Using Site Specificity to Build Better User Model from Web Browsing History

Márius ŠAJGALÍK*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 3, 842 16 Bratislava, Slovakia
sajgalik@fiit.stuba.sk

Extracting user interests from user browsing history has been already studied by several researchers. We present a novel method to enhance the quality of user interest extraction by harnessing the site specificity. We follow the idea that some sites are more generic and not so relevant to real user interests, whereas sites that are more specific are more relevant to the user interests. If there are multiple topics within a single website, it is highly probable that user is just not interested in all of these topics, but chooses to read only a few of them. In order to infer an interest of the user in a site, we propose to calculate the site specificity. The less tightly related topics are contained within a site, the more specific it is and more probable is the higher significance of the discovered topics for user interests. To be able to discover some measurable features, which might influence the overall site specificity, we need some additional knowledge about the web content. However, there is still not enough explicit semantic information of sufficient quality included in the webpage content, which forces us to incorporate some kind of ontology to understand the content of the "wild web" better [1]. Therefore, we use WordNet [3], which can be considered a lightweight ontology.

The basic idea of our approach is to compute the specificity of just a single webpage. To generalise it to an arbitrary set of webpages, we simply concatenate them and calculate the specificity over the union. To compute the website specificity, we choose several subpages within it and concatenate their content into a single piece of text. As we are focusing on enhancing a user model within the web browser, we choose only those subpages of the website, which are present in user's browsing history.

At first, we extract an article and choose only the noun terms as keyword candidates. Then we take all the noun synsets of WordNet, which contain at least one of these feasible terms. We call these synsets the basis synsets. Then we create the concept graph $G = (V, E)$, where vertices V are all the basis synsets plus those reachable by following hypernym or holonym relations. This aims to influence also the

* Supervisors: Michal Barla, Mária Bieliková, Institute of Informatics and Software Engineering

Spring 2013 PeWe Workshop, April 5, 2013, pp. 63–64.

more general concepts (WordNet synsets) to get to the broader topics discussed in the extracted article. After we have built this concept graph, we perform page ranking algorithm to infer the relevance of individual concepts inspired by [4]. We do a two-pass ranking. In the first pass, we propagate the authority of a synset to all hypernyms and holonyms to obtain the most probable word senses. Apart from [4], we consider also the information content of single concepts. Additionally, we consider collocations and link the neighbouring terms in the second-pass page rank to support the collocated word senses and thus, get the key concepts. We adapted this idea from TextRank [2].

To calculate the site specificity, we applied various measures of concept similarity to measure the topic diversity or more specifically, the semantic coherence of the topmost concepts contained in the concept graph based on [5]. We considered only the concepts ranked at the top after inferring the ranking algorithm, since those are the most probable to be the most relevant representatives of the covered topics.

We evaluated four possible measures to measure the semantic coherence; however, there was no single winner method. We believe we could enhance it further by constructing a probabilistic model (Bayesian network) based on different features of the constructed concept graph. We could train the model parameters corresponding to different features using some dataset of categorised websites. Using such model, we would be able to set the values of observed variables and do the inference to obtain the conditional probability of the website being specific. We also plan to use the results presented in this paper to devise another method, which we believe will build a better user model having taken the website specificity into account.

Extended version was published in Proc. of the 9th Student Research Conference in Informatics and Information Technologies (IIT.SRC 2013), STU Bratislava, 186-193.

Acknowledgement. This work was partially supported by the Scientific Grant Agency of Slovak Republic, grant No. VG1/0675/11.

References

[1] Bieliková, M., Barla, M., Šimko, M.: Lightweight Semantics for the "Wild Web". Keynote. In: *WWW/Internet 2011, Proc. of the IADIS Int. Conf.*, IADIS Press, (2011), pp. xxv-xxxii.

[2] Mihalcea, R., Tarau, P.: TextRank: Bringing Order into Texts. In *Conf. on Empirical Methods in Natural Language Processing* (2004), pp. 404-411.

[3] Miller, G. A.: WordNet: A Lexical Database for English. *Communications of the ACM*, Vol. 38, No. 11, (1995), pp. 39-41.

[4] Ramakrishnanan, G., Bhattacharyya, P.: Text Representation with WordNet Synsets Using Soft Sense Disambiguation. *Ingénierie des systèmes d'information*, vol. 8, (2003), pp. 55-70.

[5] Resnik, P.: Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In: *Proc. of the 14th int. joint conf. on Artificial intelligence (IJCAI'95)*, Chris S. Mellish (Ed.), Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, vol. 1, (1995), pp. 448-453.