Topic Continuity for Web Document Categorization and Ranking

B. L. Narayan, C. A. Murthy and Sankar K. Pal
Machine Intelligence Unit,
Indian Statistical Institute,
203, B. T. Road,
Kolkata - 700108, India.
E-mail:{bln_r, murthy, sankar}@isical.ac.in.

Abstract

PageRank is primarily based on link structure analysis. Recently, it has been shown that content information can be utilized to improve link analysis. We propose a novel algorithm that harnesses the information contained in the history of a surfer to determine his topic of interest when he is on a given page. As the history is unavailable until query time, we guess it probabilistically so that the operations can be performed offline. This leads to a better web page categorization and, thereby, to a better ranking of web pages.

1. Introduction

Traditional information retrieval in document analysis has largely focused on looking at the content of a document for drawing inferences regarding it. The categorization of the document is performed on the basis of the terms (or words) present in the document. The term frequency reflects the significance of the document with respect to the term whereas the inverse document frequency accounts for the relevance of a document. The TFIDF measure is a combination of the two and it provides the importance of a document with respect to a set of terms or queries.

Unlike text documents, a hypertext document is not self-contained. There is a wealth of information about the current document hidden in its neighborhood which is determined through the link structure of the available hypertext data. A hypertext document has both in-neighbors and out-neighbors, and the text associated with its neighbors proves to be useful in inferring about the content of the present document. Chakrabarti, *et al* [2] have performed experiments in this regard and have shown promising results.

Link structure, by itself, has been in use for identifying important documents across the web and ranking them. Search engines combine the ranks and the TFIDF measures of web pages and present them, in order of their importance, to the user in response to a given query. In order to satisfy the user, it is essential to obtain the best such ordering.

Recently, Richardson and Domingos [8], have proposed a directed surfer model which enhances the PageRank by taking into consideration the link structure as well as the content of the documents in the neighborhood. As the probability of following a link is assumed to be associated to the contents of the two pages that the link connects, it is necessary to have a proper text scoring function which shall be used to judge how relevant the page is to the given query.

In this article, we describe a methodology for computing the PageRank that combines the link structure, and the contents of a page and its neighborhood. Though the directed surfer model computes the probability of following a particular link based on the contents of the current page, the history of the surfer is ignored. Since all the computations are performed offline, the proposed algorithm guesses the history with the help of the contents on the backlinks of the given page. This leads to a better categorization of the web documents and improves their PageRank.

The remaining part of this article is organized as follows. First, we shall discuss the related work in Section 2. We describe our proposed methodology in Section 3. This is followed by the experimental results, in Section 4.

2 Related work on page ranking and content analysis

Simple citation ranking is used in document analysis where the number of citations of a document is considered to be a measure of importance of that page. Brin and Page [1] and Kleinberg [7] extended the idea to web page ranking where the number of inlinks of a page was taken to be a measure of its importance.

We shall use the following notation in this article. u, v

and w denote hypertext pages and k and l stand for topics. Let $L = ((l_{uv}))_{u=1,\ldots,N,v=1,\ldots,N}$ be the link matrix where N is the number of web pages, *i.e.*, $l_{uv} = 1$ if and only if page u has a link from page v. If u is a web page, $F_u := \{v \in \{1,\ldots,N\} : l_{vu} = 1\}$ is the set of pages u points to, and $B_u := \{v \in \{1,\ldots,N\} : l_{uv} = 1\}$ is the set of pages that point to u. $N_u := |F_u|$ is the number of outlinks from u. Let M denote the normalized link matrix obtained by dividing each column of L by its total.

