Identifying User Interests from Online Social Networks by Using Semantic Clusters Generated from Linked Data

Han-Gyu Ko, In-Young Ko, Taehun Kim, Dongman Lee, and Soon J. Hyun

Department of Computer Science, Korea Advanced Institute of Science and Technology
291 Daehak-ro, Yuseong-gu, Daejeon, 305-701, Republic of Korea
{kohangyu,iko,kingmbc,dlee,sjhyun}@kaist.ac.kr

Abstract. Recently, online social network services (SNSs) are being spotlighted as a means to understand users' implicit interests out of abundant online social information. Since SNS contents such as message posts and comments are however less informative comparing with news articles and blog posts, it is difficult to identify users' implicit interests by analyzing the topics of the SNS contents of users. In this paper, we propose a semantic cluster based method of combining SNS contents with Linked Data. By traversing and merging relevant concepts, the proposed method expands keywords that are helpful to understand the topic similarity between SNS contents. By using Facebook data, we demonstrate that the proposed method increases the coverage of potential interests by 28.85% and the user satisfaction by 17.24% compared to existing works.

Keywords: User interest identification, Topic analysis, Social network services, Linked Data, Semantic cluster.

1 Introduction

The proliferation of social network services (SNSs) have encouraged many researchers to investigate on understanding users' potential interests from their social contents and relationships [1]. Unlike the interests that are explicitly specified in each user's personal profile, message posts and conversations among users in SNSs need to be processed and analyzed to extract essential information about users' interests. The phenomenon called 'social correlation' is often used to understand implicit interests of users from their social contents [2, 3]. The core of this phenomenon is that SNS users are often influenced by their social neighbors when they make a decision [4]. Therefore, by identifying the correlation between a set of social neighbors' interests and a target user's SNS contents, we could successfully find and recommend potential interests for the user [5, 6]. These approaches mostly use natural language processing methods to extract keywords from SNS contents.

Despites these efforts, users' satisfaction on recommended interests is quite low – less than 65%. This is because there is a low chance of finding the keywords that represent the interests of social neighbors from SNS contents which are short and less informative in comparison to general Web documents such as news articles and blog

Q.Z. Sheng and J. Kjeldskov (Eds.): ICWE 2013 Workshops, LNCS 8295, pp. 302–309, 2013.

posts. Moreover, some keywords have multiple meanings and the existing approaches cannot deal with the semantic heterogeneity problem effectively.

In order to solve these problems, there are essential requirements to be met. Firstly, there must be a way of expanding the set of keywords extracted from SNS contents so that we can increase the chance of finding them from social neighbors' interests. Secondly, to overcome the semantic heterogeneity problem, it is necessary to identify all potential semantics of a keyword, and match them against social neighbors' interests. Since a user's SNS contents may contain keywords that are about latest issues and trends, it is essential to use ontologies and knowledge bases that reflect those while identifying semantics of a keyword.

In this paper, we propose an approach of using Linked Data[1] as the source of finding the appropriate semantics of the keywords that are extracted from SNS contents. We especially use DBpedia, Freebase and OpenCyc as the primary knowledge bases. In our approach, the concepts that are retrieved from these sources are grouped together as *semantic clusters* based on their similarity and relevance. For efficient retrieval and filtering of Linked Data, we design a concept analysis model by which we can explore Linked Data selectively based on subsumption hierarchies and concept similarity. In addition, to find the most essential concepts from which we can start exploring the relevant concepts to generate semantic clusters, we apply the centrality analysis[2] method. A set of semantic clusters are generated from the keywords that are extracted from a user's SNS contents, and another set of semantic clusters are generated from the representative terms that indicate social neighbors' interests. The correlation between these sets of semantic clusters are identified to find social neighbors' interests that might be relevant to the implicit interests of the user.

We conducted an experiment and a user study by using Facebook. The experiment result shows that our approach contributes to increase the coverage of finding potential interests by around 29%. In addition, the user study result proves that the user satisfaction on recommended interests can be improved by around 17% in comparison to the existing approaches.

