Data Mining Chapter 5
Web Data Mining

Introduction to Data Mining with Case Studies
Author: G. K. Gupta
Prentice Hall India, 2006.

Web Data Mining

Use of data mining techniques to automatically discover interesting and potentially useful information from Web documents and services.

Web mining may be divided into three categories:

1. Web content mining
2. Web structure mining
3. Web usage mining

Web details

- More than 20 billion pages in 2008
- Many more documents in databases accessible from the Web
- More than 4m servers
- A total of perhaps 100 terabytes
- More than a million pages are added daily
- Several hundred gigabytes change every month
- Hyperlinks for navigation, endorsement, citation, criticism or plain whim

Web Index Size in Pages

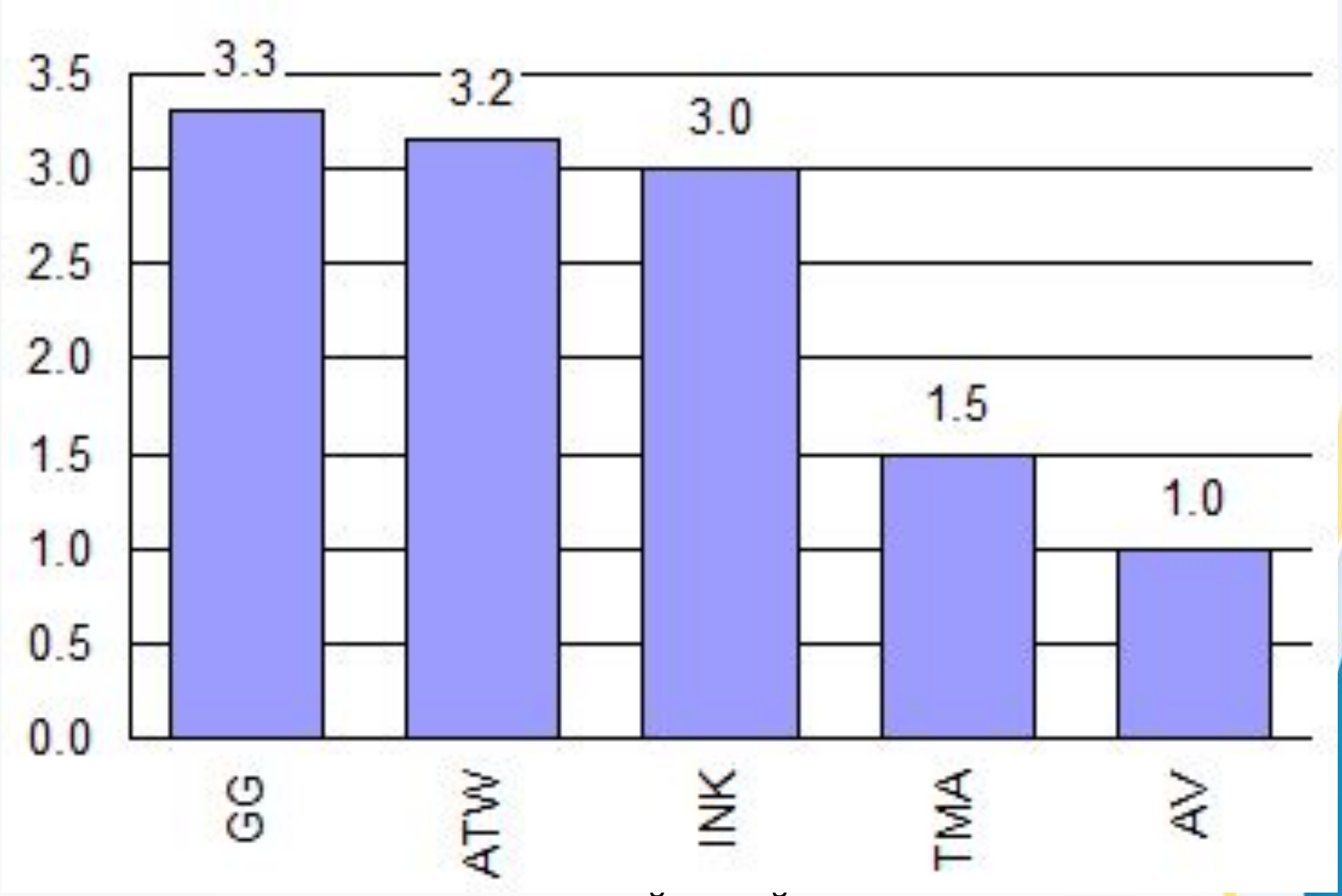
Total Web is estimated to be about 23B pages.

