Hierarchical Interest Graph from Tweets

Pavan Kapanipathi[1], Prateek Jain[2], Chitra Venkataramani[2], Amit Sheth[1]
[1]Kno.e.sis Center. (pavan, amit)@knoesis.org
[2]IBM TJ Watson Research Center.(jainpr, chitrav)@us.ibm.com

ABSTRACT

Industry and researchers have identified numerous ways to monetize microblogs for personalization and recommendation. A common challenge across these different works is the identification of user interests. Although techniques have been developed to address this challenge, a flexible approach that spans multiple levels of granularity in user interests has not been forthcoming. In this work, we focus on exploiting hierarchical semantics of concepts to infer richer user interests expressed as a *Hierarchical Interest Graph*. To create such graphs, we utilize users' tweets to first ground potential user interests to structured background knowledge such as Wikipedia Category Graph. We then adapt spreading activation theory to assign user interest score to each category in the hierarchy. The *Hierarchical Interest Graph* not only comprises of users' explicitly mentioned interests determined from Twitter, but also their implicit interest categories inferred from the background knowledge source.

Categories and Subject Descriptors

H.0 [**Information Systems**]: General

General Terms

Algorithms, Design, Experimentation

Keywords

Hierarchical Interest Graph; Personalization; Social Semantic Web; Twitter; Wikipedia; User Profiles

1. INTRODUCTION

Twitter has emerged as a prominent medium for people to communicate opinions and interests regarding events and services. Automatically determining these interests from user's topical discussions in tweets involve understanding the content of messages and finding topics and/or entities expressed in them[1]. For example, preponderance posting of messages such as *"Important to note, in current standoff, that Senate-passed plan IS a compromise, accepting lower spending levels passed by House."* indicate that the user is interested in *US Government* and *American Politics*.

Techniques such as Bag of Words and topic models do not perform so well on small, informal text as has been argued in [6]. Also, representing user interests as Bag of Concepts [3, 5] has been experimented by citing the advantages of knowledge-bases. However, exploiting these knowledge sources to generate user interests from tweets is an active area of interest. In this work, we present an approach that recognizes entities from tweets and exploits structured background knowledge to represent user interests as *Hierarchical Interest Graph (HIG)*. The structured background knowledge in our case is the Wikipedia category hierarchy. Wikipedia category hierarchy provides an ability to infer user interests which are not explicitly mentioned in tweets. Consequently, from the above example tweet, it is possible to infer that the user is interested in the *categories* "US Government", and "American Politics" through hierarchical relationships from "US Senate" and "US House of Representatives" that are mentioned via their popular names in the tweet.

The *HIG* generated using our approach extends the existing personalization and recommendation systems by providing flexibility in selecting content with varying level of abstractness. Considering the above example, the *HIG* includes specific interests such as "US Senate" and "US House of Representatives" and also provides flexibility to leverage the semantically inferred broader topics of interests such as "US Government", and "American Politics".

2. APPROACH

The goal of our approach is to construct a *Hierarchical Interest Graph* (HIG) for a Twitter user. The two primary inputs for our system are: (1) Tweets of a user to determine the basic interests of the user, and (2) Hierarchical background knowledge that can be mapped with users' basic interests to infer the HIG. We utilize the *Wikipedia Hierarchy* as the source of structured background knowledge. We opted for Wikipedia because of its vast domain coverage and timely updates.

The system performs the following steps as illustrated in Figure 1: (1) **User Interests Generator** spots and scores the *Wikipedia Entities*[2] from tweets of a user. (2) **Interest Hierarchy Generator** maps the scored *Wikipedia Entities* of interest to the *Wikipedia Hierarchy* to infer the HIG of the user.

User Interests Generator: This module determines the most specific interests of the user from his/her tweets. The process includes:

[1]In this work, we consider users' tweets as a representation of their interests. The common issue is that a user might tweet only about a limited set of his interests.

[2]We differentiate between entities and categories, where entities are the most specific (leaf) nodes in the hierarchy and categories form for the rest of the higher levels (abstract) in the hierarchy.

