Chapter 9 of BYRN (pending, outdated)

Reading

2. Finding Relevant Pages

1. Generalities

Structure

2000–01–25

Thomas Krichel

Searching on the Web

ITRO8 Lecture 3
large size •
reoccuring data •
poor data quality •
weak structure of pages •
high volatility of data •
no overall topology •
distributed data with irregular bandwidth •

Hyper text system. However there are several guiding points:

An initial idea is that the web is nothing else than a distributed
more general problem with web indexing

return pages of poor relevance •
commercially orientated •
secret •
could not find themselves •

As need by bin and page, they include

eye search engine problems in 1997

Google links to 500 Million more but they have not been indexed.

2000 OS: both Google and Fast have about 500 Million pages.
1997-98: Altavista has 20 million queries
1995-97: Altavista has 20 million pages

The history of search engines
Bill Clinton sucks and have a picture.

Let the query be "Bill Clinton" and the page say

The "Clinton sucks" problem

is high:

\[ \frac{M}{t_u} + \frac{M}{t_w} + \frac{M}{t_u} \]

Then the document is related to the query the more

\[ \frac{M}{t_u} \]

given term \( t \) that appears \( M \) times in the document. Let the be

assumption that a document has say \( M \) words. Say there is a

often in the document.

A document is related to a query if the terms of the query appear

in information retrieval.

Most search engines started with implementing the classic term

Traditional IR and the web

then we do,

text production on the Internet (just like they play better chess

that computers will eventually be able to catch up with human

however, with computer power improving, it may be the case

search is falling.

Today it appears that the fraction of the web that is indexed by

complete would be possible.

In 1994, as the start of the web, people thought indexing it

Indexing the web, a soluble problem?
If we index the whole of the web, and we search for terms on that huge collection, we need a way to select high quality pages, and display those first.

How can that be done?

What is an important page?

A page that many pages point to is important.

A page that is pointed to by other important pages.

Model of user behavior

A random surfer is given a page.

S/he keeps clicking on any link that she finds in that page with a probability 1 – d. S/he gets bored and starts with a completely new random page with probability d.

Does that surfer hit all pages on the web with the same probability?
Interpret the formula:

\[
\left( \frac{(\text{Out}_i)}{\text{In}_i} \right)^1 + \cdots + \left( \frac{(\text{Out}_j)}{\text{In}_j} \right)^1 + \left( \frac{(\text{Out}_k)}{\text{In}_k} \right)^1 \right) \cdot p + (p - 1) = (\text{A})^1
\]

The rank of a page \( i \) is given by the number of out links to \( i \). Let there be a number of pages \( \text{A} \) that have out links to \( i \). Let there be a page \( j \), let \( p(\text{A}) \) be the number of links that go to page \( j \).