Search Engine	Number of Pages in Billions
Google	16
MSN Search	7
Yahoo	50
Ask	4.2

Source
http://www.worldwidewebsize.com/
20 November 2008

Size of the Web

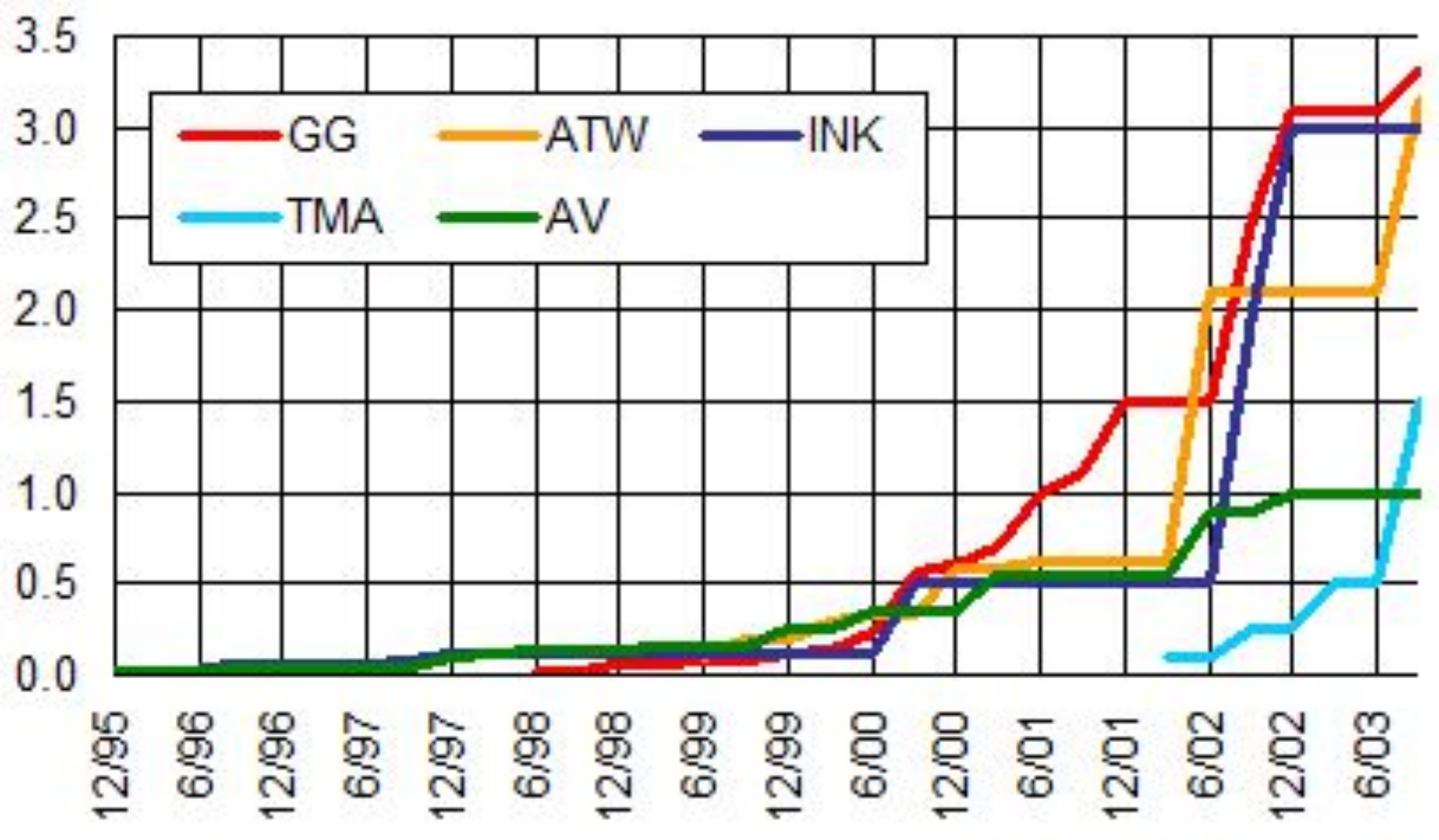
GG=Google, ATW=AllTheWeb, INK=Inktomi, TMA=Teoma, AV=AltaVista
September 2003



