Student Research Abstract: Exploiting the Semantic Similarity of Interests in a Semantic Interest Graph for Social Recommendations

Guangyuan Piao
Insight Centre for Data Analytics, National University of Ireland Galway
IDA Business Park, Lower Dangan, Galway, Ireland
guangyuan.piao@insight-centre.org

ABSTRACT

Social recommendation, a recommender system that targets social media domains, has attracted increasing attention with the growing popularity of Online Social Networks (OSNs). *User Interest Modeling (UIM)* and *Recommendation Algorithm (RA)* as two major components play significant roles in such a system. In recent studies, the *Semantic Interest Graph* (*SIG*), which represents the interests of a user as resources in DBpedia, has shown the efficiency on UIM in OSNs. However, the similarity between resources is not exploited when using *SIG* in RAs. In this work, we propose a novel semantic similarity measure for calculating the similarity between resources. We show preliminary results on the performance of calculating the similarity between general resources as well as resources in a single domain for social recommendations.

1. INTRODUCTION

Social recommendation has attracted increasing attention with the growing popularity of Online Social Networks (OSNs). As a broad definition of social recommendation, it denotes any recommender system that targets social media domains [2]. *User Interest Modeling (UIM)*, as a major component of social recommendations, has been studied based on various approaches such as *Bag of Words*, *Topic Modeling* and the *Semantic Interest Graph* (*SIG*) [3] to represent the interests of a user. *SIG* represents the interests of a user as resources in DBpedia. It can not only provide rich semantic information about an interest (resource), but also can address the problems related to a keyword-based approach such as synonymy and polysemy. On top of the *SIG*, a *Recommendation Algorithm (RA)* has to exploit it in an efficient way to provide good recommendations. However, previous studies on UIM focused on the modeling perspective and the semantic relationships among interests were not taken into consideration when adopting *SIG* into a RA. For example, Abel et. al [1] extracted weighted *SIGs* of users and news items from Twitter, and then used the *cosine similarity* measure for calculating the similarity between a user and an item. In this case, items with the topic `dbpedia:iPad`[1] cannot be recommended if a user is interested in the topic `dbpedia:iPhone`. On the other hand, existing similarity measures such as *Linked Data Semantic Distance* (*LDSD*) [4] for calculating the similarity between resources were designed for resources in the same domain. These measures cannot be applied to social recommendations where user interests can be any general resource without any domain restriction.

2. PROPOSED SIMILARITY MEASURE

We proposed a similarity measure, named ***Resim*** (***Re**source **sim**ilarity*) , which consists of two major components (equation (1)). One is $LDSD_{\gamma sim}$ (equation (2)), which measures the semantic similarity between two resources by incorporating paths. The other one is $Property_{sim}$ (equation (3)), which measures the property similarity of resources when no similarity can be found by using $LDSD_{\gamma sim}$.

$LDSD_{\gamma sim}$ consists of two C_d functions with C_{ii} and C_{io}. C_d is a function that computes the number of direct paths between resources in DBpedia. $C_d(l_i, r_a, r_b)$ equals 1 if there is a property l_i from resource r_a to resource r_b, and 0 if not. The normalizations of C_d functions are carried out using $C_d(l_i)$ that computes the global appearances of a property l_i between any two resources in DBpedia. C_{ii} and C_{io} compute the number of indirect paths between two resources. $C_{ii}(l_i, r_j, r_a, r_b)$ equals 1 if there is a linked resource r_j to both r_a and r_b via an incoming property l_i, and 0 if not. Similarly, $C_{io}(l_i, r_j, r_a, r_b)$ equals 1 if there is a linked resource r_j from both r_a and r_b via an outgoing property l_i, and 0 if not. The normalization of $C_{ii}(l_i, r_j, r_a, r_b)$ (and $C_{io}(l_i, r_j, r_a, r_b)$) is carried out using the global appearances of any two resources connect to the resource r_j with an incoming (outgoing) property l_i in DBpedia.

According to the definition of an ontology, the properties of each concept describe the various features and attributes of

[1]The prefix `dbpedia` stands for `http://dbpedia.org/resource/`

ACM 978-1-4503-3738-0.
http://dx.doi.org/10.1145/0000000.0000000

$$Resim(r_a, r_b) = \begin{cases} 1, & \text{if } URI(r_a) = URI(r_b) \text{ or } r_a \text{ } owl{:}sameAs \text{ } r_b \\ LDSD_{\gamma sim}(r_a, r_b) = 1 - LDSD_{\gamma}(r_a, r_b) & \text{if } LDSD_{\gamma sim}(r_a, r_b) \neq 0 \\ Property_{sim}(r_a, r_b) & \text{otherwise} \end{cases} \quad (1)$$

