A Luhn-Inspired Vector Re-Weighting Approach for Improving Personalized Web Search

Hanze Liu
Department of Computer Science
Memorial University of Newfoundland
St John's, NL A1B 3X5 Canada
hl5458@mun.ca

Orland Hoeber
Department of Computer Science
Memorial University of Newfoundland
St John's, NL A1B 3X5 Canada
hoeber@mun.ca

***Abstract*—A fundamental problem with current Web search technology is that in the absence of any additional information, the same query provided by two different searchers will produce the same set of search results, even if the information needs of the searchers are different. Web search personalization has been proposed as a solution to this problem, whereby the interests and preferences of individual users are modelled and used to affect the outcomes of their subsequent searches. A common approach is to generate vector-based models of searchers' interests, and re-rank the search results based on the similarity of the documents to these models. In this paper, a novel approach is proposed to automatically identify and re-weight significant dimensions in vector-based models in order to improve the personalized order of Web search results. This approach is inspired by Luhn's model of term importance, which is rooted in Zipf's Laws. Evaluations with a set of ambiguous queries illustrate the effectiveness of this approach.**

***Keywords*-search personalization; vector-based personalization models; automatic vector re-weighting.**

I. Introduction

Web search has become an essential tool for people to find information among the vast resources available on the Web. A typical search engine asks users to input a query, and returns a ranked list of documents. However, a fundamental problem with Web search is that in the absence of any additional information, the same query provided by two different searchers will produce the same set of search results, even if the information needs of the searchers are different. For example, two searchers entering the query "piracy" may be seeking very different documents (robberies at sea vs. illegal software copying). For each of these searchers, the unwanted search results may be relevant to the query, but irrelevant to the search intents of the particular user.

Personalized Web search has been proposed to address this problem by providing personalized search results for each user based on their varied information needs. A common approach for personalized Web search is to model a user's interests and preferences within a vector representation [1], [2]. Each dimension of these vectors represents a term (or stem [3]) and the value along a given dimension is commonly the term frequency (TF) found within the information used to generate the personalization vector. In order to avoid having these vectors become too bloated with irrelevant information, stop word removal is often employed, whereby common terms that have little value for differentiating between good and bad documents are ignored (e.g., "a", "the", "it").

These high-dimensional vectors can then used to re-rank the search results based on their similarity to the documents. The assumption here is that if a set of terms were used frequently in the information used to generate the personalization vector, and are used commonly in a given document in the search results list, then that document may be important for the individual searcher. Unfortunately, this is not necessarily the case. That is, the frequency of a term may not be a good indicator of the value of that term. Classical information retrieval has addressed this problem through the use of TF-IDF [4] and other related measures. However, the calculation of the inverse document frequency (IDF) is not always feasible within the context of personalized Web search since it requires knowledge of the distribution of terms across all documents in the collection.

Other classical work in the field of information retrieval may be useful for improving the TF approach to personalized Web search. One such work is that by Luhn [5], in which it was suggested that mid-frequency terms are the most useful terms for representing textual information, rather than high-frequency terms. He suggested that the significance of a term follows a normal distribution placed over the terms, when they are ranked according to their frequency.

Inspired by Luhn's work, we propose a novel approach to automatically identify and re-weight significant terms in a vector-based personalization model. The primary contribution of this research is the application of Luhn's ideas to the domain of personalized Web search, and the development of an automatic algorithm for determining both the location and variance of the normal distribution that produces the re-weighted vector. An evaluation of this approach using a set of ambiguous queries shows that it can indeed improve the order of the search results provided that the personalization vector is sufficiently well-trained and robust.

II. Related Work

Personalized Web search is a rather active field of research, with varied directions [2]. Regardless of the specific details of how the personalization approaches operate, all personalization methods require some technique for modelling searchers' interests.

Ahn et al. [6] proposed a profile-based personalized system for task-based information exploration. This system allowed users to select and save fragments of Web pages as notes while they explored information resources. Based on the top 300 important terms judged using TF-IDF, a vector-based profile for each user was created from their notes and used to re-rank search results. Dou et al. [7] proposed three profile-based approaches to offer personalized re-ranking based on different lengths of the search history. The profiles employed in all three approaches were automatically generated from the clicked Web pages in the search history, and used to calculate a personalized score for each document in the search results. The search results were then re-ranked according to their personalized scores. The user profile in the work of Sugiyama et al. [8] was generated from the user's entire Web browsing history. A weighting scheme based on TF was employed to construct the user profiles. The similarity between the profile vector and each document vector was calculated using the cosine metric, and the search results were reordered based on their similarities to the profile. The authors also presented a collaborative filtering algorithm in order to enhance user profiles that contained little information.

