Normal Distribution Re-Weighting (NDRW) for Personalized Web Search

Hanze Liu and Orland Hoeber*

Department of Computer Science,
Memorial University
St John's, N.L, Canada
{hl5458,hoeber}@mun.ca

Abstract. Personalized Web search systems have been developed to tailor Web search to users' needs based on their interests and preferences. A novel Normal Distribution Re-Weighting (NDRW) approach is proposed in this paper, which identifies and re-weights significant terms in vector-based personalization models in order to improve the personalization process. Machine learning approaches will be used to train the algorithm and discover optimal settings for the NDRW parameters. Correlating these parameters to features of the personalization model will allow this re-weighting process to become automatic.

1 Introduction

Web search is an essential tool for today's Web users. Web search systems, such as Google, Yahoo! and Bing have been introduced to the public users and achieved great success. However, traditional search engines share a fundamental problem: they commonly return the same search results to different users under the same query, ignoring the individual search interests and preferences between users. This problem has hindered conventional Web search engines in their efforts to provide accurate search results to the users. To address the problem, personalized Web search has been introduced as a way to learn the individual search interests and preferences of users, and use this information to tailor the Web search to meet each user's specific information needs [6].

Personalized Web search employs personalization models to capture and represent users' interests and preferences, which are usually stored in the form of term vectors (see [2][6] for a review of vector-based models for personalized search). High-dimensional vectors are used to represent each user's interest in specific terms that might be present in the search results. These vectors are then used to provide a personalized re-ranking of the search results. In this research, we focus on improving personalized Web search through refining such vector-based personalization models.

The goal in our research is to develop methods to automatically identify and re-weight the significant terms in the target model. This approach is inspired

* M.Sc. Supervisor.

by Luhn's seminal work in automatic text processing [5], in which he suggests that the "resolving power" of significant terms follows a normal distribution placed over a term list ranked by the frequency of term occurrence. In other words, Luhn suggests that the mid frequency terms are more content bearing than either common terms or rare terms, and so are better indicators for the subject of the text. This idea has been widely utilized in the fields of automatic text summarization [8] and Web search converge testing [1]. However, to the best of our knowledge, it has not been explored in the literature of personalized Web search. In the following sections, we will demonstrate how we could borrow Luhn's idea to improve the vector-based models used in personalized Web search.

2 Normal Distribution Re-Weighting (NDRW)

The first step in the NDRW approach is to rank the terms in the vector-based personalization model according to their frequency, resulting in a term histogram as illustrated in Fig. 1. Luhn's suggestion is that high frequency and low frequency terms are not valuable. By placing a normal distribution curve over top of the term histogram, we can assign a significance value to each term, reducing the weight of the terms near the two ends, and giving more weight to the valuable terms in the middle range.

To calculate the term significance (TS) value for each term, we employ the following formula:

$$TS(i) = normdist(s * r_i) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(s*r_i-\mu)^2/2\sigma^2} \tag{1}$$

where r_i is the rank of a given term i, and s is a predetermined step size between any two adjacent terms along the x-axis. There are three parameters in this function that affect the shape of the normal distribution curve, and therefore the TS value for a given term. μ is the mean of the distribution; it decides the location of the centre of the normal distribution curve. σ^2 is the variance of the distribution; it describes how concentrated the distribution curve is around the mean. The step size s affects the steepness of the distribution curve given a

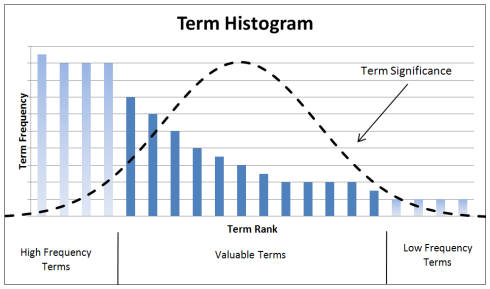

Fig. 1. NDRW re-weights the terms using a normal distribution curve.

constant variance. Once appropriate parameters are chosen for μ, σ^2 and s which specify the location and shape of the normal distribution curve, TS values can be calculated for each term and used to re-weight the personalization model.

miSearch [3] is an existing vector-based personalized Web search system that is used as the baseline system in this research. The novel feature of this personalization system is that it maintains multiple topic profiles for each user to avoid the noise which normally exists within single-profile personalization models. The topic profiles in miSearch are term vectors in which terms are extracted from the clicked result documents and weighted by term frequency. We have implemented the NDRW approach within this system to re-weight the terms in the topic profiles, and have been able to improve the accuracy of the ranked search results list by carefully choosing the NDRW parameters. However, an important part of this research is to automatically determine these parameters based on features within the target vector-based model. The process by which we plan to achieve this is outlined in the remainder of this paper.

