspiration to calculate the recommendations. This history is used to build a user profile that will help to generate recommendations.

Several approaches exist for recommender systems [2]. Collaborative filtering and content-based recommendations are among the most popular. In collaborative filtering, an item is recommended to a user if it has been appreciated by similar users to this target user. Such approaches rely on ratings and thus suffer from the lack of ratings and the cold-start (new item or user) problem. One solution is to overcome this lack using the content- or rating-based similarities between items and exploit them to deduce a rating for items without ratings [4]. Content-based approaches [9] recommend to the users items with the same characteristics as those they have formerly appreciated. Thus, these systems require a description of the recommendable items (e.g. keywords) in order to compute the similarity between the items and assess the recommendations. The limitations of such systems are the overspecialization and the new user problem but we are not going to discuss them in this paper. Since semantic web data are more and more available (e.g. Linked Data[1]), we focus, in this paper, on data described in an ontology.

In the semantic web community, the similarity measures are more often used in ontology alignment. They are based on string matching [11], lexical matching using Jaccard, Dice or overlap coefficients, and structure matching [12, 10, 3]. The authors of [7] introduced a graph matching algorithm. Given a similarity measure between two nodes, the similarity is propagated between the nodes until a fixpoint is reached. In this case, the similarity measure takes into account the neighborhood of the target nodes and adjusts the similarity accordingly. These similarity measures are defined for ontology alignment and do not suit the recommendation domain.

Some recommendation approaches exploited the semantic information described in an ontology [8, 1]. The authors of [8] defined Quickstep, a research papers recommender system that relies on a taxonomy of subjects. The interests of the subjects are computed according to the user's subject history and the hierarchy is used to reason and extract the closest subjects to those of interest to the user. On the other hand, AVATAR [1], a TV program recommender system discovers semantic associations which consist in paths composed of properties which lead to recommendations according to a certain computed degree of interest. Both Quickstep and AVATAR are domain-dependent.

In our approach, we define a domain-independent similarity measure based on the semantics of the items which no longer relies on keyword description. The domain knowl-

[1]http://linkeddata.org/

I-SEMANTICS 2011, 7th Int. Conf. on Semantic Systems, Sept. 7-9, 2011, Graz, Austria

edge is entirely contained in an ontology and the similarity measure is automatically adapted after defining the recommendable items set. Also, the semantics is used differently than in [8, 1]. We consider that an item is defined by its surrounding individuals (by means of properties) and those individuals are also defined by their neighborhood. This allows two items to be compared not only by their explicit definition, but also by the definition of their neighborhood. Thus, the similarity is assessed in an iterated way inspired by the algorithm of [7] except that our similarity measure is adapted to the recommendation task. The defined measure can be used as the base of a semantic-based recommender system, as explained in this paper, and it is flexible enough to be included in other recommendation modules as additional knowledge to enhance the user profile. Moreover, the computation of the similarity values can be done off-line which allows time to be saved in the recommendation process and then, have a real-time recommendation.

The following section introduces the semantic similarity measure. In Sect. 3, the usage of the semantic similarity in the recommendation process is discussed. The approach is experimented and evaluated in Sect. 4. Finally, the conclusion and future work are exposed in Sect. 5.

2. THE SEMANTIC SIMILARITY

In the following, an ontology is denoted $\mathcal{O}(\mathcal{C}, \mathcal{R}, \mathcal{I})$ where $\mathcal{C}$ is a set of concepts, $\mathcal{R}$ a set of properties (relations) and $\mathcal{I}$ a set of individuals (instances).

The authors of [4] defined the semantic information independently from an ontology but relative to the items' attributes, their relations with other items and their role in these relations. Thus, semantic similarity is computed with a domain dependency and real knowledge of the structure and relations between items. We define the semantic similarity between two items in a generic way. The domain ontology allows independence between the semantic similarity computation and the domain knowledge. Thus, the computation can be non-supervised.

Definition 1. Given $\mathcal{C}$, the set of concepts of the ontology, and i and j, two individuals, the semantic similarity between i and j is a function $sim : \mathcal{C} \times \mathcal{C} \rightarrow [0, 1]$ such that i and j are similar if they are related to the same individuals by means of any property, or if they are related to individuals which are similar to themselves.

2.1 The Semantic Similarity Measure

Before computing the semantic similarity, the definition of the candidate pairs for the computation is required. The selected pairs for the semantic similarity computation are those of the same nature. In other words, the pairs of individuals that instantiate the same concept. This distinction is made because there's no point in computing the similarity between two individuals of different concepts. For example, the similarity between a film and a production company would not only be minimal or null, but also, it would not make any sense.