A difficulty with the above approaches for personalized search is that most of them create a single personalization profile (vector) that is meant to capture all of the interests of the user. For a given information need, such a personalization profile will invariably include noise (i.e., terms that a user is interested in for one context, but not for other contexts). The noise in the profile may cause some of search results which are irrelevant to the user's current search interests to be promoted in the search results list, and reduce the effectiveness of the personalization.

miSearch [9] addresses this problem by allowing searchers to create distinct search topic profiles. This system automatically learns the searchers' interests based on search results that were viewed in the course of previous searches on the topic. As search results are selected, the personalization vector representing the topic profile is updated using a TF approach. Similar to the other approaches discussed above, the search results are re-ordered based on their similarity to the topic profile. The multiple topic profiles capture the user's different information needs separately, enabling the system to offer more precise re-ranking of the search results for the user's current search goal.

Because miSearch uses TF in the generation of the vector-based models for the topic profiles, it may suffer from an over-weighting of the high frequency terms. Although stop word removal is employed to address this problem, it may be possible to further improve the system performance. In this research, we employ miSearch as the baseline personalization system, apply our Luhn-inspired vector re-weighting approach, and measure its potential benefit.

III. Luhn-Inspired Vector Re-Weighting

A. Inspiration from Luhn

In his 1958 paper on the automatic creation of document abstracts, Luhn suggested that the frequency of word occurrence in a document provides a useful measure of word significance, and that the "resolving power" of significant terms follows a normal distribution placed over a word list that is ranked by the frequency of occurrence (i.e., a TF histogram) [5]. In other words, Luhn's model suggests that the mid-frequency terms are the best indicators for the subject of the text, and that common terms and rare terms are less valuable. The theoretical foundation for Luhn's model is provided by Zipf's Laws [10], which suggest that there is a power law relationship between the frequency of word occurrences and the rank of terms in a frequency table.

Luhn's model is seminal work in the field of automatic text processing, and forms the basis for many later works on automatic text analysis [11], automatic text summarization [12], and Web search coverage testing [13]. However, to the best of our knowledge, it has not been explored in the context of personalized Web search or vector re-weighting. In previous work [14], we explored the use of machine learning to optimize the parameters when using a normal distribution to re-weight personalization vectors. In the sections that follow, we explain how the parameters for applying Luhn's model can be determined based on the features of the personalization vector itself.

B. Vector Re-Weighting

Our application of Luhn-inspired vector re-weighting is based on re-weighting the topic profiles within our baseline personalization system. The first step in this approach is to rank the terms in the vector-based personalization models according to their frequency, resulting in a TF histogram as illustrated in Fig. 1. In this histogram, the terms near the left end are high-frequency terms, which usually are too common to be significant. Similarly, the terms near the right end are low-frequency terms, which are too rare to be significant (and can be considered noise). The valuable terms are located in the middle range.

Once the TF histogram is established, a normal distribution curve can then be placed on the top of the histogram to demonstrate the "resolving power of significant words" [5]. Luhn uses this phrase to refer to the ability of words to discriminate content: the greater the resolving power, the better the word can represent the characteristics of the content. As will be explained in the sections that follow,

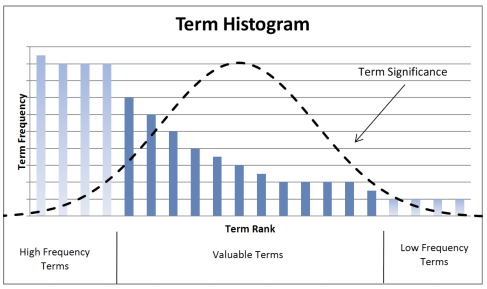


Figure 1. Luhn's model of term importance.

the primary challenge in applying Luhn's model is the development of an automatic algorithm to determine the location and variance of the normal distribution based on features of the personalization vector and its associated TF histogram.