Kleinberg proposed the Hypertext Induced Topic Selection (HITS) algorithm for ranking web pages where he introduced the notion of *hubs* and *authorities*. The hub value of a page is defined to be the sum of the authority values of the pages that it links to and the authority value of a page is the sum of the hub values of the pages that link to it. The hub and authority values are computed for a subgraph A that consists of only pages deemed relevant to the query.

$$a(u) = \sum_{v \in B_u} h(v),$$

$$h(u) = \sum_{v \in F_u} a(v).$$

Rewriting in matrix form and combining the two steps,

$$\mathbf{a} = A^T\mathbf{h} = A^TA\mathbf{a},$$

$$\mathbf{h} = A\mathbf{a} = AA^T\mathbf{h}.$$

These steps are performed till convergence. The hub and authority computations are performed afresh for each query. This makes it computationally costly. In the rest of this article, we do not consider HITS and other similar algorithms like PHITS [3].

The PageRank algorithm was introduced by Brin and Page [1], and is employed by Google (`http://www.google.com`). Every page is considered to distribute its rank equally into its outlinks. The rank of a page is recursively defined as the sum of the ranks conferred on it by its backlinks. Starting with an arbitrary vector, the PageRank vector is computed iteratively by multiplying with the normalized link matrix M. To avoid rank sinks and rank leaks, the transition matrix is modified as $M' = (1-\alpha) * M + \alpha[\frac{1}{N}]_{N\times N}$.

$$\mathbf{R} = M'\mathbf{R}$$

The PageRank vector turns out to be the dominant eigenvector of M'. The PageRank vector is computed without considering the content information and is independent of the query. At query time, only the relevance of the page to the query is computed and the results are ordered taking both the relevance and the (unconditional) rank into consideration.

The PageRank algorithm has another interpretation, known as the *random surfer model*. It is assumed that a surfer is browsing web pages by clicking on the links at random, never clicking the back button. The state of the stochastic process is the web page that the surfer is on and the click corresponds to a transition. The matrix M' corresponds to the stochastic matrix for the process. α is the probability that the surfer decides not to follow a link and moves to a new page by typing its URL. Under this model, the PageRank turns out to be the stationary probability of the process. It can also be interpreted as the unconditional probability of the surfer being on a page.

The relevance of a page to a given query is computed by calculating the TFIDF measure of the page for each term in the query. A page is considered to be significant for a term if the term frequency (TF) is high. If a term appears in almost all the documents, it might not be relevant. So, the term frequency is weighted by the inverse of the documents frequency in which it appears. This provides the TFIDF measure for each term. The TFIDF measures for each term in the query are added to provide the TFIDF value of the page for a given query.

When dealing with hypertext data, the TFIDF measure often fails to identify some relevant web pages. A celebrated example of such a situation is for the query "search engine". Home pages of most search engines do not contain the phrase "search engine" and, as a result, get a low TFIDF value. So, even though they have a huge number of inlinks, and hence a high PageRank, they do not feature at the top of the results. This sort of a problem was overcome by considering text from the neighbourhood of a given web page as its own. Chakrabarti, *et al* [2] have studied various ways of assigning text to a document from its neighbors.

Richardson and Domingos [8] have attempted to rectify the above mentioned problem by combining the content information into the page ranking algorithm itself. The probability of following a link on a page containing the query is weighted according to the presence or absence of the query in the page that the link leads to. This is called the "directed surfer model" which caters to the intuition that a surfer would more likely follow a link to a page that contains the query rather than to a page that does not.

Recently, Haveliwala [5] has proposed to compute a PageRank vector for each distinct topic. As the computation of PageRank is time consuming, it is suggested that the number of topics be kept small. For each topic, a different bias vector is used during the computation of PageRank. The topic is decided on the basis of the context of the query.

In each of the mentioned cases, the original assumption of the random surfer model, that each of the links on a page is equally likely to be followed, has been relaxed. The probability of following a link has been modified or biased in certain ways. There have been other modifications that took

into account the location or the visual characteristics of the content and links on the page. These too can be taken into account while computing the transition probabilities.

