Concise Papers

A Web Surfer Model Incorporating Topic Continuity

Sankar K. Pal, *Fellow*, *IEEE*, B.L. Narayan, and Soumitra Dutta

Abstract—This paper describes a surfer model which incorporates information about topic continuity derived from the surfer's history. Therefore, unlike earlier models, it captures the interrelationship between categorization (context) and ranking of Web documents simultaneously. The model is mathematically formulated. A scalable and convergent iterative procedure is provided for its implementation. Its different characteristic features, as obtained from the joint probability matrix, and their significance in Web intelligence are mentioned. Experiments performed on Web pages obtained from WebBase confirm the superiority of the model.

Index Terms—Web intelligence, probabilistic surfer history, page ranking, stochastic processes, context identification, categorization.

1 Introduction

The present article deals with a methodology based on the principle of surfer models that simultaneously performs page ranking and context extraction.

Surfer models model a surfer who browses the Internet. The sequence of pages that the surfer visits is modeled as a stochastic process $\{X_t\}$, where X_t denotes the page the surfer is on at time t. The state space for this process consists of all Web documents, each page being a state that the process may attain. The transition probabilities $P(X_{t+1} = v | X_t = u)$ are defined as the probability of reaching page v, given that the surfer is currently on page u, by either clicking on a link available in u or by typing the URL.

In general, one would be interested in knowing some of the properties of the process $\{X_t\}$. One such property of interest is the convergence of this process to a stationary distribution. These properties may be used to draw inferences about, among others, the ranks and categories of Web documents.

The Random Surfer Model assumes that the surfer is browsing Web pages at random by either following a link from the current page chosen uniformly at random or by typing its URL. On the contrary, the Directed Surfer Model assumes that, when the surfer is at any page, he jumps to only one of those pages that are relevant to the context, the probability of which is proportional to the relevance of each outlink. Both models guarantee the convergence of this stochastic process to a stationary distribution under mild assumptions like the irreducibility of the transition probability matrix. In practice, these assumptions are enforced by pruning or ignoring some links.

This paper is an attempt at demonstrating the significance of incorporating the information derived from another aspect, namely, the history of a surfer for ascertaining the transition probabilities in the surfer model for ranking a page. Here, the surfer is assumed to follow, more often than not, links on topics contained on the pages that he had visited earlier, thus maintaining a continuity of topics.

It is shown to be possible to simultaneously estimate both the rank and categorization of the available pages, unlike the earlier models. As a result, both categorization and ranking improve. A mathematical framework of the model is provided along with its convergence and scalable properties. Other applications of the model, as obtained from the joint probability matrix, are also listed. The superiority of the model over some related ones is demonstrated on a data set obtained from WebBase [1].

- *S.K. Pal and B.L. Narayan are with the Machine Intelligence Unit, Indian Statistical Institute, 203, B.T. Road, Calcutta 700108, India. E-mail: {sankar, bln_r}@isical.ac.in.*
- *S. Dutta is with INSEAD, Boulevard de Constance, 77305 Fontainebleau, Cedex, France. E-mail: soumitra.dutta@insead.edu.*

Manuscript received 4 Mar. 2004; revised 4 Nov. 2004; accepted 1 Feb. 2005; published online 17 Mar. 2005.
For information on obtaining reprints of this article, please send e-mail to: tkde@computer.org, and reference IEEECS Log Number TKDE-0063-0304.

2 Surfer Models

A variety of surfer models, such as *random surfer* [2], *HITS (Hypertext Induced Topic Selection)* [3], *directed surfer* [4], *topic-sensitive pagerank* [5], etc., are available in the literature. More recently, another model called *WPSS (Web Page Scoring Systems)*, which generalizes all the above mentioned models, has been proposed in [6]. We describe here the Directed Surfer Model which encompasses the models which allow only forward walks.

Richardson and Domingos [4] have modeled a surfer who probabilistically chooses the next page to be visited depending on the content of the pages and the query terms the surfer is looking for. This model is an extension of the one introduced in [7].