3 Automatic Algorithm for NDRW

In order to develop the automatic algorithm for NDRW, we plan to employ a supervised machine learning scheme. There are three main steps in this plan: preparing the training data and test data for the learning process, defining the evaluation metrics to guide the learning, and training the optimum parameters and the algorithm.

Twelve queries were selected from the TREC 2005 Hard Track [7] for previous evaluations on the baseline miSearch system [3]. We will continue to use this test collection as the training data for our experiments. These queries were intentionally chosen because of their ambiguity. For each query, 50 search results have been collected and judged for relevance. The value of the personalization approach will be decided based on whether the relevant documents can be moved to the top of the search results list. For the test data, we will select another 12 ambiguous queries from this test collection and provide relevance judgements on the documents retrieved.

We will use average precision (AP) measured over the top-10 and top-20 documents as the evaluation metric. In order to facilitate the experiments, a test program will be implemented to automatically apply NDRW to the target personalization models with associated test queries, and directly output the resulting AP values, given a set of NDRW parameters.

To train the optimum parameters for each set of training data, Particle Swarm Optimization (PSO) [4] will be employed. The test program mentioned above will play the role of the fitness function in the PSO. The fitness value will be calculated by 60% of the top-10 AP value plus 40% of the top-20 AP value. Each particle contains three parameters (μ, σ^2 and s), and the optimum parameters are achieved when particles converge to the global best fitness value for a given set of training data.

After gathering the optimum parameters for each set of training data, it may be possible to discover relationships between the optimum parameters and the features within the corresponding personalization models. Furthermore, by analyzing these relationships, we may be able to establish general rules for choosing the NDRW parameters.

With the established rules, the algorithm for automatically choosing the parameters can be constructed. We can then verify its quality by applying it to the test data, measuring the degree to which the AP is improved and how close the parameters are to the optimal parameters for each test query.

4 Conclusion and Future Work

In order to improve personalized Web search, we proposed a novel Normal Distribution Re-Weighting (NDRW) approach to identify and re-weight significant terms in vector-based personalization models. Currently, we are working on the main task of this research, which is to develop an automatic algorithm for choosing NDRW parameters based on the features of the target model. In the future, we plan to conduct user evaluations to measure the benefit of using the NDRW technique for improving personalized Web search in realistic settings.

References

1. Dasdan, A., D'Alberto, P., Kolay, S. and Drome, C.: Automatic retrieval of similar content using search engine query interface. In: Proceedings of the ACM Conference on Information and Knowledge Management, pp. 701-710. (2009)
2. Gauch, S., Speretta, M., Chandramouli, A. and Micarelli, A.: User profiles for personalized information access, In: Brusilovsky, P., Kobsa, A., Nejdl, W. (eds.). The Adaptive Web: Methods and Strategies of Web Personalization. Springer-Verlag, Berlin Heidelberg New York. pp.54-89. (2007)
3. Hoeber, O. and Massie, C.: Automatic topic learning for personalized re-ordering of Web search results. In: Proceedings of the Atlantic Web Intelligence Conference, pp.105-116. (2009)
4. Kennedy, J. and Eberhart, R.: Particle Swarm Optimization. In: Proceedings of IEEE International Conference on Neural Networks. IV. pp. 1942-1948. (1995)
5. Luhn, H.P.: The automatic creation of literature abstracts. IBM Journal of Research and Development, 2, pp. 159-165. (1958)
6. Micarelli, A., Gasparetti, F., Sciarrone, F., and Gauch S.: Personalized search on the World Wide Web. In: Brusilovsky, P., Kobsa, A., Nejdl, W. (eds.). The Adaptive Web: Methods and Strategies of Web Personalization. Springer-Verlag, Berlin Heidelberg New York. pp. 195-230. (2007)
7. National Institute of Standards and Technology. TREC 2005 Hard Track. `http://trec.nist.gov/data/t14_hard.html`
8. Shen, D., Chen, Z., Yang, Q., Zeng, H., Zhang, B., Lu, Y., and Ma, W.: Webpage classification through summarization. In: Proceedings of the International ACM/SIGIR Conference on Research and Development in Information Retrieval, pp. 242-249. (2004)