C. Approach Formalization

More formally, the goal of the Luhn-inspired vector re-weighting is to replace the TF value in the personalization vector with a term significance (TS) value based on the normal distribution placed over the TF histogram. To calculate the TS value for each term in the vector, we use the probability density function:

$$TS(i) = f(r_i; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(r_i-\mu)^2/2\sigma^2} \qquad (1)$$

where r_i is the rank of the given term i in the TF histogram, and $TS(i)$ is the significance value of that term i. Equation 1 contains two parameters that affect the location and the shape of the normal distribution, and therefore the TS value for a given term: the mean value μ and the variance σ^2. The mean value μ decides the location of the centre of the normal distribution, and the variance σ^2 describes how concentrated the distribution is around the mean. How these parameters are chosen is the fundamental challenge of making this approach an automatic method, and will be discussed in detail later in this paper.

Once the appropriate values are determined for μ and σ^2, the TS values for the terms in a given vector-based personalization model can simply be calculated using (1). A re-weighted vector can then be created by replacing the frequency of term i with $TS(i)$. An alternate approach is to multiply the term frequency by the term significance, resulting in a $TF * TS$ approach for re-weighting. We will compare and discuss these two different re-weighting approaches in the evaluation section of this paper.

D. Automatic Parameter Selection

A critical aspect of this research is the use of the features of the term vector and associated TF histogram to choose the location and shape of the normal distribution that generates the TS values. Our goal is to be able to automatically generate appropriate mean (μ) and variance (σ^2) values by analyzing the features of the source vector-based model. We discuss the issues of choosing these two parameters separately.

1) Determining the Location: In order to choose the mean value for the normal distribution used in the vector re-weighting process, we find that Zipf's Laws, and more importantly, Goffman's theory regarding the transition region [15], provides valuable information for our purposes. Consistent with Luhn's model, Goffman pointed out that there is a transition region between Zipf's First Law of high-frequency words [10] and Booth's revision of Zipf's Second Law of low-frequency words [16]. It is at this transition region that the most content-bearing words of a given text occur.

Pao formalized this theory into a simple equation [15]:

$$n = (-1 + \sqrt{1 + I_1})/2; \qquad (2)$$

where n is the frequency of the word that is located at the centre of Goffman's transition region, and I_1 is the number of the words that only appear once in the target text. Using this equation, one can easily identify the words around Goffman's transition region, which are considered the terms that have the highest resolving power.

We use Pao's equation for determining Goffman's transition region to decide the mean value for our approach. Given a topic profile, we count the number of terms that occur only once within the topic profile vector, and use (2) to calculate the frequency value n. The term in the profile that has the nearest frequency value to n (called "mean term") is selected as the centre of the transition region, and its rank is used as the mean value.

2) Determining the Shape: The shape of the normal distribution is controlled by the variance parameter σ^2, which is the square of the standard deviation σ of the distribution. Increasing σ makes distribution flat and broad; decreasing σ makes the distribution steep and narrow.

Whether it is better to have a flat and broad or a steep and narrow normal distribution depends upon the features of the histogram near the mean term. If there are many other terms nearby that have similar values, then a high σ value is desirable since it will flatten the distribution to include these terms that have a similar frequency. On the other hand, if the terms near the mean term have very different TF values, then it may be better to have a low σ so that the distribution is narrowly focused on the mean term.

The slope s of the histogram at the mean term can be estimated numerically using the 5-point central difference formula. If the mean term is located at the second or first terms in the TF histogram, then the 3-point and 2-point central difference formulas are used, respectively. However, there is a difficulty with using this estimated slope within a finite calculation: it has an unlimited range of values. Since ordered histograms are always monotonically decreasing, the slope s will be in the range $(0, \infty)$. We address this

problem by calculating the angle of the secant line and use this angle to determine the standard deviation. The angle θ can be calculated from the slope s using the $arctan$ function, resulting in a value that is in the range $(0, \pi/2)$, measured in radians.

Using this angle θ, we calculate the standard deviation σ using the following formula:

$$\sigma = a + b/\theta \tag{3}$$

where a and b are two tuning parameters that can be used to control the minimum value and the rate at which it changes as a result of a change in θ. A large θ indicates a steep slope, resulting in a low standard derivation that produces a narrow normal distribution around the mean term. A small θ indicates a flat slope, resulting in a high standard deviation that produces a broad distribution around the mean term.