The transition probability m'_{uv} is computed as

$$m'_{uv} = (1 - \beta)P'_q(u) + \beta P_q(v \to u),$$

where, u and v are Web documents, q is a query term, β is a constant, and $P'_q(u)$ and $P_q(v \to u)$ are arbitrary distributions. $P'_q(u)$ is the probability that the surfer reaches page u without following a link in the context of q. $P_q(v \to u)$, on the other hand, is the probability of choosing u in the context of q from among the links provided on page u. In practice, $P'_q(u)$ and $P_q(v \to u)$ may be derived from a relevance measure as

$$P'_q(u) = \frac{R_q(u)}{\sum_{v=1}^{N} R_q(v)}$$

and

$$P_q(v \to u) = \frac{R_q(u)}{\sum_{u \in F_v} R_q(u)},$$

where $R_q(u)$ is the relevance of u to q.

The choice of the relevance function is arbitrary. If $R_q(u) = 1, \forall q, u$, it is the random surfer model. Other suggestions for $R_q(u)$ provided in [4] include an indicator function for the presence of q in u and TFIDF-like scores for q in u. The latter ones make the model more efficient.

In general, by considering P'_q to be derived from R'_q, which need not be the same as R_q, Haveliwala's recent work on topic-sensitive ranking [5] of Web pages, which employs the directory listing of the Open Directory Project (ODP) [8], becomes its special case with the choices of $R'_q(u)$ and $R_q(u)$ being $\mathcal{I}[u \text{ appears under category } q \text{ in ODP}]$ and 1, respectively.

3 Surfer Model Incorporating History

3.1 Theory

The transition probabilities depend on the pages visited prior to reaching the current page. In order to incorporate this dependency, we propose a new surfer model, where a surfer moves on to pages that match his topic of interest. We assume that every page may

have content on one or more topics and the surfer chooses one of them as his topic of interest. Usually, the surfer moves on to a new page in keeping with his topic of interest. However, occasionally he may also visit other pages, say, out of curiosity.

The topic of interest is guessed by looking at the pages from which the page under consideration is reached. The knowledge of pages visited previously may be utilized by an online algorithm that computes the transition probabilities each time the surfer visits a new page.

Our primary interest being in offline applications, we, probabilistically, guess the history of the surfer and thereby estimate the topic of interest. We compute a set of transition probabilities under the assumption that a surfer generally browses with a particular topic of interest in his mind and is more likely to browse pages on similar topics rather than dissimilar ones. We formally introduce our model as follows.

Let X_t and I_t denote the page the surfer is on and his topic of interest, respectively, at time t. We assume that $I_t \in T_{X_t}$. We also assume that the probability of the surfer changing his topic of interest is γ ($\gamma < 0.5$), i.e., $P(I_{t+1} \neq I_t) = \gamma$. Then,

$$\begin{aligned} &P(X_{t+1} = z | X_t = v) \\ &= \sum_{k \in T_z} P(X_{t+1} = z, I_{t+1} = k | I_t = k, X_t = v) P(I_t = k | X_t = v) \\ &+ \sum_{k \in T_z} P(X_{t+1} = z, I_{t+1} = k | I_t \neq k, X_t = v) P(I_t \neq k | X_t = v). \end{aligned} \tag{1}$$

In (1), we have split the probability of a one-step transition (from v to z) into that of within-topic transitions and cross-topic transitions. Let N_{vk} denote the number of outlinks from v to pages containing topic k. Now, substituting γ for cross-topic transition and simplifying, we have:

$$\begin{aligned} P(X_{t+1} = z | X_t = v) = &\sum_{k \in T_z} \frac{\gamma}{N_{vk} |T_z|} \left(1 - \gamma - \frac{\gamma}{|T_z|}\right) \\ &\sum_{k \in T_z \cap T_v} \frac{P(I_t = k | X_t = v)}{N_{vk}}. \end{aligned} \tag{2}$$