The similarities are initialized so that two identical individuals have the maximum similarity which is 1, while the similarity between two different instances is initialized to 0.

$$sim_0(i,j) = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

The next step consists in computing the semantic similarity between two individuals i and j in an iterative way. Once again, if the two instances are identical, the similarity equals 1. Otherwise, it is defined as being the average of the semantic similarities of the pairs (i', j') related to (i, j).

$$sim_{k+1}(i,j) = \begin{cases} 1 & \text{if } i = j \\ \sum_{E_{i,j}} \frac{sim_k(i',j')}{|E_{i,j}|} & \text{otherwise} \end{cases} \quad (2)$$

with $E_{i,j} = \{(i', j') | \exists R \in \mathcal{R}\ R(i, i') \wedge R(j, j')\}$

For $k = 0$, the semantic similarity $sim_1(i, j)$ between two individuals i and j only takes into account the individuals that are both related to i and j by a property, due to equation 1. In other words, only the individuals which are at a distance 1 from i and j (We say that an instance a is at a distance k from b if there is a sequence of k properties linking a to b).

For $k > 0$, the computed similarities sim_{k+1} in step k are taken into account. Intuitively, this means that in step $k+1$, instances being at a distance k from i and j impact on the similarity of these two instances. Indeed, at step k, the similarity of the items being at a distance $k-1$ from i and j (sim_1) is computed only based on the individuals that are both related to them, while at step $k+1$, their similarity (sim_2) also takes into account the similar individuals.

The function sim is convergent. Thus, the semantic similarity computation reaches a fixpoint. The proof can be found in the technical report [6].

2.2 Adding New Data in the System

Updating the system with new data makes the semantic similarity no longer correct because adding a new individual or a new relation between two individuals influences the semantic similarity values in the neighborhood. The actions of adding an individual and adding a relation between two individuals are distinguished. Thus, adding an individual with relations to other individuals consists in first adding the individual and then adding each relation one by one.

2.2.1 New Individuals

We consider that adding a new individual means adding an individual without any relation with other individuals. Consequently, adding an individual in the ontology implies computing all the pairs containing this individual a and the individuals b of the same nature. The semantic similarity of each pair (a, b) (except for (a, a)) at every step k equals zero since there is no property linking a to the other individuals.

2.2.2 New Relations Between Individuals

Adding a relation linking an individual a to an individual b influences the semantic similarity of each pair at a distance k from (a, b). The algorithm to update the similarity values is described below.

Algorithm 1 Update_sim(i, j, k_{max})

for all $k \in]0, k_{max}]$ **do**
  $sim_{k+1}(i,j) = \sum_{E_{i,j}} \frac{sim_k(i',j')}{|E_{i,j}|}$
  $propagate(E_{i,j}, k+1, k_{max})$
end for

2.3 Complexity

The complexity of assessment of the semantic similarity values for all the pairs of the system is about $n^2 \times k \times m^2$,

Algorithm 2 propagate($E_{i,j}$, k, k_{max})

if $k_{max} > k$ **then**
  for all $(i', j') \in E_{i,j}$ **do**
    $sim_{k+1}(i', j') = \sum_{E_{i',j'}} \frac{sim_k(i'', j'')}{|E_{i',j'}|}$
    $propagate(E_{i',j'}, k+1, k_{max})$
  end for
end if

where k is the maximum number of steps, n is the number of individuals and m the average number of properties for each individual. On the other hand, the complexity for adding a relation between two individuals is about $k^3 \times m^4$. Thus, the complexity for adding a new individual with his m relations is $k^3 \times m^5$, and then, the complexity for adding n individuals is $n \times k^3 \times m^5$. Consequently, when $n > k^2 \times m^3$, it is better to use the second approach.

3. SEMANTIC SIMILARITIES IN THE RECOMMENDATION PROCESS

In a recommender system, the similarity can help to find items of interest to the users, based on their history. Basically, this approach is entirely content-based. It is based on the semantics expressed in an ontology which makes it possible to reuse semantic web data that are more and more spread and available.

The recommendation, in this paper, consists in getting similar items to those appreciated by the user. Two items are considered as being similar if their semantic similarity is higher than a certain threshold. This threshold can be either fixed by an expert or learned by the system, by adapting the threshold according to the users' feedback. So, given a rank K and a user u who appreciated an item i, and given i' similar to i, the user interest e for i' is measured so that:

$$e(u, i') = sim_K(i, i') \times eval(u, i) \quad (3)$$

where $eval(u, i)$ is an explicit (item ratings, information entered in forms) or implicit (clickstream data, research) valuation of the item i by the user u.