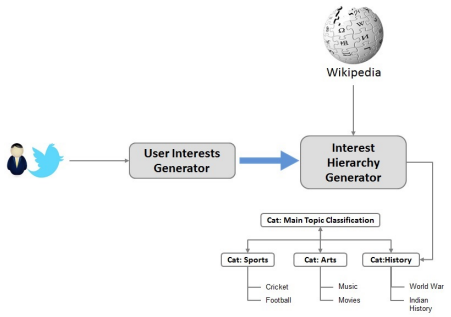

Figure 1: Architecture

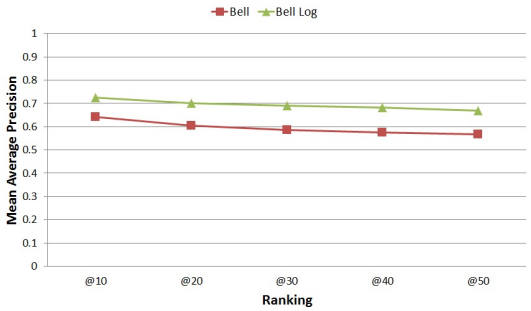

Figure 2: Mean Average Precision of Activation Functions

(1) recognizing Wikipedia entities from users' tweets, and (2) scoring them to represent the extent of users' interests for the entity. Entity recognition from tweets is performed using Zemanta web service[3]. We opted for Zemanta due to its superior performance as evaluated in [2] and their web service's higher rate limit[4]. The scores of the recognized entities is determined based on its normalized frequency using the following equation:
$nf_i = frequency(e_i)/frequency(e_{max})$, where e_{max} is the entity that is mentioned most number of times by the user.

Interest Hierarchy Generator: The final step is to generate the HIG for a user, given the *Wikipedia Hierarchy* and the user's scored specific interests from *User Interest Generator*. In order to accomplish this task, (1) the specific interests are linked to their appropriate categories in the *Wikipedia Hierarchy*, and (2) an adaptation to Spreading Activation theory [1] is used spread the scores of the specific interests to the higher nodes in the hierarchy.

A naive spreading of scores (with empirical decays) up the *Wikipedia hierarchy* infers the *HIG* of the user. However, we discovered that it does not determine appropriate scores for categories in the hierarchy. Analyzing the scores determined by the naive approach, we found that the issues were an impact of the structure of the *Wikipedia Hierarchy* on the spreading activation. Specifically the uneven distribution of nodes at each levels in the hierarchy drastically increases the score up the hierarchy. Therefore, we adapted the spreading activation theory by introducing the parameters below to nullify the impact of the structure:
Normalizers: In order to normalize the scores and hence reduce the propagation of scores up the hierarchy, we utilized the distribution of the nodes in the hierarchy. We experimented with the following two parameters:

Bell: Dividing the value based on the raw number of nodes in the child level.

$$F_i = \frac{1}{nodes_{(h_i+1)}} \quad (1)$$

where $nodes_{(h_i)}$ is the no. of nodes at hierarchical level of node i.

Bell Log: Log value of the distribution, to reduce the impact of high number of categories.

$$FL_i = \frac{1}{\log_{10} nodes_{(h_i+1)}} \quad (2)$$

3. EVALUATION

We experimented the system with three different spreading activation functions that are a combination of the parameters explained in Section 2. (1) Bell (2) Bell Log . In order to evaluate the system, we performed a user study with 37 participants. The participants were provided with the *top-50* categories of interests obtained by employing each of the above activation functions. The *top-50* were determined based on the scores of each category. The participants were asked to mark *Yes/No/Maybe* reflecting the relevance of the interest category. The *Maybe* option was provided due to the abstractness of some categories such as *Category:Technology, Category:Sports* that might not completely be a user's interest. We then calculated the Mean Average Precision [4] for the results obtained from each activation function as shown in Figure 2. We can then conclude that the system employed with *Bell Log* activation function performs better in scoring each category in the *HIG*.

4. CONCLUSION

In this work, we have introduced a novel representation of user interests as a *Hierarchical Interest Graph* and an approach to generate HIG for a twitter user. The approach leverages tweets of users and introduces new parameters to adapt the spreading activation theory, in order to score the interest categories in a user's HIG. The scores represent the user's extent of interest which is demonstrated with an evaluation where the mean average precision for the top ten interest categories is close to 75%.

Acknowledgment

The first author was visiting IBM TJ Watson as a research intern when this work was performed. This work is also supported by NSF (IIS-1111182, 09/01/2011-08/31/2014) SoCS program under the grant titled "Social Media Enhanced Organizational Sensemaking in Emergency Response".

5. REFERENCES

[1] A. M. Collins and E. F. Loftus. A spreading-activation theory of semantic processing. *Psychological review*, 1975.

[2] L. Derczynski, D. Maynard, N. Aswani, and K. Bontcheva. Microblog-genre noise and impact on semantic annotation accuracy. HT '13.

[3] P. Kapanipathi, F. Orlandi, A. P. Sheth, and A. Passant. Personalized filtering of the twitter stream. In *SPIM Workshop at ISWC 2011*.

[4] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[5] F. Orlandi, J. Breslin, and A. Passant. Aggregated, interoperable and multi-domain user profiles for the social web. I-SEMANTICS '12.

[6] B. Sriram, D. Fuhry, E. Demir, H. Ferhatosmanoglu, and M. Demirbas. Short text classification in twitter to improve information filtering. SIGIR '10.

[3] http://developer.zemanta.com/

[4] We thank Zemanta for their support for our research.