In the above, only the nonzero terms have been retained in the summation. Thus, the problem of estimating the transition probability has now been restated in terms of $P(I_t = k | X_t = v)$, the surfer's topic of interest given his current location. This conditional probability is expanded as:

$$\begin{aligned} &P(I_t = k | X_t = v) \\ &= \sum_{u \in B_v} \frac{P(I_t = k, X_t = v, X_{t-1} = u)}{P(X_t = v)} \\ &= \sum_{u \in B_v} \sum_{l \in T_u} \frac{P(I_t = k, X_t = v | X_{t-1} = u, I_{t-1} = l) P(I_{t-1} = l, X_{t-1} = u)}{P(X_t = v)}. \end{aligned} \tag{3}$$

In the above, we have expressed the quantity in terms of the possible pages and topics that could have been attained at time $t-1$. We again substitute γ for the probability of the topic having changed during a transition (and $1-\gamma$ for no change) and obtain the following:

$$\begin{aligned} P(I_t = k | X_t = v) = &\sum_{u \in B_v} \frac{1-\gamma}{N_{uk}} P(I_{t-1} = k | X_{t-1} = u) \frac{P(X_{t-1} = u)}{P(X_t = v)} \\ &+ \frac{\gamma}{|T_v|} \sum_{u \in B_v} \sum_{l \in T_u \setminus \{k\}} \frac{P(I_{t-1} = l | X_{t-1} = u)}{N_{ul}} \\ &\frac{P(X_{t-1} = u)}{P(X_t = v)}. \end{aligned} \tag{4}$$

We have thereby expressed the conditional probability, at time t, of the topic of interest given the page, in terms of the same quantities at time $t-1$. In this manner, we have obtained an iterative procedure for computing the conditional probabilities. Note that this iterative procedure also allows us to compute the joint probabilities of (I_t, X_t) by multiplying both sides by $P(X_t = v)$.

The proof of convergence of the above iterative procedure follows by noting that the situation is the same as that of the Random Surfer Model, with each state now becoming a pair of values.

3.2 Obtaining Initial Estimates

For faster convergence of the above iterations, a good initial estimate of the joint probability distribution is necessary. The joint probability $P(I_0, X_0)$ is estimated as $P(I_0|X_0)P(X_0)$, where $P(X_0)$ is obtained using an existing version of PageRank. The quantities $P(I_0|X_0)$ have been estimated from the ODP data using the Naive Bayes algorithm [9], where the document X_0 is treated as the term vector $(x_1, x_2, \ldots, x_K)$ and C_j is the jth topic listed under the Open Directory. The topic-conditional probabilities for each term, $P(x_i|C_j)$, and the prior probabilities of the topics $P(C_j)$, are estimated as the corresponding frequencies obtained from the pages listed under the Open Directory.

These initial estimates are then plugged into the iterative procedure.

3.3 Different Uses and Characteristic Features

As described above, we have obtained the stationary (joint) distribution of (I, X). By considering appropriate quantities like the marginals and the conditionals of the joint probability matrix of (I, X), as has been done in [7], the *page ranks*, *topic-specific page ranks*, *relevance to topics*, etc., may be computed. The parameter γ may be varied to reflect the behavior of a particular surfer and may be utilized for personalization.

Earlier investigations had generally focused on either page ranking or page categorization. What the current investigation does is to perform both these interdependent tasks simultaneously. This notion had been mentioned in [10] and had also been independently reported in a preliminary form in [11].

It is interesting to note that the proposed algorithm does not actually categorize, in its true sense, a Web page's contents. It just estimates what a surfer would be interested in when he reaches a page. This means that, even though the terms appearing in a document are suggestive of some particular category, the topics and ranks of the pages linking to it play a major role in determining if it indeed is relevant for that category. The scores that we compute for each page can be considered equivalent to categorization in the sense that this is what a surfer visiting this page would be thinking about.

We mention here that, though both Haveliwala's topic-sensitive PageRank algorithm and the proposed one make use of the topic information available under the ODP directory, there is a difference in the manner in which it is employed. While the former needs information about which topics a URL is listed under, the latter needs some text categorization mechanism, which in this particular case, happens to be derived from the ODP data. Since Haveliwala's algorithm does not need to categorize each available page, it is computationally more efficient compared to our algorithm. However, by virtue of this extra effort, the proposed methodology counters topic drift by restricting the transfer of rank between dissimilar pages.

