MATHEMATICS BEHIND

SEARCH ENGINE

INTRODUCTION

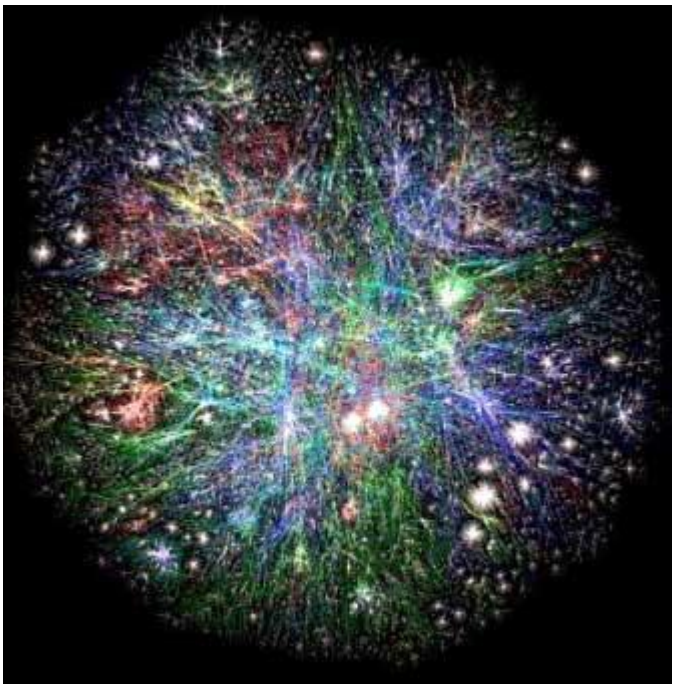
We live in a computer era. Internet is part of our everyday lives and information is only a click away. Just open your favourite search engine, like Google, AltaVista, Yahoo, type in the key words, and the search engine will display the pages relevant for your search. But how does a search engine really work?

At first glance, it seems reasonable to imagine that what a search engine does is to keep an index of all web pages, and when a user types in a query search, the engine browses through its index and counts the occurrences of the key words in each web file. The winners are the pages with the highest number of occurrences of the key words. These get displayed back to the user. This used to be the correct picture in the early 90s.

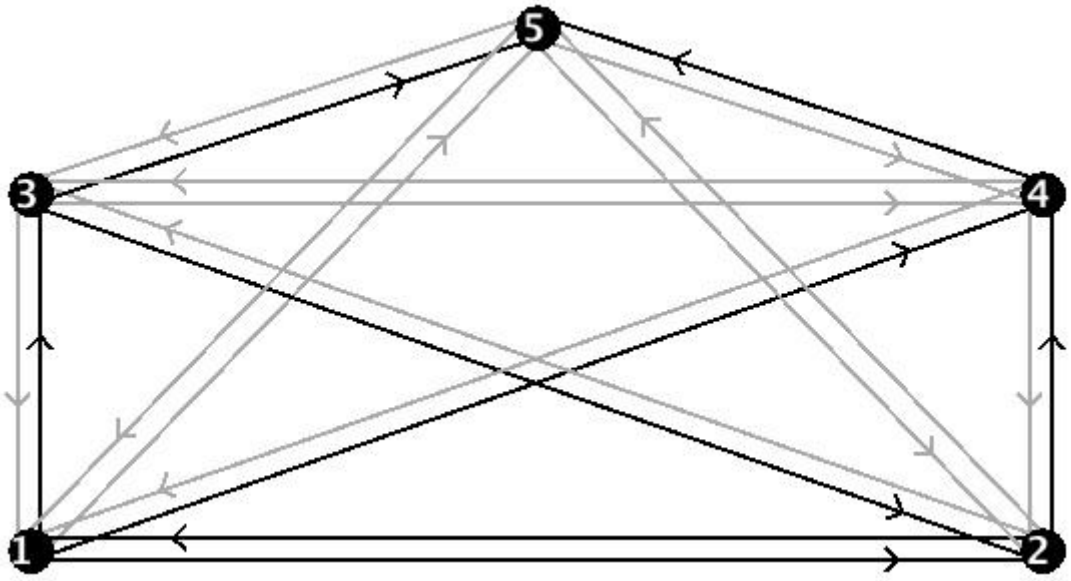
Modern search engines employ methods of ranking the results to provide the "best" results first that are more elaborate than just plain text ranking. One of the most known and influential algorithms for computing the relevance of web pages is the Page Rank algorithm used by the Google search engine. It was invented by Larry Page and Sergey Brin while they were graduate students at Stanford, and it became a Google trademark in 1998. The idea that Page Rank brought up was that, the importance of any web page can be judged by looking at the pages that link to it. If we create a web page i and include a hyperlink to the web page j, this means that we consider j important and relevant for our topic. If there are a lot of pages that link to j, this means that the common belief is that page j is important. If on the other hand, j has only one back link, but that comes from an authoritative site k, (like www.google.com, www.cnn.com, www.cornell.edu) we say that k transfers its authority to j; in other words, k asserts that j is important. Whether we talk about popularity or authority, we can iteratively assign a rank to each web page, based on the ranks of the pages that point to it.

