# Modelling User Interests in Latent Feature Vector Space based on Document Categorisation

Márius ŠAJGALÍK*

*Slovak University of Technology in Bratislava*
*Faculty of Informatics and Information Technologies*
*Ilkovičova 2, 842 16 Bratislava, Slovakia*
marius.sajgalik@stuba.sk

User modelling includes modelling various different characteristics like user goals, interests, knowledge, background and much more [1]. However, evaluation of each of these characteristics can be very difficult, since every user is unique and objective evaluation of each modelled feature often requires huge amount of training data. That requirement cannot be easily satisfied in public research environment, where personal information is too confidential to be publicly accessible. In a common research environment, we are confronted with training the model on only a small sample of data. It mostly requires humans to evaluate the model manually, which is often very subjective and time-consuming.

We liken the problem of personalised keyword extraction problem to document categorisation. We consider users as being different category labels and web pages visited by user as documents labelled by the respective label. By extracting personalised (discriminative) keywords for each document, we can easily infer user interests by aggregating document keywords contained within the web browsing history of the respective user. We examine a novel approach to quantitatively evaluate user interests by formulating an objective function on quality of the model. To accomplish that, we make several assumptions:

- User is interested in all visited web pages in their web browsing history.
- We can represent each such visited web page by a set of extracted keywords to represent user interests that made user to visit that particular page.

Since every user tends to be unique in their interests, we focus on user interests as a discriminative factor between different users. We formulate our objective function to propagate more discriminative (possibly unique) interests, which are supposed to be more informative than the generic ones. Another view on this formulation is that given

* Supervisor: Mária Bieliková, Institute of Informatics and Software Engineering

*Spring 2015 PeWe Workshop, April 11, 2015, pp. 101–102.*

a particular domain of interest, we are interested only in such user interests that can be used effectively to provide the user with personalised content or services.

Our keyword extraction method [3] does not rely on any externally supplied human-labelled semantics. Instead, it uses distributed representation of words, where each word is represented by a vector of latent features that contain semantic and lexical information [2]. Word feature vectors have been already proven to improve results of many NLP tasks. They are assumed to possess an important property that regardless of the task we want to use them for, we can learn them in advance without any prior knowledge of the task. We want to broaden the spectrum of application tasks and propose to use the feature vectors to extract keywords. Our goal is to obtain highly discriminative representation of each document. Our assumption while doing so is that we care only about the descriptive words that discriminate categories or topics within our working domain. We are not interested in generic words that define the domain as a whole, since we know we are working in that particular domain and therefore it does not have any real information value for us. Since the extracted keywords are personalised (discriminative) and thus reflect user interests local to given web page, using the distributed representation we can easily aggregate these words and calculate global user interests.

We also take advantage of having feature vectors for each word and simplify the representation of a document from a list of keywords to a single feature vector. This helps classifiers, since it transforms variable-sized list of keywords into the feature vector of fixed size, so we always have a reasonable number of input features to be used in training. It helps mostly in cases when the number of keywords is very small.

By developing an automated evaluation method, we can assess the quality of user models on much larger scale, since we are no more dependent on leveraging manual assessment. The proposed quantitative evaluation method may have a big impact on user modelling research. By automating evaluation, the development of new methods and models can progress at much higher speed. That means that researchers can make more iterations to improve both their user modelling methods and the user models, which may result in bigger and faster research improvements in the field of user modelling.

*Extended version was published in Proc. of the 11th Student Research Conference in Informatics and Information Technologies (IIT.SRC 2015), STU Bratislava, 52-59.*

*Acknowledgement.* This work was partially supported by the Scientific Grant Agency of Slovak Republic, grant No. VG 1/0752/14.

## References

[1] Brusilovsky, P. Millán, E.: User models for adaptive hypermedia and adaptive educational systems. In: LNCS 4321. Springer, 2007, pp. 3-53.

[2] Mikolov, T., Yih, W., Zweig, G.: Linguistic Regularities in Continuous Space Word Representations. In: NAACL HLT, ACL, 2013, pp. 746-751.

[3] Šajgalík, M., Barla, M., Bieliková, M.: Exploring multidimensional continuous feature space to extract relevant words. In: SLSP, Springer, 2014, pp. 159-170.