*

Size Trends

GG=Google, ATW=AllTheWeb, INK=Inktomi, TMA=Teoma, AV=AltaVista
September 2003



Web

- Some 80% of Web pages are in English
- About 30% of domains are in .com domain

Graph terminology

- Web is a graph – vertices and edges (V,E)
- Directed graph – directed edges (p,q)
- Undirected graph - undirected edges (p,q)
- Strongly connected component - a set of nodes such that for any (u,v) there is a path from u to v
- Breadth first search
- Diameter of a graph
- Average distance of the graph

Graph terminology

- Breadth first search - layer 1 consists of all nodes that are pointed by the root, layer k consists of all nodes that are pointed by nodes on level k-1
- Diameter of a graph - maximum over all ordered pairs (u,v) of the shortest path from u to v

Web size

- In-degree is number of links to a node
- Out-degree is the number of links from a node
- Fraction of pages with i in-links is proportional to $1/i^{2.1}$
- With i out-links, it is $1/i^{2.72}$

i	In-links	Out-links
2	23%	15%
3	10%	5%
4	5%	2%
5	3%	1%

Citations

Lotka's Inverse-Square Law - Number of authors publishing n papers is about $1/n^2$ of those with only one.

60% of all authors that make a single contribution.

Less than 1% publish 10 or more papers.

Most web pages are linked only to one other page (many not linked to any). Number of pages with multiple in-links declines quickly.

Rich get richer concept!

Web graph structure

Web may be considered to have four major components

- Central core – strongly connected component (SCC) – pages that can reach one another along directed links - about 30% of the Web
- IN group – can reach SCC but cannot be reached from it - about 20%
- OUT group – can be reached from SCC but cannot reach it - about 20%

Web graph structure

- Tendrils – cannot reach SCC and cannot be reached by it - about 20%
- Unconnected – about 10%

The Web is hierarchical in nature. The Web has a strong locality feature. Almost two thirds of all links are to sites within the enterprise domain. Only one-third of the links are external. Higher percentage of external links are broken. The distance between local links tends to be quite small.

Web Size

- Diameter over 500
- Diameter of the central core SCC is about 30
- Probability that you can reach from any node a to any b is 24%
- When you can, on average the path length is 16

Terminology

- Child (u) and parent (v) – obvious
- Bipartite graph – BG(T,I) is a graph whose node set can be partitioned into two subsets T and I. Every directed edge of BG joins a node in T to a node in I.
- A single page may be considered an example of BG with T consisting of the page and I consisting of all its children.

Terminology

- Dense BG – let p and q be nonzero integer variables and t_c and i_c be the number of nodes in T and I. A DBG is a BG where
  - each node of T establishes an edge with at least p $(1 \le p \le i_c)$ nodes of I and
  - at least q $(1 \le q \le t_c)$ nodes of T establish an edge with each node of I

- Complete BG or CBG – where $p = i_c$ and $q = t_c$.

Web Content mining

Discovering useful information from contents of Web pages.

Web content is very rich consisting of textual, image, audio, video etc and metadata as well as hyperlinks.

The data may be unstructured (free text) or structured (data from a database) or semi-structured (html) although much of the Web is unstructured.

Web Content mining

Use of the search engines to find content in most cases does not work well, posing an abundance problem. Searching phrase “data mining” gives 2.6m documents.

It provides no information about structure of content that we are searching for and no information about various categories of documents that are found.

Need more sophisticated tools for searching or discovering Web content.

Similar Pages

Proliferation of similar or identical documents on the Web.

It has been found that almost 30% of all web pages are very similar to other pages and about 22% are virtually identical to other pages.

Many reasons for identical pages (and mirror sites), for example, faster access or to reduce international network traffic.

How to find similar documents?

Web Content

- New search algorithms
- Image searching
- Dealing with similar pages

Web Usage mining

Making sense of Web users' behaviour.

Based on data logs of users' interactions of the Web including referring pages and user identification. The logs may be Web server logs, proxy server logs, browser logs, etc.

Useful is finding user habits which can assist in reorganizing a Web site so that high quality of service may be provided.

Web Usage mining

Existing tools report the number of hits of Web pages and where the hits came from. Although useful, the information is not sufficient to learn user behaviour. Tools providing further analysis of such information are useful.

May involve usage pattern discovery e.g. the most commonly traversed paths through a Web site or all paths traversed through a Web site.

Web Usage mining

These patterns need to be interpreted, analyzed, visualized and acted upon. One useful way to represent this information might be graphs. The traversal patterns provides us very useful information about the logical structure of the Web site.