3.4 Complexity and Scalability

We now discuss the complexities involved in the proposed algorithm. Let K be the number of topics under consideration. Then, the disk space required is K times that required for ordinary PageRank and is the same as that for topic-sensitive PageRank [5]. The time complexity is as follows: From (4), it is obvious that, in

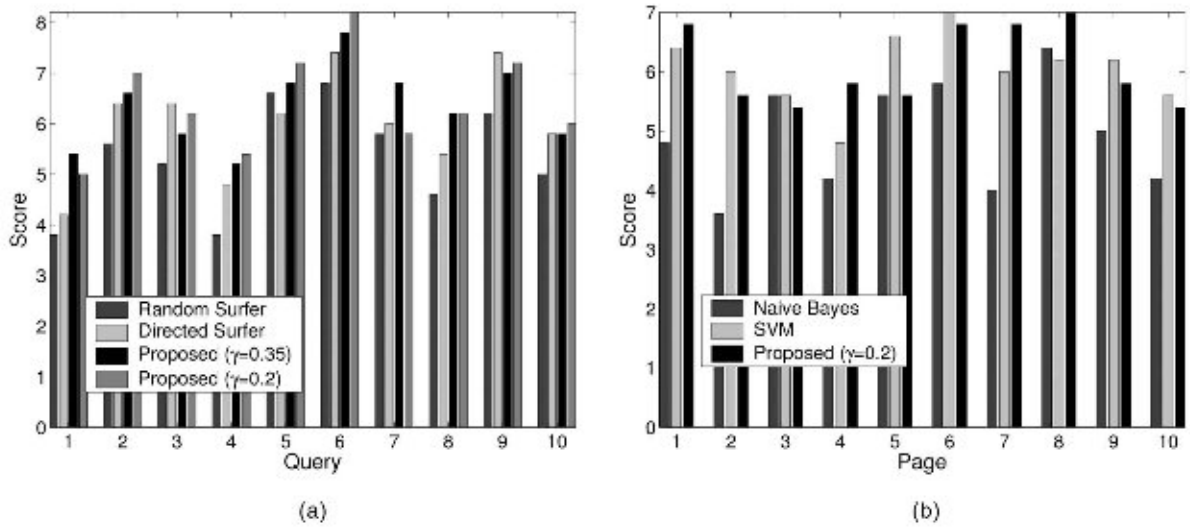

Fig. 1. Experimental results: (a) page rank comparison and (b) comparison of categorization.

each iteration and for each k and v, $3 * |T_u|$ computations are required for each backlink u of v. This makes a total of $3 * \overline{T} * |B_v|$ computations, where $\overline{T}$ is the average number of topics contained in a page. Thus, the total number of computations needed for each iteration is $3 * \overline{T}^2 * N * \overline{B}$, where $\overline{B}$ is the average number of backlinks.

This is about the same as that for topic-sensitive PageRank if we assume that $\overline{T}^2$ is comparable to K. In practice, this is a reasonable assumption as, on average, the number of topics represented on a page is much smaller compared to K. Note that we do not count the preprocessing steps like stemming, stopword removal, and creation of an inverted index as they are the same for any such algorithm.

4 Experimental Results

The performance of the proposed methodology has been evaluated and compared with some of the existing algorithms. Here, we discuss the data sets used and the methods of implementation and evaluation.

4.1 Data Sets Used

A training data set is obtained from the ODP [8] in the form of a file in RDF format. This file consists of URLs and their description organized into 17 distinct topics. The words available in the description are assumed to represent, to some extent, the topics under which they appear.

The test data set, consisting of approximately five million pages, was obtained from WebBase [1]. These pages form a connected neighborhood of the Web. The headers were ignored and only the contents of each page were used in our experiments. The links were normalized and self-links were removed.

For estimating the value of γ, we have used the *msnbc.com anonymous Web data* (available at http://kdd.ics.uci.edu/databases/msnbc/msnbc.html), which has transition information between categories.