DESCRIPTION

Google's PageRank algorithm, though mathematically elegant and very powerful, is very easy to understand. The first step to understanding PageRank is to view the Web as a giant graph. Each webpage is a node in the graph and each hyperlink is a directed link connecting two nodes.



The figure above is a graphical representation of the Web, but since there are so many pages the view gets easily obscured. Instead, let's consider a much smaller web, a web with only five WebPages as depicted below; here all black links are active and gray links are inactive, in other words, not part of the network. Notice, for instance, that webpage 5 has unlinks from webpages 3 and 4 but has no out links. Similary, webpage 2 has an in link from and an out link to webpage 1.



SKILL AND CHALLENGES

This project requires knowledge about various properties of matrix and basic knowledge about probability .Idea is to consider each web graph as matrix .so the project requires fluency in properties of matrix like eigenstates,eigenvector.

It requires study of stochastic matrix and in-depth analysis on damping factor which is used in such matrix.

will calculate page rank using matrix operation. will have consider tricky cases such as dangling node case(node with no outgoing edges),disconnected components. will study about perron-frobenius theorem, power method convergence theorem, graph theory, dynamical system, designed to answer stringent problem about how information propagates over the net.

PAST ATTEMPT

some of past attempts are listed below

Power method

Page, Brin, Motwani & Winograd 1999
Bianchini, Gori & Scarselli 2003

Surveys of numerical methods:

Langville & Meyer 2004
Berkhin 2005
Langville & Meyer 2006 (book)

Methods that adapt to web graph
Broder, Lempel, Maghoul & Pedersen 2004 Kamvar, Haveliwala & Golub 2004
Haveliwala, Kamvar, Manning & Golub 2003
Lee, Golub & Zenios 2003
Lu, Zhang, Xi, Chen, Liu, Lyu & Ma 2004
Ipsen & Selee 2006

TIMELINE

1st week: will start with study of all theorem mentioned above and properties of matrix and probability.will also make rough skeleton of different issues related to Google search engine.

2nd week: will analyse how page rank algorithm is followed for different cases and will analyse about how different features affects page rank.

3rd week: will study about how maths is related to google Earth ,google Map, gmail etc.

4th week: will study about web as a graph means the graph structure in the web.

REFERENCES

1. **Anthony Bonato, *A Course on The Web Graph*, AMS-AARMS, Graduate Studies in Mathematics v. 89, 2008. http://www.math.ryerson.ca/~abonato/webgraph.html**
2. **Kurt Bryan, Tanya Leise, *The $25,000,000,000 Eigenvector: The Linear Algebra behind Google*, SIAM Review, Vol. 48, No. 3. (2006). http://www.siam.org/journals/sirev/48-3/62328.html**

3. **Sergey Brin, Lawrence Page, *The Anatomy of a Large-Scale Hypertextual Web Search Engine*, Seventh International World-Wide Web Conference (WWW 1998). http://infolab.stanford.edu/pub/papers/google.pdf**
4. **M. Brin, G. Stuck, *Introduction to Dynamical Systems* , Cambridge University Press, 2002.**
5. **David Austin, *How Google Finds Your Needle in the Web's Haystack*. http://www.ams.org/featurecolumn/archive/pagerank.html**
6. **Jon Kleinberg, *Authoritative sources in a hyperlinked environment,* Proc. 9th ACM-SIAM Symposium on Discrete Algorithms, 1998. http://www.cs.cornell.edu/home/kleinber/auth.pdf**
7. **Internet Mathematics Journal http://www.internetmathematics.org/**
8. **http://www9.org/w9cdrom/160/160.html**