Association rule methodology may also be used in discovering affinity between various Web pages.

Web data mining

- Early work by Kleinberg in 1997-98
- Links represent human judgement
- If the creator of p provides a link to q then it confers some authority to q although that is not always true
- When search engine results are obtained should they be ranked by number of in-links?

Web Structure mining

Discovering the link structure or model underlying the Web.

The model is based on the topology of the hyperlinks. This can help in discovering similarity between sites or in discovering authority sites for a particular topic or discipline or in discovering overview or survey sites that point to many authority sites (called hubs).

Authority and Hub

- A hub is a page that has links to many authorities
- An authority is a page with good content on the query topic and pointed by many hub pages, that is, it is relevant and popular

Kleinberg's algorithm

- Want all authority pages for say "data mining"
- Google returns 2.6m pages, may not even include topics like clustering and classification
- We want s pages such that
  1. s is relatively small
  2. s is rich in relevant pages
  3. s contains most of the strongest authorities

Kleinberg's algorithm

- Conditions 1 and 2 (small and relevant) are satisfied by selecting 100 or 200 most highly ranked from a search engine
- Needs to use an algorithm to satisfy 3

Kleinberg's algorithm

- Let R be the number of pages selected from search result
- S = R
- For each page in S, do steps 3-5
- Let T be the set of all pages pointed by S
- Let F be the set of all pages pointed to S
- Let S := S + T + some or all of F
- Delete all links with the same domain name
- Return S

Kleinberg's algorithm

- S is the base for a given query – hubs and authorities are found from it
- One could order the pages in S by in-degrees but this does not always work
- Authority pages should not only have large in-degrees but also considerable overlap in the sets of pages that point to them
- Thus we find hubs and authorities together

Kleinberg's algorithm

- Hubs and authorities have a mutually reinforcing relationship: a good hub points to many good authorities; a good authority is pointed to by many good hubs.
- We need a method of breaking this circularity

An iterative algorithm

- Let page p have an authority weight x_p and a hub weight y_p
- Weights are normalized, squares of sum of each x and y weights is 1
- If p points to many pages with large x weights, its y weight is increased (call it O)
- If p is pointed to by many pages with large y values, its x weight is increased (call it I)

An iterative algorithm

- On termination, the pages with largest weights are the query authorities and hubs

Theorem: The sequences of x and y weights converge.

Proof: Let A be the adjacency matrix of G, that is, (i,j) entry of A is 1 if (pi, pj) is an edge in G, 0 otherwise.

I and O operations may be written as

$$x \leftarrow A^T y$$

$$y \leftarrow A x$$

Kleinberg's algorithm

- I and O operations may be written as
  - $x \leftarrow A^T y$
  - $y \leftarrow A x$
- x_k is the unit vector given by
  - $x_k = (A^T A)^{k-1} x_0$
- y_k is the unit vector given by
  - $y_k = (AA^T) y_0$
- Standard Linear algebra shows both converge.

Intuition

A page creator creates the page and displays his interests by putting links to other pages of interest. Multiple pages are created with similar interests. This is captured by a DBG abstraction.

A CBG would extract a smaller set of potential members although there is no guarantee such a core will always exist in a community.

Intuition

A DBG abstraction is perhaps better than a CBG abstraction for a community because a page-creator would rarely put links to all the pages of interest in the community.

Web Communities

- A Web community is a collection of Web pages that have a common interest in a topic
- Many Web communities are independent of geography or time zones
- How to find communities given a collection of pages?
- It has been suggested that community core is a complete bipartite graph (CBG) –can we use Kleinberg's algorithm?

Discovering communities

- Definition of community

  A Web community is characterized by a collection of pages that form a linkage pattern equal to a DBG(T, I, α, β), where α, β are thresholds that specify linkage density.

Discovering communities

- Assert that Web communities are characterized by a collection of pages that form a linkage pattern equal to DBG
- A seed set is given
- Apply a related page algorithm (e.g. Kleinberg) to each seed, derive related pages

cocitation

- Citation analysis is study of links between publications
- Cocitation measures the number of citations in common between two documents
- Bibliographic coupling measures the number of documents that cite both of the two documents

cocite

- Cocite is a set of pages is related, if there exist a set of common children
- n pages are related under cocite if these pages have common children at least equal to cocite-factor
- If a group of pages are related according to cocite relationship, these pages form an appropriate CBG.