4.2 Implementation

Initial categorizations of pages from WebBase were obtained using the Naive Bayes algorithm and the ODP data set, as mentioned earlier. We observed that the number of cross-topic transitions in the msnbc data were about 35 percent. Accordingly, we chose the value of γ to be 0.35. Note that, this data set did not capture any requests that had been served from the user's cache. Had these requests been included in the data set, the number of within-transitions would have been higher, i.e., γ value would have been lower. In order to reflect this, we also conducted an experiment for a lower value of γ (= 0.2, say).

Our implementation is similar to the one suggested in [12], where accesses to the hard disk are minimized by reading chunks of the transition matrix into memory. Extrapolation methods with a higher degree of efficiency may also be used in the computation of the principal eigenvector of the matrix S [13].

4.3 Evaluation

The process of evaluation consists of two parts: The first part deals with comparison of the ranks of the pages and the second with their categorizations.

The ranks obtained by our method were compared against those obtained by PageRank [2] and the directed surfer model [4]. Ten queries from those used in [5] and five volunteers were chosen. The queries are *architecture*, *bicycling*, *computer vision*, *gardening*, *gulf war*, *java*, *rock climbing*, *table tennis*, *vintage wine*, and *volcano*. The top 10 pages obtained in response to each query by the four algorithms, namely, Random Surfer Model (or PageRank), Directed Surfer Model (or QD-PageRank), Proposed approach with $\gamma = 0.35$, and Proposed approach with $\gamma = 0.2$, were studied by the five volunteers. They provided a rating (or score) between 0 and 10, a rating of 10 being the best, to each algorithm for each query. The average values obtained for each query are presented in Fig. 1a. We have performed pairwise comparisons testing for difference of the means using a t-test with 9 degrees of freedom. The null hypothesis was taken to be that the means were equal, while the alternate hypothesis was that the second method (the one appearing later in the bar-plot) fared better. The tests gave the following results:

- The proposed method (for both the above mentioned values of γ) and the directed surfer model significantly outperform the random surfer model at a confidence level of 95 percent.
- The scores obtained by the proposed method with $\gamma = 0.2$ show a significant improvement over those obtained by the directed surfer model at a confidence level of 95 percent.
- The improvement in scores over the directed surfer model obtained by the proposed method with $\gamma = 0.35$ is

TABLE 1
Running Times

Method	Time (secs)
PageRank	60
QD-PR	150
Topic-sensitive PR	1000
Proposed	1300

significant at a confidence level of 90 percent, but not at a confidence level of 95 percent.

- No significant difference was observed between the scores of the proposed algorithm for the two choices of γ.

In the second part of our evaluation, we compared the categorization of Web pages by our method with those of Naive Bayes [9] and SVMLight [14] with a polynomial kernel.

All of these methods including the proposed one were then used to obtain the topic categorization of 10 randomly chosen pages into the ODP topics, and the earlier volunteers were asked to rate them. A score of 10 to a page denotes that the volunteer viewed the topic categorization as totally appropriate, while a score of 0 denotes a complete mismatch with the volunteer's categorization. The results are shown in Fig. 1b only for $\gamma = 0.2$, as an example. Both the proposed algorithm and SVMLight produced significantly better categorization than the Naive Bayes algorithm at 95 percent confidence level. Since our algorithm has used the Naive Bayes algorithm for initial estimates, this indicates that our method has improved the categorization, as expected. However, it is seen that both ours and SVMLight are at par even at a 90 percent confidence level.

It may be noted that we had already obtained the joint probability matrix during our experiments on page ranking. Therefore, the categorization experiment only needed to compute the marginal and conditional probabilities, which are computationally inexpensive.

For checking the scalability of the proposed algorithm, we measured the time taken by it and compared to those of PageRank, QD-PageRank, topic-sensitive PageRank (Table 1). The times mentioned are only those for the actual computation of the ranks and not for the preprocessing steps which are common to all. None of the above computations employed the extrapolation methods [13] mentioned above. The average number of topics turned out to be about three per page, which resulted in the proposed algorithm having a running time almost equal to that of the topic-sensitive PageRank. This also confirms our observation in Section 3.4.

