Describing Web Pages better in Search Results

Deepak P,
Model Engg: College, Kochi, India
deepak-p@eth.net

Jyothi John,
Model Engg: College, Kochi, India
jyothijohn@mec.ac.in

Introduction

This paper presents an algorithm for generating web page descriptions automatically in situations where the need is for a lot of information within a few sentences, a typical example being the descriptions of web pages in search results. Such descriptions often prove to be very useful aids for navigation. Many search engines provide extracts from the page as a description. But studies say that extracts from pages do not usually serve the cause. A page is often described by other pages better that by itself. This work presents an algorithm which retrieves descriptions about a page from those linking to it and ranks them. The descriptions generated on test sets have been compared to those generated by acclaimed search engines. It may well be noted that Google (http://www.google.com) and many other search engines have not publicized the algorithm that they use.

The Approach

Table 1. The Algorithm

```
Algorithm Generate_Descriptions(Page A)
{
        for each page, page_i that links to A
        {
                description by page_i = anchor text of the link to A and all that follows until
                the next hyperlink in page_i;
                score (description by page_i)=0;
        }
        for every pair of descriptions, di and dj /*(i!=j)*/
        {
                k=number of common words between di and dj
                if(k>upper limit)
                        k=upper limit;
                score (di)+=k;
                score (dj)+=k;
        }
        sort descriptions in the descending order of scores and output a couple or more (or
        less) of the best descriptions;
}
```

This Algorithm does not use any information from a page to generate a description for that page. It relies on the descriptions provided by pages that link to the page in question. A page linking to another is generally considered as an endorsement of the latter by the former. The general human tendency of a hypertext author to create links to pages of his interest in the page that he creates is exploited here. Earlier studies opine that the paragraph which starts with a URL describes the target of the URL. Hypertext authors generally do not give large descriptions of the target pages and they tend to give only as much information as will enable the user to decide on whether to go to

0-7803-8482-2/04/$20.00 ©2004 IEEE.

that page or not. As such descriptions are human authored; they can be expected to be of considerable accuracy. The algorithm uses all the text that follows the link from (and including) the anchor text till the next link as a description for the target page.

Having extracted all the information as a collection of descriptions, the next issue is to decide on a scheme to rank them. Clues from another study have been considered in ranking the obtained descriptions. The score of each description is initialized to zero. Then each description, di is compared with other descriptions dj, (i not equal to j) and the score of both di and dj are incremented by the number of common words between them or the upper limit, whichever is lesser. The upper limit enforces an upper bound on the amount of influence that can be exerted by one description on another. The upper bound should be kept high enough for the scores to reflect the reality and low enough so that contaminations do not occur. In the absence of the upper limit, two very similar large descriptions (possibly created by the same author in different pages authored by him) could boost each other's scores much beyond the reach of other descriptions. After the score computations, the best couple (or more or less) descriptions are output as the description of the page. Some sample test results follow

Table 2. Sample test Results

URL: http://www.roadahead.com/ (from 11 inward links) **Title:** Bill Gates: The Road Ahead **Description by algorithm proposed:** the road ahead the homepage of gates' 1996 book. the road ahead book by gates first published in 1996 reviews contents and information about second edition **Description by Google:** Book by Gates first published in 1996 Reviews contents and information about the second edition **Description by AltaVista:** The Road Ahead Explore The Road Ahead Kids on The Road Ahead Tools for The Road Ahead Credits Penguin Publishing This site is best viewed with Microsoft Internet Explorer 3.0, in thousands+ colors .	**URL:** http://www.keralapolitics.com/ (from 6 inward links) **Description by the algorithm proposed:** kerala politics kerala government departments kerala politics features political history parties former and current ministries poll **Description by Google:** A site on politics of Kerala, current and former governments, and events **Description by AltaVista:** Keralapolitics.com-An exclusive site on politics of Kerala,Current Ministry,FormerMinistry,Events ... Historical movements that influenced Kerala Politics, influence of regional, linguistic, religious ...

Further, it was seen that the algorithm consistently outperformed all search engines other than Google. But the results compare very well with Google. Further, the technology that Google uses has not yet been published. The algorithm worked very well with the test collections. This algorithm does not depend on the language of the web pages considered. Further, the score computation, with the pair-wise similarity computation is seen to have been successful in filtering out noisy descriptions like, "this page is bad" (which could have been inserted by people bearing some ill-will to the author of the target page). Thus it can be concluded that the assumption that pages are described better by those linking to them is seen to have worked well.