In the next section, we introduce existing approaches to infer user interests from SNS contents and social network structures. In Section 3, we describe our approach of generating semantic clusters from Linked Data, and recommending potential interests for users. In Section 4, the effectiveness of the proposed method for identifying users' implicit interests from SNS contents is described, followed by the conclusion and future work in Section 5.

2 Related Works

There have been some researches on inferring user interests from users' activities in SNSs. These researches can be categorized into two types. The first category of works measures the correlation among users by using the structural features of social networks such as popularity, similarity, and interaction strength with social neighbors.

[1] http://www.w3.org/standards/semanticweb/data
[2] http://en.wikipedia.org/wiki/Centrality

White et al. [7], Sharma and Cosley [8] are examples. The limitation of these approaches is that they do not consider topics or semantics in social relationships which are essential to accurately infer users' implicit interests.

Another category of works is about analyzing users' SNS contents as well as structural features of social networks. To improve the accuracy of inferring implicit user interests, Zhen et al. [2, 5] and Ahn et al. [6] proposed an approach that combines topic similarity among users' SNS contents with network features such as familiarity with their social neighbors. In these approaches, however, they depend on keyword sets from SNS contents to analyze users' topics, which may be not enough to find the keywords that represent the interests of social neighbors. Hence, their results do not show much improvement in terms of user satisfaction.

3 Identifying User Interests via Semantic Clusters

In this section, we describe a user interest identification method that combines SNS contents with Linked Data. By retrieving concepts that are related to SNS contents and finding and associating more concepts from Linked Data, we can expand the keywords that are extracted from SNS contents and improve the possibility of finding implicit interests of users.

As shown in Figure 1, we handle information from SNSs in two different ways. A *social content* is the combination of a post and its associated comments. An *interest-content* is the content that consists of its name and descriptions. We extract keywords from social contents of a user and interest-contents from his or her social neighbors and retrieve relevant concepts of the topic keywords from Linked Data. We then group semantically relevant concepts together and form semantic clusters. Finally, we identify a list of interest-contents that show high correlation with the user's implicit interests found from their SNS contents.

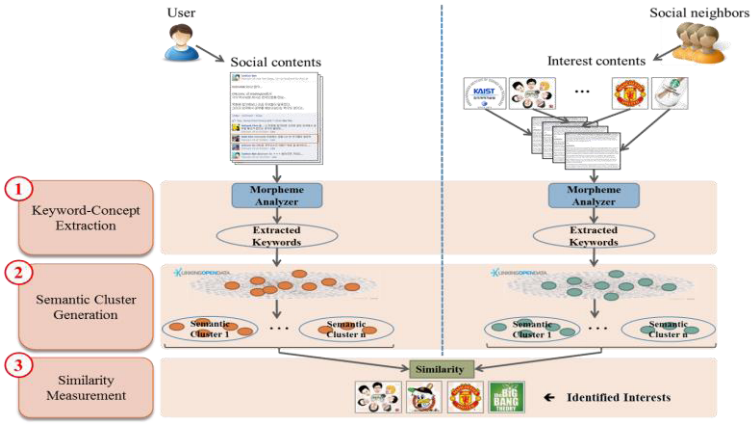

Fig. 1. The process of generating and comparing semantic clusters

The main characteristic of the proposed method is the utilization of semantic clusters rather than using keywords as the source to measure topic similarity between social contents and interest-contents. In the following subsections, we describe the detail steps to generate and analyze semantic clusters from SNS sources and recommend potential interests to users.

3.1 Retrieving Concepts from Linked Data

As the first step to generate semantic clusters, it is necessary to retrieve concepts related to keywords in a user's SNS contents from Linked Data. From Facebook users, we collected all social contents and interest-contents. By using a morpheme analyzer[3], we extracted nouns and noun phrases from those contents as keywords removing stop words and duplicated words.