This measure allows the recommendations to be sorted. Indeed, a 50% similarity should not be treated the same as a 90% one. Also, on a 1 to 5 rating scale, a history item rated 4 by the user, should not be treated the same as a 5 star rating. So, given a set of similar items to those appreciated by the user, the recommendation is made according to the implied order of e.

4. EXPERIMENTATION AND EVALUATION

We experimented the semantic similarity measure on movies and research papers. For each domain, every possible pair of individuals was first generated for each concept. Then, the semantic similarity was computed for each pair. Instead of running the computation process until reaching a fixpoint, we fixed the rank to $k = 10$, which is sufficient to analyze the evolution of the computed similarities from one rank to another. Finally, a set of volunteers evaluated the generated recommendations of the movie domain.

4.1 Research Papers

We used the Semantic Web Dog Food Corpus[2] for the research paper ontology. Each paper is related to a conference. It has one or many subjects and a list of authors. Moreover, each author has an affiliation. The research paper dataset is composed of 209 papers, 1282 authors, 216 affiliations, and 726 subjects. They belong to the 17th and 18th International World Wide Web Conference.

One can see on table 1 the number of paper pairs belonging to a certain similarity range for the ranks 0 to 10. The numbers show that the semantic similarity values evolve between the ranks 1 and 10. It also shows important values for the maximum similarity against lower values for the average and the standard deviation. This means that while the pairs that are not actually similar keep a low similarity value, the similarity of the similar pairs keep growing. This semantic similarity helps to find out similar pairs that are not necessarily related to the same individuals. For examples, let's consider two authors with two different affiliations and who don't have shared papers. A classical similarity measure will classify them as being not similar while the presented similarity measure will take into account their publications. Thus, if they have similar publications, they would be considered as being similar which is logical.

4.2 Movies

For the experimentation, we used the MovieLens[3] dataset, which a classical dataset for recommendation and the information available on IMDB[4]. We focused only on the concepts Film, Person, Actor, Director, Writer, Country, Language and Genre so that each movie is related to a certain number of persons who can be actors, directors or writers. Also, a movie has one or many genres (Action, Adventure, Animation, Children, Comedy, Crime, etc.), languages and countries. The movie dataset is composed of 523 movies, 1813 main actors, 413 directors, 876 writers, 54 languages, 33 countries and 24 genres.

Table 1 depicts the number of film pairs belonging to a certain similarity range for every rank. One can see that similarity values keep increasing from one step to another. Even if similarity values keep increasing, the increase is important at the beginning to become slight as we go along.

We had the recommendations applied on the movie domain evaluated by seventeen 16-40-year-old volunteers. The evaluation consisted exclusively in explicit valuations (ratings between 1 and 5) of the recommended items. First, each user rated about 20 movies. For each recommended item, the user was asked to check the like, dislike or N/A button. The similarity threshold has been fixed to 40% according to the analysis of the obtained similarities for rank 10. Also, some pairs, which were wrongly detected as non similar with the classical measure, have been correctly identified as similar after the propagation step. We presented the users 10 recommendations from the semantic similarity at rank 10 and 10 recommendations from the semantic similarity at rank 1 (classical measure).

The results of this evaluation are depicted on Fig. 1. We can see that 63,8% of the recommendations satisfy the users when the semantic similarity is propagated against 56,9% for a classical measure. These results are encouraging and we believe that it would perform better in a larger dataset with various data.

[2]http://data.semanticweb.org/

[3]http://www.movielens.org/

[4]http://www.imdb.com/

Table 1: Similarity distribution, standard deviation, max and average similarities by rank

			Sim											
			0	]0,0.1[	[0.1,0.2[	[0.2,0.3[	[0.3,0.4[	[0.4,0.5[	[0.5,0.6[	[0.6,1[	1	Std. dev.	Avg.	Max
Papers	Rank	0	21736	0	0	0	0	0	0	0	0	0	0	0
		1	20668	758	240	68	1	1	0	0	0	0,022	0,004	0,400
		2	18994	2120	454	123	32	7	3	3	0	0,034	0,008	0,775
		6	4422	16046	725	305	150	52	19	17	0	0,059	0,017	0,989
		10	3055	17140	718	431	203	128	29	32	0	0,071	0,024	0,999
Movies		0	136503	0	0	0	0	0	0	0	0	0	0	0
		1	4535	124618	7289	46	12	3	0	0	0	0,025	0,059	0,440
		2	28	128566	7784	89	23	9	4	0	0	0,026	0,060	0,532
		6	0	15228	97405	23245	542	47	22	14	0	0,049	0,159	0,847
		10	0	3992	22976	93598	15431	436	38	32	0	0,059	0,238	0,918