The directed surfer model assumes an accurate classification of any given web page into the topics that are contained in it. Simple text based classification does not perform well due to the very nature of a language where the same words may appear in different topics. The need for context sensitive page categorization, therefore, arises for this reason. In the following section, we describe the problem that we consider here and provide a solution to it. We also discuss how it differs from other related approaches.

3. Methodology for incorporating history into content analysis

3.1. Assumptions

We assume that every page consists of content on one or more of a set of predefined topics and that pages on similar or same topics are more likely to be linked to each other than pages on totally unrelated topics. We also assume that a surfer would browse pages not just randomly, but with some objective in mind. There is a topic of interest (ToI) for the surfer at any time t and he is more likely to be interested in the same topic at time $t + 1$. With a small probability, he does change his topic of interest, possibly, out of curiosity. Another related assumption is that, if a surfer is on a particular page, his topic of interest is one of the set of topics available on the page. For the time being, we consider the case where the surfer always follows a link on the current page.

3.2. Problem under consideration

With the above assumptions in mind, we consider the following example. A page v leads to two pages w_1 and w_2 (see Fig. 1). If a surfer is on page v at time t, what is the probability that he is on page w_1 at time $t + 1$? The random surfer model had assigned a probability of 0.5 to that (assuming the surfer follows one of the links available on page v). If, after analyzing the textual content, it is found that v belongs equally to topics k_1 and k_2, w_1 belongs to k_1 and w_2 belongs to k_2, then what is the above probability? The directed surfer model prescribes a value of 0.5. However, had it been known that the surfer had followed the link from u_3, would it still be as likely for the surfer to reach w_1? In such a situation, it seems to be more likely that the surfer is interested in k_2 and, hence, quite probably, would visit w_2 rather than w_1.

From this example, it is evident that both the transition probabilities and the topic categorization depend not just on the contents of the current page but on the PageRank and the contents of the backlinks too. We make use of the information hidden in the surfer's history to compute both the above quantities simultaneously.

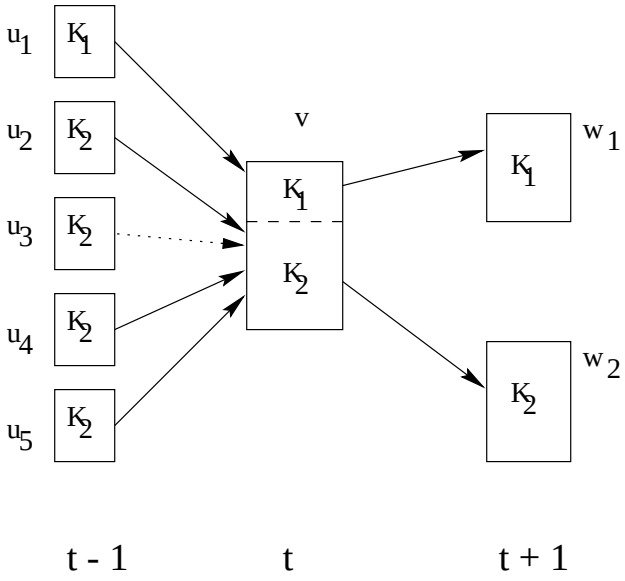

Figure 1. Given that the surfer is on page v at time t what is the probability of him being on page w_2 at time $t + 1$? What if it is known that the surfer had followed the link shown by the dotted line?

Though it is evident that the history of the surfer plays an important role in determining the transition probabilities, it is not likely to be obtained before query time. Our intention being a quick response at query time, the computations are required to be performed offline. To this end, we probabilistically guess the history of a surfer on a given page. As we consider only the case where the surfer follows a link to reach the current page, we apply Bayes' Theorem to find the probability of the surfer having reached here from a particular backlink. The ToI too is determined by a similar approach.

It may be noted here that though Chakrabarti, *et al* [2] had considered text from neighbouring pages, they had not taken into consideration the rank of the pages that lead to the current page. A backlink with a low PageRank, although with the same content as the current page, is not as likely to have been visited as compared to a backlink with a high PageRank.

