Optimizing WebPage Interest Research Proposal

Willem Elbers (0441570, w.j.m.elbers@student.ru.nl)

Information Retrieval & Information Systems (IRIS)
Radboud University Nijmegen, The Netherlands

April 2008

Abstract

In the rapidly evolving and growing environment of the internet, web site owners aim to maximize interest for their web site. How do you position a site optimally on the internet? Where does advertising generate the biggest increase in interest for your website? This document proposes a study to research these issues and develop an answer to these questions.

Keywords: webgraph; website interest; centrality; prestige; estimate website traffic

Introduction

The internet is a rapidly changing and evolving environment. The size of the internet is incredible, more than 11 billion web pages [1] exist today and this number is ever increasing (Gulli & Signorini, 2005).

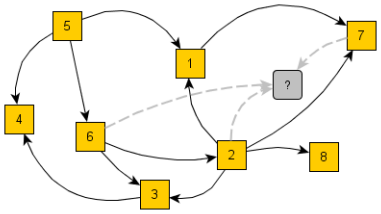
Several theories have been developed to study the internet. One of these theories models the internet as a graph, this graph is called the web graph (Broder et al., 2003). Web pages are represented as nodes in this graph and hyperlinks between pages as edges in this graph. The graph can both be directed, incorporating the direction of the hyperlink, or undirected. Since the internet is now modeled as a normal graph, all known graph theory can be applied to it. Another important aspect of the web graph to note, is that page content is not incorporated in the graph model. Nevertheless, studies have proven that clusters (in the webgraph) of websites usually are about the same topic. Discarding the actual content of the website and only looking at the graph structure, still allows us to make claims about the content of the pages.

This web graph can be used to derive all kinds of interesting properties. Several techniques exist to determine an importance value for a node (website) in the graph. This is known as the centrality of a node. In a more social context, prestige can also be used as a measure of importance (Rupnik, 2006). Identifying communities in this graph is also a well studied topic (Hinne, 2007).

Problem Description

An important issue for website owners is: how to maximize the interest for their webpage. This can be seen as the problem to optimally position a website in the webgraph. See figure 1 for an example. The grey node is the website that has to be inserted as optimal as possible. The question now is, which links are required to achieve this goal? [1]

Figure 1: Example graph.



Several website properties can be used to define what an optimal position for a website is:

- *Content*
  The content of a website has to be of high quality for it's customers. If the content does not meet the customers requirements, they loose interest for the website. But, before content comes into play, customers have to reach the website first.

- *Links*
  Links from other pages to a website define the reachability of a website. A link from a website with high traffic will increase interest more than a link from a low traffic website.

- *Advertisement*
  The more people know about a website, the more people will be interested in it. Again the traffic of the potential candidate for advertisement is important to maximize the yield of the advertising. Advertising is basically creating links to your website. However, the purpose of the advertisement links is different. Normal links are used to link-to and get-links from authority sites, where advertisement is primarily aimed at increasing (relevant) traffic to your site.

- *Other*
  There are other aspects that might improve the interest for a website, e.g. search engine rankings, but those are not considered in this research project.

[1] The term web page and website are used synonymously in this document

[1] The links in the webgraph define the position of a node in the graph

This research project will focus on combining existing theories with partially known information to maximize interest for a website.

Research Question

The main research question:

- How can we optimize the interest of a website?

Sub questions, raised from the main question:

- How can we measure the importance of a website?
- How can we incorporate, partially, unknown information to maximize the interest for a website?

Products

These products will be the final result of the research.

1. Master Thesis
   The master thesis will be a written report describing the results of the research project. The study is conducted by two members but they will both study a separate subject. This will result in a master thesis with a shared introduction and a individual section and conclusion about this research.

2. Presentation
   A presentation aimed at presenting the key achievements of the research project.

Global Planning

The following table shows the global project planning:

Task	Period
Orientation	March + April
Research: importance measure	May
Research: incorporating unknown information	June
Writing the thesis	July
Finishing + Presentation	August

The first three tasks, each yield a rough version of the corresponding section in the master thesis (the orientation task results in the project proposal, which will be the basis for the introduction in the master thesis). Writing the master thesis itself will consist of merging the available chapters, writing the conclusion and polishing the merged parts into the first version of the master thesis.

The research project yields three, solid, deliverables. A project proposal after the orientation period. A master thesis after the research and writing period and a final presentation when the project is finished.

Overview of holidays during the research project period:

Holiday	Period
Goede vrijdag	Fri 21-03-2008
Tweede Paasdag	Mon 24-03-2008
Meivakantie	Mon 28-04-2008 t/m Mon 05-05-2008
Tweede pinksterdag	Mon 12-05-2008
Diesviering	Mon 15-05-2008
Zomervakantie	Mon 14-07-2008 t/m Fri 29-08-2008

NB: We continue working on the project during the majority of these holidays.

Project Conditions

The research is conducted partly together with Frank Koopmans and the research project is supervised by Theo van der Weide.

Meetings are scheduled roughly every two weeks. In the beginning there will be more meetings, once a week, to prevent problems. During the research there will be less meetings and during the end of the project the two week interval will be used again.

During these meetings we discuss the progress and the current state of the research and try to identify possible problems. Besides these meetings any problems or questions can also be discussed with the supervisor by email.

References

Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., et al. (2003). Graph structure in the web.

Gulli, A., & Signorini, A. (2005). The indexable web is more than 11.5 billion pages.

Hinne, M. (2007). Local identification of web graph communities.

Rupnik, J. (2006). Finding community structure in social network analysis - overview.