User Interest Modelling Based on Microblog Data

Miroslav BIMBO*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova, 842 16 Bratislava, Slovakia
bimbo08@student.fiit.stuba.sk

This paper is focused on extracting information about users who are writing the posts, i.e. creating their user model. User model can be used with advantage to recommend or filter content for a particular user, helping him to overcome information overload and allowing him to focus attention on relevant information resources.

In many works the interests of particular user are extracted only from posts written by given user. Interesting way of improving the user model is microblog post enrichment, where interests are extracted from documents, which are only related to original user post. Abel et al. proposed method, where user interests are extracted from news articles, what produced fuller and richer user interest model [1]. Bernstein et al. proposed method for extracting topics of interests leveraging Yahoo search, achieving better results compared to extracting the topic from a post itself [2].

In our work we create the user model based on similar approach as [1, 2]. However, rather than focusing on one enrichment method, we build on the intuition, that we can create better user model by aggregating results from several different enrichment methods.

Proposed method of creating the user model is following:

1. Classification: Classifier compute *interest relevance i* of posts. It is trained by a supervised machine learning algorithm. We create train set (pairs <post, interest relevance of post>) for classifier by manual annotating of posts, and train the classifier using several features of posts.
2. Enrichment: To enrich posts by external documents, we employ several methods: Hashtag (documents are microblog posts, which contain same hashtag as given post), Tagdef (documents are descriptions of hashtag found on Tagdef service[1]), URL (document is text of URL included in given post), News (document is most similar news article, method proposed in [1]), Youtube (documents are descriptions of Youtube videos, retrieved after transformation of posts to queries).

* Supervisor: Marián Šimko, Institute of Informatics and Software Engineering
[1] http://tagdef.com

Spring 2013 PeWe Workshop, April 5, 2013, pp. 45–46.

For some of these methods, we are able to compute *confidence of relation c* between posts and documents.

3. Interest extraction: We extract interests (represented as semantic web entities) from given documents using the OpenCalais web service (web service returns *weight w*).
4. Weighting: We compute importance of each particular interest for a user as:

$$score(user, interest, post, method) = w^{\lambda_1} * c^{\lambda_2} * i^{\lambda_3}, \lambda_i \epsilon \{0,1\}$$

5. Aggregation: In situation, when one interest is found in one post by more methods, or when one interest is found in more posts several aggregation methods, we proposed several aggregation methods.
6. Filtration: We filter out low score interests and mostly repeated repeating false positive interests.

For evaluation of our method, we used UMAP 2011 dataset[2]. We divided posts of one user into 5 equal groups. Four of these groups are used to create a model, last part is test set – viewed as text representation of its posts. Then, we can compute precision (how many of interests from model have its text representation in test set) and recall (how many of words from test set are found in model). We applied 5-fold cross validation, so the final results are averaged from the 5 partial results.

The preliminary results shown, that method aggregating Youtube, News and Tagdef method is more successful according to F1 measure, than baseline method (i.e. no enrichment), what supports our hypothesis.

Our second result is that filtering out all interests found in documents with *confidence c* lower than some threshold can improve results. Furthermore, we found that same filtering based on *weight w* is not useful.

Extended version was published in Proc. of the 9th Student Research Conference in Informatics and Information Technologies (IIT.SRC 2013), STU Bratislava, 119-124.

Acknowledgement. This work was partially supported by the Slovak Research and Development Agency under the contract No. APVV-0208-10.

References

[1] Abel, F., Gao, Q., Houben, G., Tao, K.: Semantic enrichment of twitter posts for user profile construction on the social web. In: *Proc. of the 8th extended semantic web conf. on Thesemanic web: research and applications*, (2011), Springer-Verlag, pp. 375–389.

[2] Bernstein, M. S., Suh, B., Hong, L., et al.: Eddi: interactive topic-based browsing of social status streams. In: *Proc. of the 23nd annual ACM symposium on User interface software and technology*, (2010), pp. 303–312.

[2]http://wis.ewi.tudelft.nl/umap2011/