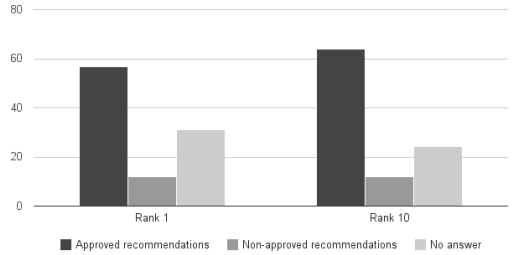

Figure 1: Users' evaluation of the semantic similarity-based recommendations

5. CONCLUSION AND FUTURE WORK

In this paper, a domain-independent semantic similarity measure has been defined. It allows the semantic similarity between a pair of individuals to be estimated and propagated. Thus, if applied to a pair of recommendable items, it helps to identify interesting items for the user. Indeed, finding similar items to those appreciated by the target user can lead to interesting recommendations. The assessment of the similarities is independent from the recommendation. Consequently, it can be run offline which allows execution time to be saved and then, have real-time recommendations.

We experimented the semantic similarity measure on the movie and the research paper domains. The propagation of the similarity measure detects similar pairs of individuals which have not been detected by classical similarity measures. Also, the semantic similarity has been applied in the recommendation process of the movie domain and the recommended items have been evaluated by a set of volunteers. The results showed that more satisfying items are recommended to the users when the semantic similarity is propagated. We believe that the results could be even better using a larger dataset with various data.

The semantic similarity assessment, being independent from the recommendation process, can be exploited in other ways. We plan to include this similarity measure in a previous work where we defined three recommendation modules: a collaborative, a semantic-based and a frequency module [5]. The collaborative module is based on association rule mining and including the semantic similarity to it could make the process more flexible. Also, the semantic-based module computes the interest of individuals according to the instances contained in the user profile. This assessment suffers from a lack of information which, can be alleviated by including the semantic similarity measure. Also, we plan to use the semantic similarity measure as an independent recommendation module and will improve the recommendation selection by adding a trust value for each individual so that an item which is similar to an item disliked by the user, will not be used for the recommendation.

6. REFERENCES

[1] Y. Blanco-Fernández, J. J. P. Arias, A. Gil-Solla, M. R. Cabrer, M. L. Nores, J. G. Duque, A. F. Vilas, R. P. D. Redondo, and J. B. Muñoz. A flexible semantic inference methodology to reason about user preferences in knowledge-based recommender systems. *Knowl.-Based Syst.*, 21(4):305–320, 2008.

[2] R. Burke. Hybrid recommender systems: Survey and experiments. *User Modeling and User-Adapted Interaction*, 12(4):331–370, 2002.

[3] P. Ganesan, H. Garcia-Molina, and J. Widom. Exploiting hierarchical domain structure to compute similarity. *ACM Trans. Inf. Syst.*, 21(1):64–93, 2003.

[4] X. Jin and B. Mobasher. Using semantic similarity to enhance item-based collaborative filtering. In *2nd IASTED International Conference on Information and Knowledge Sharing*, Scottsdale, Arizona, 2003.

[5] R. Lémdani, N. Bennacer, G. Polaillon, and Y. Bourda. A Collaborative and Semantic-based Approach for Recommender Systems. In *Proceedings of the 10th International Conference on Intelligent Systems Design and Applications (ISDA'10)*, pages 469–476, Cairo, Égypt, 2010. IEEE.

[6] R. Lémdani, G. Polaillon, N. Bennacer, and Y. Bourda. A semantic similarity measure for recommender systems. Technical report, SUPELEC Systems Sciences (E3S), 2011.

[7] S. Melnik, H. Garcia-Molina, and E. Rahm. Similarity flooding: A versatile graph matching algorithm and its application to schema matching. In *18th International Conference on Data Engineering (ICDE 2002)*, 2002.

[8] S. E. Middleton, H. Alani, N. R. Shadbolt, and D. C. D. Roure. Exploiting synergy between ontologies and recommender systems. In *Semantic Web Workshop 2002 At the Eleventh International World Wide Web Conference*, 2002.

[9] M. J. Pazzani and D. Billsus. Content-based recommendation systems. In *The Adaptive Web*, volume 4321, pages 325–341. Springer Berlin / Heidelberg, 2007.

[10] P. Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *IJCAI*, pages 448–453, 1995.

[11] W. E. Winkler. The state of record linkage and current research problems. Technical report, Statistical Research Division, U.S. Bureau of the Census, 1999.

[12] Z. Wu and M. Palmer. Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, ACL '94, pages 133–138, Stroudsburg, PA, USA, 1994.