1Q-04

Ranking friends' posts on Facebook using hypernyms and tf-idf based on cosine similarity

Siwan Chen[1] Chengguang Shen[2] Ryo Nishide[1] Ian Piumarta[1] Hideyuki Takada[1]

[1]*Graduate school of information science and engineering, Ritsumeikan University*
[2]*School of software, Dalian University of Technology*

1 Introduction

On Facebook, it is possible to miss posts from friends who have similar interests but who are only weakly associated with the user, since posts are ranked based on affinity and type of post but not their content. We are therefore trying to discover friends with similar interests and recommend their posts to the user, connecting the user with weakly-associated friends whose posts' contents reveal similar interests.

Identifying common ground and allowing more opportunities to engage with weak ties can strengthen those ties[1]. One way to detect common interests between weakly-tied friends is to discover the exact interests of a user from their posts and calculate the similarities between that user and their friends. One problem with this approach is that no perfect method exists that can capture the fields of interest to which a word belongs. The existing method, *hypernyms* from WordNet (a large lexical database of English), gives a more general meaning for a word[3] but not the field of interest to which it probably belongs. To detect the exact interests of a user, it is usual to generate manually an *interest-words* pair database. Because this database may be incomplete and/or subjective, we instead use a *word tree structure* within the WordNet database containing 'is-a' and 'is-part-of' relationships[4].

2 Method

We use hypernyms to expose connections between words within a user's posts. A given word can have several hypernyms belonging to different meanings. Since we expect to encounter words related to common topics in a post, we can reject hypernyms that occur for only one word.

We also extract more general relationships from the user's posts, similar to keyword categorization[2]. A user interested in 'music' might also mention 'song', 'piano', and so on, creating is-part-of relationships within the word tree structure. Consequently we compare not only hypernyms but also *holonyms* and *meronyms* within the tree structure. (Y is a holonym of X if X is a part of Y; Y is a meronym of X if Y is a part of X.)

We combine hypernyms from the word tree structure with the original posts' word patterns using tf-idf (short for *term frequency-inverse document frequency*). This results in a *bag* that quantifies the importance of each word in posts by the user and by their friends which, combined with cosine similarity, provides a measure of similarity between the users' posts.

3 Evaluation

To evaluate similarity metrics we compared hypernym similarity with similarities calculated from two other word groups. The first contains all original nouns (ON) appearing more than once in the user's posting history. The second combines the holonyms and meronyms (HM), representing the 'is-part-of' relationship in the word tree structure.

Short test We first calculated these three similarities using fewer than ten posts from seven users and evaluated the results. Inspecting the posts we see that User 1 (U1) mentions having a meal with someone, playing with kids, and famous people in sport. U2 writes about makeup, dress, and a little about tea. U3 writes about sports events, flowers, festivals and parties. Table 1 summarizes the results.

Long test We then compared users based on all their posts. With many thousands of words posted from a user, it is not feasible to inspect the central themes of each post and assess the result. We considered quantifying the result by comparing it with users' 'likes', assuming that these correspond to their interests. We selected five users (U1–U5) whose pages were public, and whose 'likes' were both stable and related to their posts. These users belonged to the categories 'sport', 'fashion', 'photographer', 'chef and travel' and 'kitchen and cooking'. Similarities were calculated between these users and two more users belonging to categories 'athlete' (U6) and 'photographer' (U7). Figure 1 shows the results.

4 Results and discussion

Short test Using tf-idf and cosine similarity, the measured similarity for ON was 0. This is because the posts contain no words in common—a frequent sit-