3.3. Methodology

Based on the aforesaid concept we provide here an algorithm to compute the transition probabilities in terms of the ToI. As these two quantities depend on each other, we com-

pute both the ToI and the transition probabilities recursively. Once the transition matrix is available, the PageRank vector is obtained in the usual manner as the dominant eigenvector of the transition matrix.

We assume that the surfer strays from his current ToI with a small probability of γ. The transition probabilities are computed as follows.

$$
\begin{aligned}
& P(X_{t+1} = z | X_t = v) \\
= & \sum_{k \in T_z} P(X_{t+1} = z, ToI_{t+1} = k | X_t = v) \\
= & \sum_{k \in T_z} [P(X_{t+1} = z, ToI_{t+1} = k | ToI_t = k, X_t = v) \\
& \qquad P(ToI_t = k | X_t = v) \\
& + \sum_{k \in T_z} P(X_{t+1} = z, ToI_{t+1} = k | ToI_t \neq k, X_t = v) \\
& \qquad P(ToI_t \neq k | X_t = v)] \\
= & \sum_{k \in T_z} \frac{1-\gamma}{N_{vk}} P(ToI_t = k | X_t = v) \\
& + \sum_{k \in T_z} \frac{\gamma}{N_v} (1 - P(ToI_t = k | X_t = v)),
\end{aligned}
$$

where $P(ToI_t = k | X_t = v)$ is taken to be 0 if $k \notin T_v$.

In this manner, the transition probabilities can be calculated once the quantities $P(ToI_t = k | X_t = v)$ are known. They are normalized each time so that

$$\sum_{k=1}^{17} P(ToI_t = k | X_t = v) = 1$$

The topic of interest of the surfer given that the surfer is on page v is calculated as:

$$
\begin{aligned}
& P(ToI_t = k | X_t = v) \\
= & \frac{P(ToI_t = k, X_t = v)}{P(X_t = v)} \\
= & \sum_{z \in B_v} \frac{P(ToI_t = k, X_{t-1} = u, X_t = v)}{P(X_t = v)}
\end{aligned}
$$

The denominator can be ignored as it will disappear during normalization. Now,

$$
\begin{aligned}
& P(ToI_t = k, X_{t-1} = u, X_t = v) \\
= & \sum_{l \in T_u} P(ToI_t = k, X_t = v, ToI_{t-1} = l, X_{t-1} = u) \\
= & \sum_{l \in T_u} P(ToI_t = k, X_t = v \mid ToI_{t-1} = l, X_{t-1} = u) \\
& \qquad P(ToI_{t-1} = l, X_{t-1} = u) \\
= & \sum_{l \in T_u} P(ToI_t = k, X_t = v \mid ToI_{t-1} = l, X_{t-1} = u) \\
& \qquad P(ToI_{t-1} = l \mid X_{t-1} = u) P(X_{t-1} = u)
\end{aligned}
$$

