

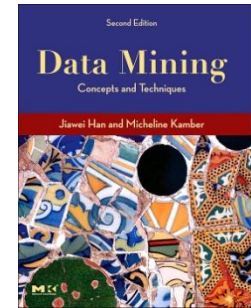


Web Mining

Data Mining and Text Mining (UIC 583 @ Politecnico di Milano)

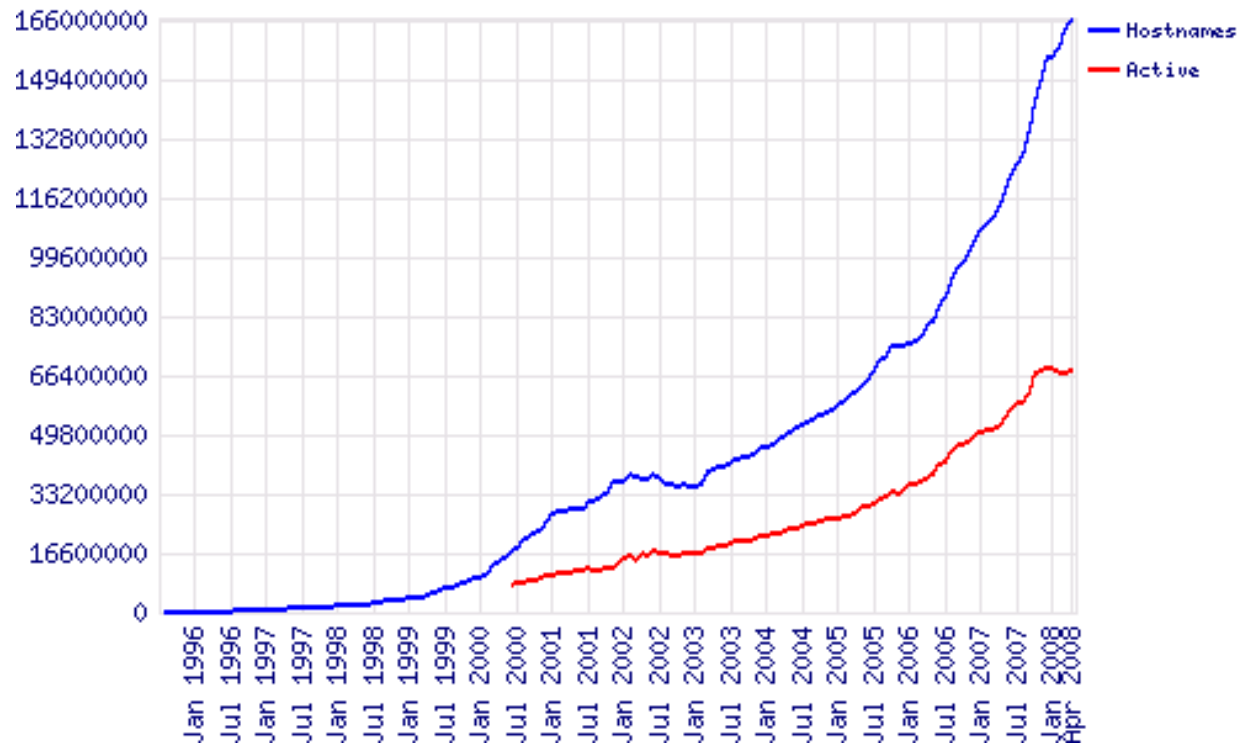
References

- Jiawei Han and Micheline Kamber, "Data Mining: Concepts and Techniques", The Morgan Kaufmann Series in Data Management Systems (Second Edition)
 - ▶ Chapter 10
- **Web Mining Course by *Gregory-Platesky Shapiro*** available at www.kdnuggets.com
- **Federico Facca and Pier Luca Lanzi. Mining Interesting Knowledge from Weblogs: A Survey. *Journal of Data and Knowledge Engineering*, 53(3):225–241, 2005.**



How big is the Web?

165,719,150 Web Sites @Apr 2008 (Netcraft Survey)