$$LDSD_{\gamma}(r_a, r_b) = \frac{1}{1 + \sum_i \frac{C_d(l_i, r_a, r_b)}{1+\log(C_d(l_i))} + \sum_i \frac{C_d(l_i, r_b, r_a)}{1+\log(C_d(l_i))} + \sum_i \sum_j \frac{C_{ii}(l_i, r_j, r_a, r_b)}{1+\log(C_{ii}(l_i, r_j))} + \sum_i \sum_j \frac{C_{io}(l_i, r_j, r_a, r_b)}{1+\log(C_{io}(l_i, r_j))}} \quad (2)$$

$$Property_{sim}(r_a, r_b) = \frac{\sum_i \frac{C_{sip}(l_i, r_a, r_b)}{C_d(l_i)}}{C_{ip}(r_a) + C_{ip}(r_b)} + \frac{\sum_i \frac{C_{sop}(l_i, r_a, r_b)}{C_d(l_i)}}{C_{op}(r_a) + C_{op}(r_b)} \quad (3)$$

that concept. Hence, the property similarity is important when a relationship between two resources is not indicated by $LDSD_{\gamma sim}$. $Property_{sim}$ is defined in equation (3). C_{sip} and C_{sop} are functions that compute the number of distinct shared incoming and outgoing properties between resources. $C_{sip}(l_i, r_a, r_b)$ (and $C_{sop}(l_i, r_a, r_b)$) equals 1 if there is an incoming (outgoing) property l_i that exists for both r_a and r_b. C_{ip} and C_{op} compute the number of incoming and outgoing properties respectively of a resource.

3. PRELIMINARY RESULTS

Firstly, we evaluated the performance of calculating similarities on general resources since the *SIG* of a user might contain any topical resource that the user is interested in. For example, the similarity between the two resources `dbpedia:Cat` and `dbpedia:Dog` should be higher than that between `dbpedia:Cat` and `dbpedia:Human`, and a test pair can be created as `sim(dbpedia:Cat, dbpedia:Dog) > sim (dbpedia:Cat, dbpedia:Human)`. In order to get the gold standard test pairs, we use the WordSim353 dataset. WordSim353 is a dataset containing English word pairs along with human-assigned similarity judgements, and is used to train and/or test algorithms implementing semantic similarity measures. We then retrieve the corresponding DBpedia resources and construct test pairs such as `sim(dbpedia:Car, dbpedia:Automobile) > sim(dbpedia:Car, dbpedia:Flight)`. The result shows that 23 out of 28 test pairs can be satisfied using ***Resim*** while 13 pairs can be satisfied using *LDSD* [6].

Secondly, we evaluated the similarity measure in the context of social recommendations where user interests are in a single domain. We use a dataset from the second Linked Open Data-enabled recommender systems challenge. The dataset was collected from Facebook profiles about "liked" items in the music domain that have been mapped to their corresponding DBpedia resources. For each user, 5 liked items were blinded out to construct a candidate list for recommendations while the rest of the items were used to construct the *SIG* of the user. An evaluation based on the F1 score at N (see Table 1) shows that ***Resim*** performs significantly better than *LDSD* ($p < 0.05$) [5].

Table 1: F1 score of recommendations

	F1@5	F1@10	F1@20
Resim	0.0608	0.0642	0.0611
LDSD	0.0484	0.0522	0.0563

4. CONCLUSION AND FUTURE WORK

In this paper, we proposed a measure for calculating the semantic similarity between resources that can represent user interests in *SIG*. Preliminary results have shown that ***Resim*** can not only calculate the similarity between general resources, but also can yield better performance for calculating the similarity between resources in the same domain compared to *LDSD*. In future work, we plan to extend ***Resim*** for calculating the semantic similarity between *SIGs* rather than resources. In addition, we intend to evaluate it on a collected dataset from OSNs, which can extract the *SIGs* of users with various topical interests.

5. ACKNOWLEDGMENTS

This publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289 (Insight Centre for Data Analytics).

6. REFERENCES

[1] F. Abel, Q. Gao, G.-J. Houben, and K. Tao. Analyzing user modeling on twitter for personalized news recommendations. In *User Modeling, Adaption and Personalization*, pages 1–12. Springer, 2011.

[2] I. Guy and D. Carmel. Social recommender systems. In *Proceedings of the 20th international conference companion on World Wide Web*, pages 283–284, 2011.

[3] B. Heitmann. An open framework for multi-source, cross-domain personalisation with semantic interest graphs. In *Proceedings of the sixth ACM conference on Recommender systems*, pages 313–316. ACM, 2012.

[4] A. Passant. Measuring Semantic Distance on Linking Data and Using it for Resources Recommendations. In *AAAI Spring Symposium: Linked Data Meets Artificial Intelligence*, volume 77, page 123, 2010.

[5] G. Piao and J. G. Breslin. Measuring Semantic Distance for Linked Open Data-enabled Recommender Systems. In *The 31st ACM/SIGAPP Symposium on Applied Computing*, 2016.

[6] G. Piao, S. showkat Ara, and J. G. Breslin. Computing the Semantic Similarity of Resources in DBpedia for Recommendation Purposes. In *Semantic Technology*. Springer International Publishing, 2015.