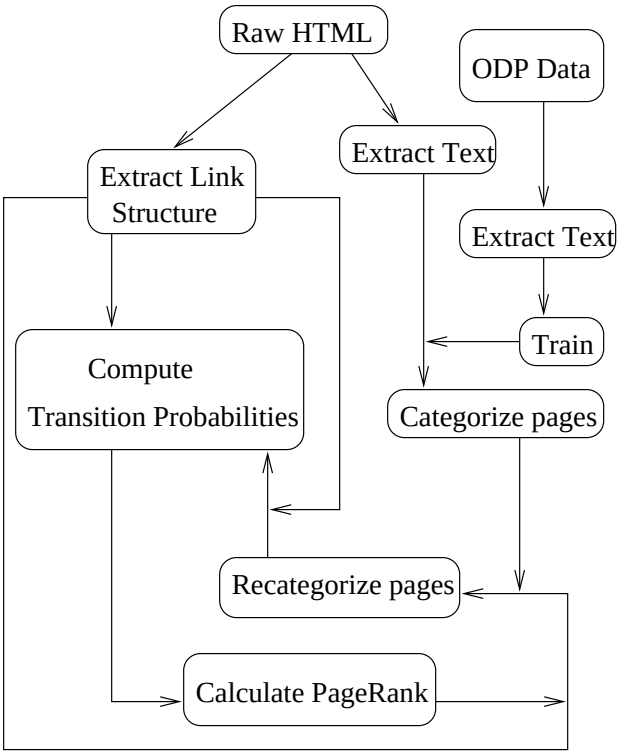

Figure 2. Flowchart of the proposed algorithm

The initial values of $P(ToI_t = k | X_t = v)$ can be estimated by text based scoring. Once the ToI has been calculated accurately, the directed surfer model can be employed to enhance the PageRank. A flowchart of the proposed algorithm is provided in Fig. 2. It has been shown that the directed surfer model is scalable. The above computations too, are scalable. The operations are performed in exactly the same manner as for efficiently computing the PageRank [4].

4. Experimental Results

To deonstrate the effectiveness of our method, one million pages have been obtained from Stanford's WebBase [6] available online at `www-diglib.stanford.edu/~testbed/doc2/WebBase/`. Since most of these pages belong to sites related to Stanford University, we added "Stanford" to the list of stopwords.

The training data for the analysis of the textual content of all these pages has been taken from

the Open Directory Project (ODP) available online at `http://rdf.dmoz.org/rdf/content.rdf.u8.gz`. About 3.6 million web pages are available under the seventeen categories in the ODP data. We have, however, ignored the categories *regional* and *world* due to the high number of non-English words that they generate. The classification is performed by manual volunteers and is considered to be of high quality.

Due to lack of time for training, only the description of each page (and not the content) has been considered. Each word is stemmed with Porter's Stemming Algorithm prior to its use in the analysis. The frequency of a stem for each topic is computed by counting the number of times the stem appears under that particular topic. Stopwords are ignored at the outset itself. Rare stems, *i.e.*, those appearing fifteen times or less, are treated as spelling mistakes and are ignored.

For each page v and topic k, we compute $P(ToI = k|Page = v)$. After normalizing, if any of these probabilities is found to be less than 0.1, we consider it to be noise and forcibly put it to zero. If a page does not appear to be significantly related to any of the fifteen topics, we replace the ToI vector by a *prior* vector which is just the unconditional probability of a page being on a particular topic. This vector is estimated from the ODP data as the proportion of topics under each category. The scores are normalized once more. The initial estimates are based on a simple text scoring function which uses the frequencies of the stems in each of the fifteen topics to decide the set of topics that are available on the current page.

Refinements are then made on the basis of the link structure and an initial estimate of PageRank (the one computed assuming equal probabilities for all outlinks). Once the categorization is performed accurately, the desired transition probability matrix is obtained. The dominant eigenvector of this matrix is taken to be the final PageRank vector.

A group of three volunteers were requested to study the categorizations obtained before and after our algorithm was applied. The distances between the two categorization vectors was computed. Fifteen such pages were obtained for which the distance turned out to be larger than 0.7 (Table 1. These were made available to the volunteers and they were asked to rate the extent of their agreement with each categorization. The ratings (Table 2) were analysed and the improvement has been found to be significant by a t-test (with 14 degrees of freedom) (see Table 3). The null hypothesis that the proposed method does not provide an improvement in topic categorization was rejected at a confidence level of 97.5%.