We then query with each keyword to Linked Data to retrieve the corresponding concept by using SPARQL. We assume that RDF triples which describe the corresponding concept, contain the keyword as one of their property values such as name, title, or label. Because each dataset in Linked Data may use different kinds of predicates to indicate these properties, we handle multiple predicates such as `rdf:label` and `skos:prefLabel` used in target datasets. After that, we query again to retrieve all triples that describe each subject and group them as a concept.

Rather than querying to all the datasets in Linked Data, we selectively query to the datasets that cover various domain knowledge. Because DBpedia is the most central dataset in Linked Data, in which Wikipedia articles are represented as concepts on a large coverage of domains, we use DBpedia and other datasets such as Freebase and OpenCyc, which provide links to DBpedia covering various domains.

3.2 Generating Semantic Clusters

After retrieving all concepts from Linked Data with given keywords, we construct groups of relevant concepts called semantic clusters to filter out irrelevant concepts. We use a concept analysis model shown in Figure 2. Each concept has a label for presentation and a URI for reference. Both properties are literal and connected to the concept through a 'hasKeyword' or 'hadURI' properties, respectively.

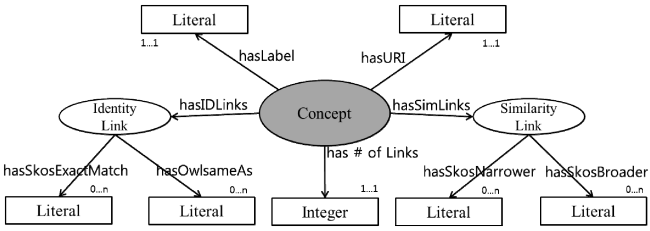

Fig. 2. Concept analysis model

[3] `http://developer.yahoo.com/search/content/V1/termExtraction.html`

Identity-links such as `owl:sameAs` and `skos:exactMatch` are used to consolidate concepts that have the same meaning. We then find and merge semantically relevant concepts by using *similarity-links* such as `skos:broader` and `skos:narrower`. Since those links are used to represent the relative hierarchies of concepts within the same dataset in Linked Data, we need to use similarity-links along with identity-links in a complementary manner to group relevant concepts generated from different datasets. Except for the identity-links and similarity-links which are necessary for analyzing semantic hierarchies among the concepts, others are counted as the number of links.

As the next step, we find and merge more concepts that are semantically relevant to make semantic clusters richer. We apply the breadth-first search method starting from a representative concept for each semantic cluster. The number of hops (the number of edges in the concept graph) from the representative concept is inversely proportional to the semantic relevance. Therefore, it is possible to retrieve concepts that are most semantically relevant by using this method. To select the representative concept, we measure the centrality of concepts and choose the concept that has the highest centrality value as the representative concept in each semantic cluster. We use the PageRank[4] algorithm for measuring the centrality of each concept since it is one of the most effective ways of measuring the relative importance of a node in a set of linked data. We also chose the maximum number of hops as 2 since more than 80% of relevant concepts can be found empirically in the range of 2 hops from the representative concepts [9].

Once the semantic clusters are generated, the ambiguity of a keyword can be resolved by associating it with a set of relevant concepts. For example, the meaning of the keyword "Apple" will become clear when it is associated with the relevant concepts such as "Steve Jobs" and "iPhone".

3.3 Comparing Similarity among Semantic Clusters

By using the semantic clusters generated, we now can measure the similarity between a user's social contents and his or her interest-contents. The interest-contents are from the user's social neighbors in SNS and useful to identify the user's implicit interests.

The similarity measure can be implemented by aggregating the number of overlapped concepts between the semantic clusters of social contents and the interest-contents. When we count the overlapped concepts, each concept's centrality value (C) is considered since it implies how much the concept is important within each semantic cluster.

$$Sim(sc, ic_j) = \sum C(k_{sc\text{-}ic}) \tag{1}$$

where sc is the semantic clusters generated for the social contents, ic_j is a semantic cluster for an interest-content, and $k_{sc\text{-}ic}$ is all concepts that are co-occurred in both sides. After measuring the similarity for interest contents, we generate a list of interest-contents as the user's implicit interests ordered by the similarity value in descending manner.