5 Conclusions

The problem of modeling the interrelationship between page categorization and ranking in terms of topic continuity has been addressed in this article. An offline algorithm developed for this purpose probabilistically estimates the surfer's history and, thus, his/her current topic of interest. The incorporation of surfer history (or topic continuity) is a unique feature of this methodology. This resulted in a scalable iterative procedure that provides page categorizations as well as ranking simultaneously. The convergence of the iterative procedure is proved. The merits of the methodology have been established. Although we have presented experimental results only for page ranking and categorization, the method can be made applicable for topic-sensitive page ranking, topic representation on the Web, and personalization.

Acknowledgments

The authors are thankful to the WebBase Group at Stanford for providing the WebBase data. Mr. B.L. Narayan is the recipient of the ISI-INSEAD (France) Fellowship to carry out doctoral research.

References

[1] http://www-diglib.stanford.edu/~testbed/doc2/WebBase/, 2005.
[2] S. Brin and L. Page, "The Anatomy of a Large-Scale Hypertextual Search Engine," Stanford Univ., technical report, 1998.
[3] J. Kleinberg, "Authoritative Sources in a Hyperlinked Environment," *J. ACM*, vol. 46, no. 5, pp. 604-632, 1999.
[4] M. Richardson and P. Domingos, "The Intelligent Surfer: Probabilistic Combination of Link and Content Information in Pagerank„" *Advances in Neural Information Processing Systems*, vol. 14, pp. 1441-1448, MIT Press, 2002.
[5] T.H. Haveliwala, "Topic-Sensitive Pagerank: A Context-Sensitive Ranking Algorithm for Web Search," *IEEE Trans. Knowledge and Data Eng.*, vol. 15, no. 4, pp. 784-796, Jan./Feb. 2003.
[6] M. Diligenti, M. Gori, and M. Maggini, "A Unified Probabilistic Framework for Web Page Scoring Systems," *IEEE Trans. Knowledge and Data Eng.*, vol. 16, no. 1, pp. 4-16, Jan. 2004.
[7] D. Rafiei and A.O. Mendelzon, "What Is This Page Known for? Computing Web Page Reputations," *Proc. Ninth Int'l World Wide Web Conf.*, pp. 823-835, 2000.
[8] http://dmoz.org/about.html, 2005.
[9] D.D. Lewis, "Naive (Bayes) at Forty: The Independence Assumption in Information Retrieval," *Proc. ECML-98, 10th European Conf. Machine Learning*, pp. 4-15, 1998.
[10] B.D. Davison, "Unifying Text and Link Analysis," *Proc. Int'l Joint Conf. Artificial Intelligence Workshop Text-Mining & Link-Analysis (TextLink)*, 2003.
[11] B.L. Narayan, C.A. Murthy, and S.K. Pal, "Topic Continuity for Web Document Categorization and Ranking," *Proc. 2003 IEEE/WIC Int'l Conf. Web Intelligence*, pp. 310-315, 2003.
[12] T.H. Haveliwala, "Efficient Computation of Pagerank," Stanford Univ., technical report, 1999.
[13] S.D. Kamvar, T.H. Haveliwala, C.D. Manning, and G.H. Golub, "Extrapolation Methods for Accelerating Pagerank Computations," *Proc. 12th Int'l World Wide Web Conf.*, pp. 261-270, May 2003.
[14] T. Joachims, "Text Categorization with Support Vector Machines: Learning with Many Relevant Features," *Proc. ECML-98, 10th European Conf. Machine Learning*, pp. 137-142, 1998.
[15] S. Chakrabarti, M. Joshi, K. Punera, and D. Pennock, "The Structure of Broad Topics on the Web," *Proc. 11th Int'l World Wide Web Conf.*, pp. 251-262, 2002.
[16] P. Baldi, P. Frasconi, and P. Smyth, *Modeling the Internet and the Web: Probabilistic Methods and Algorithms*. J. Wiley and Sons, 2003.
[17] S. Chakrabarti, *Mining the Web: Discovering Knowledge from Hypertext Data*. Morgan Kaufman, 2002.