Table 1. URLs whose categorization vectors were at a distance greater than 0.7

No.	URL
1.	tour.stanford.edu/cgi/search.prl/d/
2.	pangea.stanford.edu/students/jobfair2.html
3.	labrea.stanford.edu/gnu/cvs/
4.	cis.stanford.edu/programs/talks/cauth/abstract.html
5.	w6yx.stanford.edu/ paulf/
6.	kzsu.stanford.edu/pguide/1996spring/page05.html
7.	cellwall.stanford.edu/cellwall/species/ arabidopsis_atcsla/graphics.shtml
8.	cellwall.stanford.edu/cellwall/species/ arabidopsis_atcsle/graphics.shtml
9.	kzsu.stanford.edu/pguide/1996winter/page19.html
10.	tehran.stanford.edu/literature/poetry/classic/ hafez.0.html
11.	kzsu.stanford.edu/pguide/1996winter/page24.html
12.	sloan.stanford.edu/evonline/profiles.htm
13.	labrea.stanford.edu/gnu/gnurobots/
14.	labrea.stanford.edu/gnu/goose/
15.	assu.stanford.edu/speakers/bios/clark.html

Table 2. Ratings of categorizations averaged over all users

URL No.	Rating	
	Text based	Link based
1	0.4333	0.3667
2	0.3333	0.9333
3	0.6000	0.3333
4	0.1333	1.0000
5	0.3667	0.7333
6	0.2333	0.8667
7	0.4000	0.8667
8	0.4000	0.8667
9	0.1667	0.5667
10	0.1667	0.4000
11	0.2000	0.6333
12	0.5667	0.6000
13	0.7000	0.3000
14	0.7000	0.3000
15	0.1667	0.6667

Table 3. Summary of results

	Mean (s.d.)	Mean (s.d.)
Average rating by each user	0.3200 (0.1897)	0.6533 (0.2642)
	0.4133 (0.2446)	0.6200 (0.2366)
	0.3800 (0.2569)	0.6133 (0.3314)
Overall average	0.3711 (0.1971)	0.6289 (0.2452)
$Rating_2 - Rating_1$	0.2578 (0.3905)	
t-statistic	2.56	
$t_{14,0.975}$	2.14	

5. Conclusions

Page ranking is greatly enhanced by considering the textual content of web pages along with the link structure. Proper categorization of web pages is required to obtain a correct ranking technique. However, by its very nature, a language consists of words that appear in multiple categories. The confusion that such words add to categorization of web pages can be reduced by considering the context of the web page. We have suggested one such approach where the context is automatically obtained from the backlinks of a page. We have also shown how the categorization can be refined starting with an initial classification.

The present investigation shows the improvement of categorization of web documents because of the incorporation of history. this can be subsequently used to enhance the page rank using the algorithm of Richardson and Domingos [8].

6. Acknowledgements

We would like to thank Wang Lam for helping us with retrieving documents from the Stanford Webbase. We are also grateful to the volunteers for sparing their valuable time. The first author's research is funded by INSEAD, Fontainebleau, France.

References

[1] S. Brin and L. Page. The anatomy of a large-scale hypertextual search engine. Technical report, Stanford University, 1998.

[2] S. Chakrabarti, B. Dom, and P. Indyk. Enhanced hypertext categorization using hyperlinks. In *Proceedings of SIGMOD-98, ACM International Conference on Management of Data*, 1998.

[3] D. Cohn and H. Chang. Learning to probabilistically identify authoritative documents. In *Proceedings of the 17th International Conference on Machine Learning*, 2000.

[4] T. H. Haveliwala. Effi cient computation of pagerank. Technical report, Stanford University, 1999.

[5] T. H. Haveliwala. Topic-sensitive pagerank. In *Proceedings of the Eleventh International World Wide Web Conference*, 2002.

[6] J. Hirai, S. Raghavan, H. Garcia-Molina, and A. Paepcke. Webbase: A repository of web pages. In *Proceedings of Ninth International World Wide Web Conference*, 2000.

[7] J. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.

[8] M. Richardson and P. Domingos. The intelligent surfer: Probabilistic combination of link and content information in pagerank,. In *Advances in Neural Information Processing Systems*, volume 14. MIT Press, 2002.