IV. Evaluation

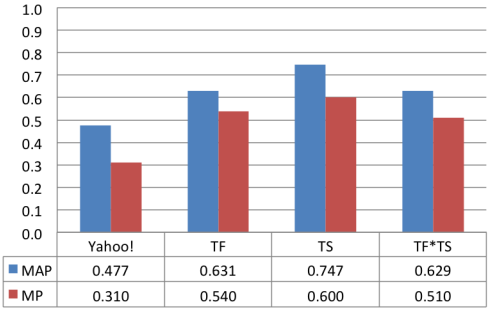
A study was conducted to evaluate the effectiveness of the Luhn-inspired vector re-weighting approach. With the automatic algorithm for parameter selection, the system is able to choose parameters for the normal distribution and re-weight the term vectors based on the features of the TF histogram. We wish to determine whether the approach can indeed result in an improvement over the original order of the search results and the order produced by the existing baseline TF approach.

A. Evaluation Methodology

Ten ambiguous topics were selected from the TREC 2005 Hard Track [17] for use in these experiments. For each query, 50 search results were retrieved from Yahoo! and assigned relevance scores by a panel of reviewers resulting in ground truth relevance. Since the queries were ambiguous in nature, each search result list contained a mixture of relevant and irrelevant documents. The effectiveness of the personalization approach can be judged based on whether the relevant documents can be identified and moved near the top of the list after the re-ranking process.

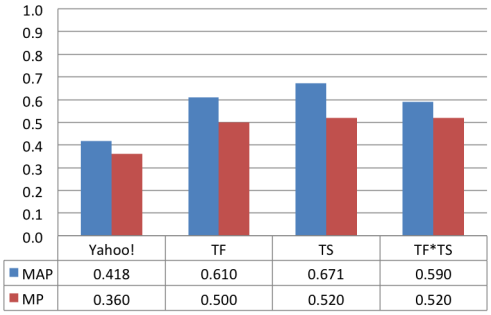
Both precision and average precision [18] measured over the top 10 and top 20 documents are used as the primary evaluation metrics (i.e, P-10, P-20, AP-10 and AP-20). Precision is the ratio of relevant documents to the total documents retrieved. Average precision provides a score that not only takes into account the relevance of the documents, but also their placement within an ordered list. This metric provides a measure of the quality of the ranked search results list, indicating the extent to which the relevant documents are placed in the high positions in the list.

To start, separate topic profiles were created in the miSearch system for each of the test topics. In the initial stage, these ten topic profiles were all empty, and the ranking order of the search results under each test query reflected the Yahoo! rank without any personalization. In the second step, we conducted five searches under each topic profile, using queries that were derived from the test topic, but different than the test query. The goal here was to mimic a searcher's past interest and search activity in a topic. In each of these searches, we clicked on the first five relevant documents to update the topic profile, and then used the updated profile to re-rank the search results under the test query.



	Yahoo!	TF	TS	TF*TS
MAP	0.477	0.631	0.747	0.629
MP	0.310	0.540	0.600	0.510

(a) Top 10 search results



	Yahoo!	TF	TS	TF*TS
MAP	0.418	0.610	0.671	0.590
MP	0.360	0.500	0.520	0.520

(b) Top 20 search results

Figure 2. Mean average precision (MAP) and mean precision (MP) scores over the top 10 and top 20 search results.

Three different re-ranking algorithms are evaluated in these experiments. The TF approach represents the method employed in the original miSearch system. The TS approach is a result of performing the Luhn-inspired vector re-weighting method proposed in this paper. As an alternative to replacing the TF vector values with TS values, the third approach under investigation scaled the TF values by the TS factor, resulting in TF*TS. For the TS and TF*TS approaches, we set the tuning parameters in (3) to $a = 0.1$ and $b = 5.7$. These parameters were selected based on experimentation with the approach.

B. Evaluation Results

The general theme that emerged as a result of these experiments is that the baseline personalization system (TF) improved upon the original order of the Yahoo! search results; the Luhn-inspired work described in this paper (TS) showed even further improvement. However, the performance of the TF*TS approach was not very different from the TF approach. The mean values for AP and P across all of the ten test topics are illustrated in Fig. 2.

Viewing the mean results for the TS approach in more detail, we can see that it achieved the highest values in all of the four evaluation metrics. Since the improvements of the AP were simultaneously supported by improvements on P, not only was the order of the search results improved when viewing the top 10 and top 20 search results, but more relevant documents from the remainder of the set were moved to prominent locations in the search results list. Therefore, we conclude that the TS approach produced the best overall results among the four approaches.

In tuning the parameters for the shape of the normal distribution (i.e., a and b), we were focused on improving the AP metric on the TS approach. Therefore, it is not surprising that the TS approach outperformed the TF*TS approach in general. However, by analyzing the performance data for each of the test topics individually, we observed that in two cases the TF*TS approach performed best among the four approaches. This shows the promise of the TF*TS approach, warranting further study.