Table 1: Similarity Tests Using Hypernyms, ON, and HM

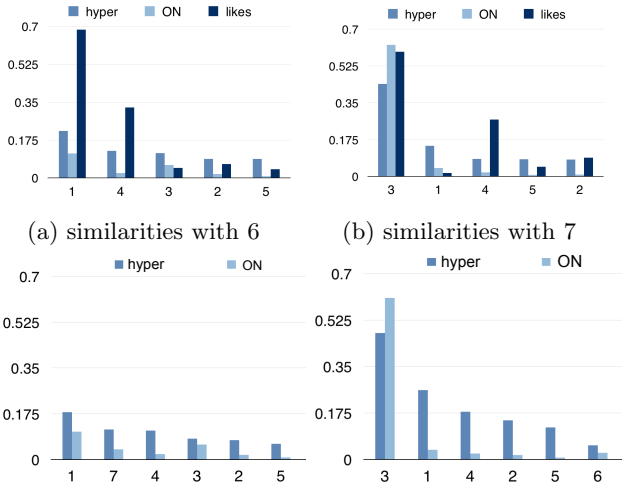
User	Keywords	Hyper	ON	HM
		Compared with User3		
1	meal football sport	0.589	0	0
2	makeup hair dress tea	0.030	0	0
3	flower sport party	-	-	-
		Compared with User5		
2	makeup hair dress tea	0.038	0	0.219
4	pie cake recipe	0.410	0	0
5	dinner food chef	-	-	-
		Compared with User7		
6	tour travel nature	0	0	0
3	flower sport party	0.411	0	0
7	party vacation foreigner	-	-	-

uation when posts are too small or too few in number. Using hypernyms, original words such as 'football' (U1) and 'lacrosse' (U3) belong to the same hypernym meaning 'field_game', and so we obtain a positive similarity even though there are no common original words. Compared with U5, U2 showed a lower similarity than U4, but when compared using holonyms and meronyms the opposite result was obtained.

When comparing fewer than ten posts, hypernym similarity showed a better result than comparing original words or holonyms/meronyms; posts with common themes had higher similarities than those having fewer or no common themes.

Long test When comparing all posts from one user, similarity based on 'likes' and hypernyms indicated the same 'most similar user'. Because U6 and U1 both belong to 'athlete', and U7 and U3 to 'photographer', higher similarities exist between these two pairs.

U6 (athlete) showed the second-highest similarity with U4 (chef) because U6 wrote about traveling, fashion and foods, and U4 travels between France and America. Compared with U6, U1 showed a discrepancy between 'likes' similarity and hypernyms similarity. U1 posts about not just sports but also parties, meals and fashion—which has some shared themes with U6. The 'likes' similarity is exaggerated by their common 'athlete' category because there are only a few hundred such categories, resulting in a coarse-grained, imprecise tree structure compared with the more precise parent-child relationships we extract from WordNet. Another observation is that, when using tf-idf, adding another user to the group does not greatly affect the ranking of similarities for the original group's members. Also, since the tf-idf score is based on all documents in a bag, if one document is highly similar to the document being compared, it will adversely affect the scores of the other documents in the bag; this can be seen in Figure 1d where the high similarity between U3 and U7 reduces the apparent similarity between U6 and U7, which we expect to be higher considering the similarity between U7 and U6 in Figure 1c.



(a) similarities with 6 (b) similarities with 7

(c) similarities with 6 (d) similarities with 7

Figure 1: 'Likes', Hypernyms, and ON Similarities

5 Conclusion and Future Work

Hypernym similarity appears promising for recommending Facebook posts, but further verification is needed. We had planned an experiment using public figures from ten different fields, recommending friends' pages based on similarity of posts and surveying the satisfaction with the results. However, since Facebook changed its policy for the Graph API it is only possible to retrieve the identity of friends who are registered to use the API, making this experiment difficult to perform. We are now considering using 'statuses', 'links', and 'photos' to track contents that users post. Secondly, instead of using only one level of the 'word-tree' to find direct siblings, we are considering improving our recommendations by widening our search to two levels and giving each level a weight. Finally, the users' 'likes' and 'bios' also show a relationship with the contents of their posts, which might be usefully exploited.

References

[1] Catherine Grevet, Loren Terveen, Eric Gilbert, "Managing political differences in social media", Proc. of CSCW, social media and politics, p.1400-1408, (2014)

[2] Prantik Bhattacharyya, Ankush Garg, S. Felix Wu, "Social network model based on keyword categorization", IEEE, 2009 advances in social network analysis and mining, p170-175, (2009)

[3] Paige H. Adams and Craig H. Martell, "Topic detection and extraction in chat", IEEE intl. conference on semantic computing, p581-588, (2008)

[4] WordNet. http://wordnet.princeton.edu/

[5] Prantik Bhattacharyya, Ankush Garg, Shyhtsun Felix Wu, "Analysis of user keyword similarity in online social networks", Social Network Analysis and Mining, p 143-158, (2011)