[4] `http://ilpubs.stanford.edu:8090/422/1/1999-66.pdf`

4 Evaluation

In order to prove the effectiveness of the proposed method, we performed two different evaluations. First, we measured how many interest-contents from each user's social neighbors were found to validate the effectiveness of the semantic clusters. In addition, we performed a user study to validate the user satisfaction on the interests recommended by the proposed method. Both evaluation results were compared against the keyword based method [6] that is considered as the state-of-the-art approach so far.

For our evaluation, we collected 1,043,000 posts and 123,000 interests from 50 Facebook users who have volunteered provide their account data for this research. The volunteers consist of graduate students in the department of computer science at KAIST. We extracted around 10,000 ~ 250,000 keywords for each user by using the morpheme analyzer.

4.1 Finding Potential Interests from Social Neighbors

We measured the coverage of users' potential interests that were found by analyzing the correlation between social-contents and interest-contents. For each user, we calculated the recall value to measure the coverage of the user's potential interests. The recall value is SOC over INT (SOC/INT), where INT is a set of the interest-contents from each user's social neighbors and SOC is a set of social-contents that have overlapped keywords with the interest-contents from each user.

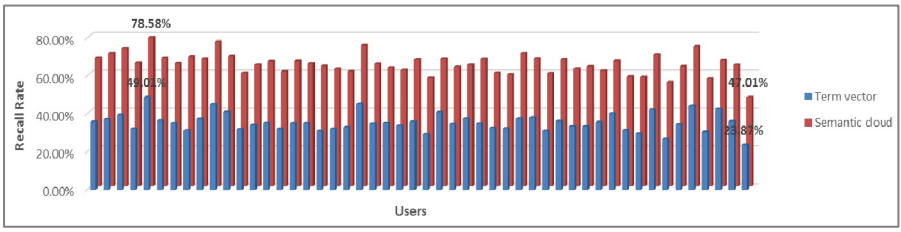

Fig. 3. The coverage of candidate interests

Figure 3 shows the recall rates of the existing approach that uses term vectors and the proposed approach that uses semantic clusters. For all cases, the proposed method shows better results in finding potential interests. The proposed method finds 64.54% of potential interests while the existing approach finds only 35.69% of the interests in average. This result implies that the proposed method is more effective to identify users' implicit interests that can be found by using the semantic similarity measure.

4.2 User Satisfaction on Recommended Contents of Interest

To check if the users actually satisfy with the interests that are recommended by our approach, we conducted a user study. We asked the users to rate the recommended interests in a Likert scale, ranging from 0 to 5. For each participant, we provided two

sets of interests, one is recommended by our approach and another is generated from the existing work [6].

We compared the results in terms of the average rating of the users as shown in Figure 4. For 35 cases out of 50, the proposed method shows better user satisfaction than the existing approach (improves the user satisfaction by 17.24%). In addition, the some results from our approach show much higher ratings than the ones from the existing approach. This is because there are many users who rarely reveal their interests in their Facebook pages.

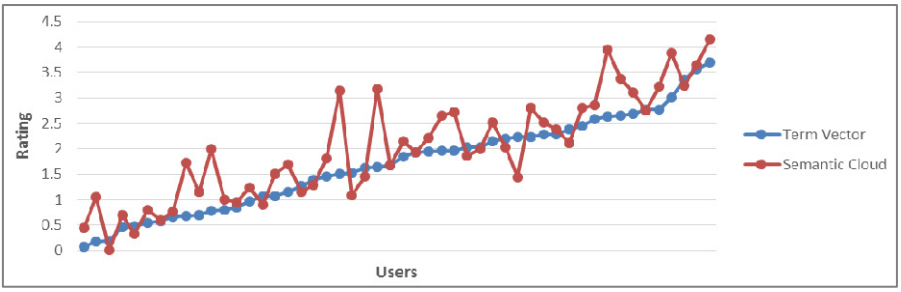

Fig. 4. Ratings on recommended interests (ordered)

4.3 Threats to Validity

Our experiment and user study have the following threats to validity.