V. Conclusion and Future Work

A novel Luhn-inspired vector re-weighting approach for improving vector-based personalized Web search has been proposed and studied in this paper. This method employs a normal distribution to identify and re-weight terms within personalization vectors. An algorithm was developed to automatically determine the location and the shape of the normal distribution based on features of a TF histogram generated from the source vector. The study showed that this approach can result in improved precision and average precision scores.

Further study is currently underway to evaluate the approach over a larger collection of queries, as well as to study the potential benefits of the TF*TS approach. The exploration of other possible applications of the proposed approach are also under consideration. Although implemented within the content-based multiple-profile framework of miSearch, Luhn-inspired vector re-weighting could be applied to any personalization method that employs a TF-based vector modelling of information, including collaborative-based personalization frameworks. This approach may also be helpful to improve TF-based models used in other fields beyond the scope of personalized Web search.

References

[1] S. Gauch, M. Speretta, A. Chandramouli, and A. Micarelli, "User profiles for personalized information access," in *The Adaptive Web: Methods and Strategies of Web Personalization*, P. Brusilovsky, A. Kobsa, and W. Nejdl, Eds. Springer-Verlag, 2007, pp. 54–89.

[2] A. Micarelli, F. Gasparetti, F. Sciarrone, and S. Gauch, "Personalized search on the world wide web," in *The Adaptive Web: Methods and Strategies of Web Personalization*, P. Brusilovsky, A. Kobsa, and W. Nejdl, Eds. Springer-Verlag, 2007, pp. 195–230.

[3] M. Porter, "An algorithm for suffix stripping," *Program*, vol. 14, no. 3, pp. 130–137, 1980.

[4] C. J. van Rijsbergen, *Information Retrieval*. London: Butterworths, 1979.

[5] H. Luhn, "The automatic creation of literature abstracts," *IBM Journal of Research and Development*, vol. 2, pp. 159–165, 1958.

[6] J. Ahn, P. Brusilovisky, D. He, J. Grady, and Q. Li, "Personalized Web exploration with task models," in *Proceedings of the International World Wide Web Conference*, 2008, pp. 1–10.

[7] Z. Dou, R. Song, and J. Wen, "A large-scale evaluation and analysis of personalized search strategies," in *Proceedings of the International World Wide Web Conference*, 2007, pp. 581–590.

[8] K. Sugiyama, K. Hatano, and M. Yoshikawa, "Adaptive Web search based on user profile constructed without any effort from user," in *Proceedings of the International World Wide Web Conference*, 2004, pp. 675–684.

[9] O. Hoeber and C. Massie, "Automatic topic learning for personalized re-ordering of Web search results," in *Proceedings of the Atlantic Web Intelligence Conference*, 2009, pp. 105–116.

[10] H. P. Zipf, *Human Behaviour and the Principle of Least Effort*. Cambridge, MA: Addison-Wesley, 1949.

[11] H. Edmondson and R. Wyllys, "Automatic abstracting and indexing survey and recommendations," *Communications of the ACM*, vol. 4, pp. 226–234, 1961.

[12] D. Shen, Z. Chen, Q. Yang, H. Zeng, B. Zhang, Y. Lu, and W. Ma, "Webpage classification through summarization," in *Proceedings of the International ACM/SIGIR Conference on Research and Development in Information Retrieval*, 2004, pp. 242–249.

[13] A. Dasdan, P. D'Alberto, S. Kolay, and C. Drome, "Automatic retrieval of similar content using search engine query interface," in *Proceedings of the ACM Conference on Information and Knowledge Management*, 2009, pp. 701–710.

[14] H. Liu and O. Hoeber, "Normal distribution re-weighting for personalized web search," in *Proceedings of the Canadian Conference on Artificial Intelligence*, 2011, pp. 281–284.

[15] M. Pao, "Automatic text analysis based on transition phenomena of word occurrences," *Journal of the American Society for Information Science*, vol. 29, no. 3, pp. 121–124, 1978.

[16] A. Booth, "A law of occurrences for words of low frequency," *Information and Control*, vol. 10, no. 4, pp. 386–393, 1967.

[17] National Institute of Standards and Technology, "TREC 2005 hard track," http://trec.nist.gov/data/t14_hard.html, 2005.

[18] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval: The Concepts and Terminology Behind Search*, 2nd ed. Addison-Wesley, 2011.