1) Limited participants: the number of participants was 50 from a specific area may be insufficient to prove the effectiveness of the propose method to cover various interest domains. However, their social neighbors are not domain-specific people, which may alleviate this threat.
2) Limited datasets: we only used a limited number of datasets such as DBpedia, Freebase and OpenCyc, even there are more than 300 datasets according to the statistics of Linked Data. However, they are a large volume of datasets covering various kinds of interest domains and they are highly interlinked.
3) Outliers in the user study result: an important hypothesis of this research is that users reveal their intentions and interests on their SNS contents. However, there is also a large portion of users who never or rarely do that. This is the reason for the outliers in the user study result.

5 Conclusion and Future Work

In this paper, we proposed a method of identifying and recommending potential interests of a user by analyzing semantically enriched topic keyword sets (semantic clusters) that are generated from the user's SNS contents and social neighbors' interests. In our approach, topic keywords are expended to semantic clusters by associating relevant concepts that are retrieved from Linked Data.

The main contribution of our work are as follows. Firstly, we proposed a framework of utilizing Linked Data to expand topic keywords extracted from SNS contents to semantically enriched sets. Secondly, we developed a method of applying centrality measures and generating semantic clusters to solve the semantic heterogeneity problem of comparing topic keywords. Finally, we proposed a way of comparing semantic clusters to identify the correlation between users' implicit interests extracted from their SNS contents and social neighbors' explicit interests.

In our future research, we will focus on enhancing the quality of semantic cluster generation in terms of scalability, performance, and personalization. In order to access more datasets in Linked Data to cover various topic interests of users, these scalability and performance issues are critical. We are currently working on developing a distributed and iterative approach of accessing and analyzing Linked Data to meet these qualities. In addition, we will produce personalized semantic clusters to improve the users' satisfaction on recommended interests by considering users' preferences.

Acknowledgments. This research was supported by the KCC (Korea Communications Commission), Korea, under the R&D program supervised by the KCA (Korea Communications Agency) (KCA-2013-11913-05005).

References

1. Liu, K., Tang, L.: Large-scale behavioral targeting with a social twist. In: Proceedings of the 20th ACM Conference on Information Knowledge Management, pp. 1815–1824 (2011)
2. Wen, Z., Lin, C.-Y.: On the quality of inferring interests from social neighbors. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 373–382 (2010)
3. Mislove, A., Viswanath, B., Gummadi, K.P., Druschel, P.: You are who you know: inferring user profiles in online social networks. In: Proceedings of the Third ACM International Conference on Web Search and Data Mining, pp. 251–260 (2010)
4. Anagnostopoulos, A., Kumar, R., Mahdian, M.: Influence and correlation in social networks. In Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 7–15 (2008)
5. Wen, Z., Lin, C.-Y.: Improving User Interest Inference from Social Neighbors. In: Proceedings of the 20th ACM Conference on Information Knowledge Management, pp. 1001–1006 (2011)
6. Ahn, D., Kim, T., Hyun, S.J., Lee, D.: Inferring User Interest using Familiarity and Topic Similarity with Social Neighbors in Facebook. In: Proceedings of the 2012 IEEE/WIC/ACM International Conference on Web Intelligence, pp. 196–200 (2012)
7. White, R.W., Bailey, P., Chen, L.: Predicting user interests from contextual information. In: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 363–370 (2009)
8. Sharma, A., Cosley, D.: Network-Centric Recommendation: Personalization with and in Social Networks. In: Proceedings of the 2011 IEEE 3rd International Conference on Social Computing, pp. 282–289 (2011)
9. Mirizzi, R., Ragone, A., Di Noia, T., Di Sciascio, E.: Semantic tag cloud generation via dBpedia. In: Buccafurri, F., Semeraro, G. (eds.) EC-Web 2010. LNBIP, vol. 61, pp. 36–48. Springer, Heidelberg (2010)