Discovering interesting and useful information from Web content and usage

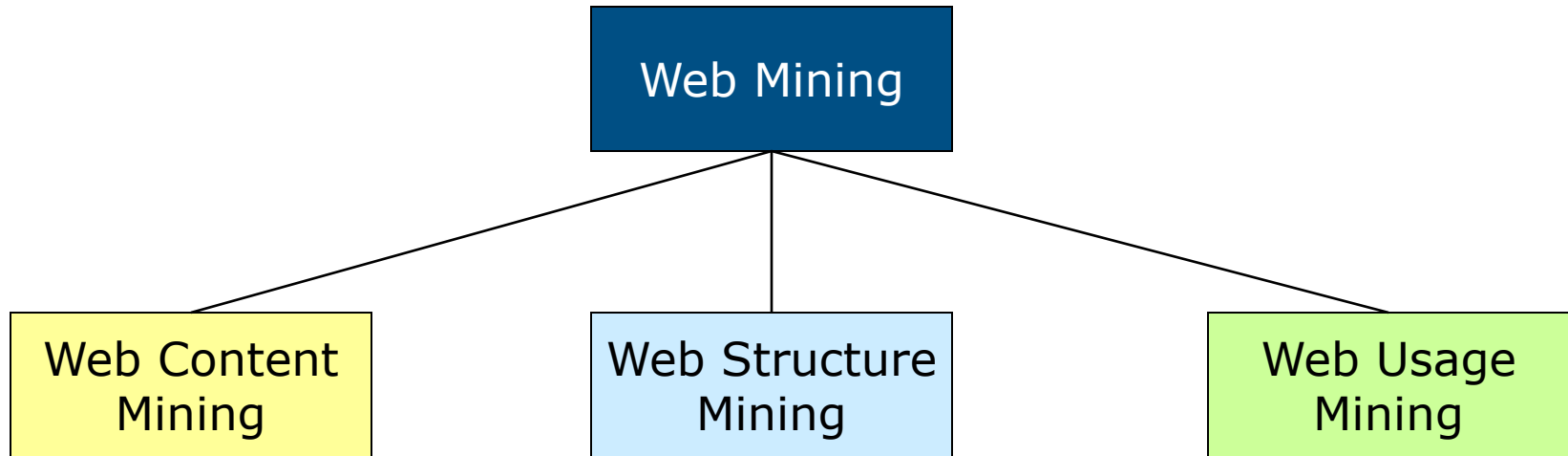
□ Examples

- ▶ Web search, e.g. Google, Yahoo, MSN, Ask, ...
- ▶ Specialized search: e.g. Froogle (comparison shopping), job ads (Flipdog)
- ▶ eCommerce
- ▶ Recommendations (Netflix, Amazon, etc.)
- ▶ Improving conversion rate: next best product to offer
- ▶ Advertising, e.g. Google AdSense
- ▶ Fraud detection: click fraud detection, ...
- ▶ Improving Web site design and performance

Web Mining Challenges

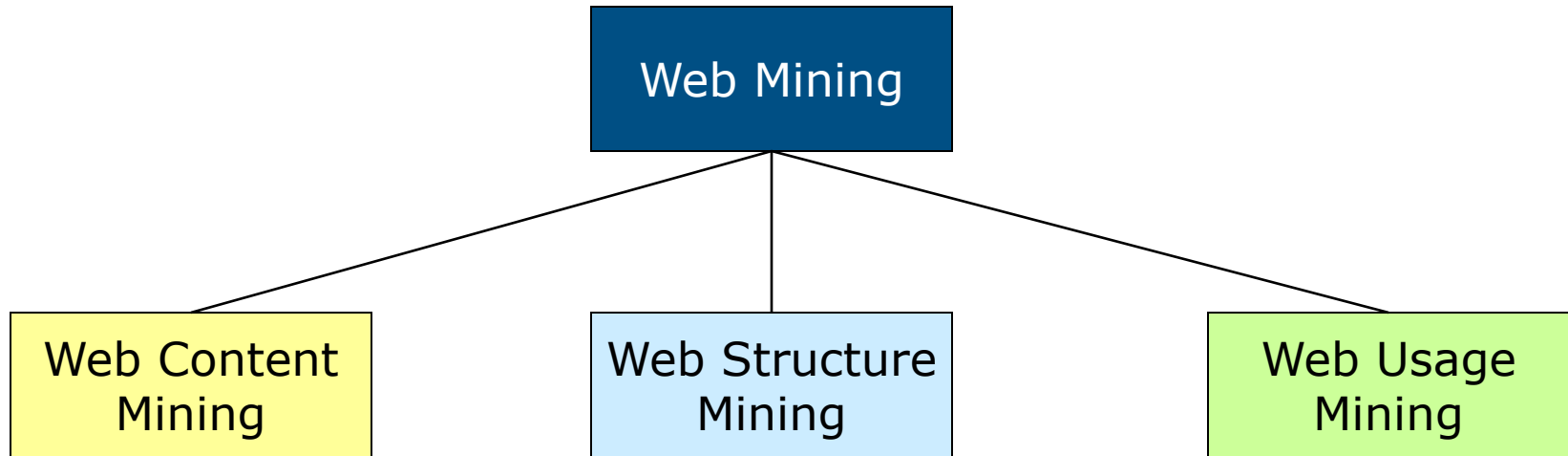
- ❑ Huge amount of data
- ❑ Complexity of Web pages
 - ▶ Different styles
 - ▶ Different contents
- ❑ Highly dynamic and rapidly growing information
 - ▶ Number of sites is rapidly growing
 - ▶ Information is constantly updated
- ❑ Web serves many user communities
 - ▶ Users with different interests, background and purposes
 - ▶ “99% of the Web information is useless to 99% of Web users”

Web Mining Taxonomy



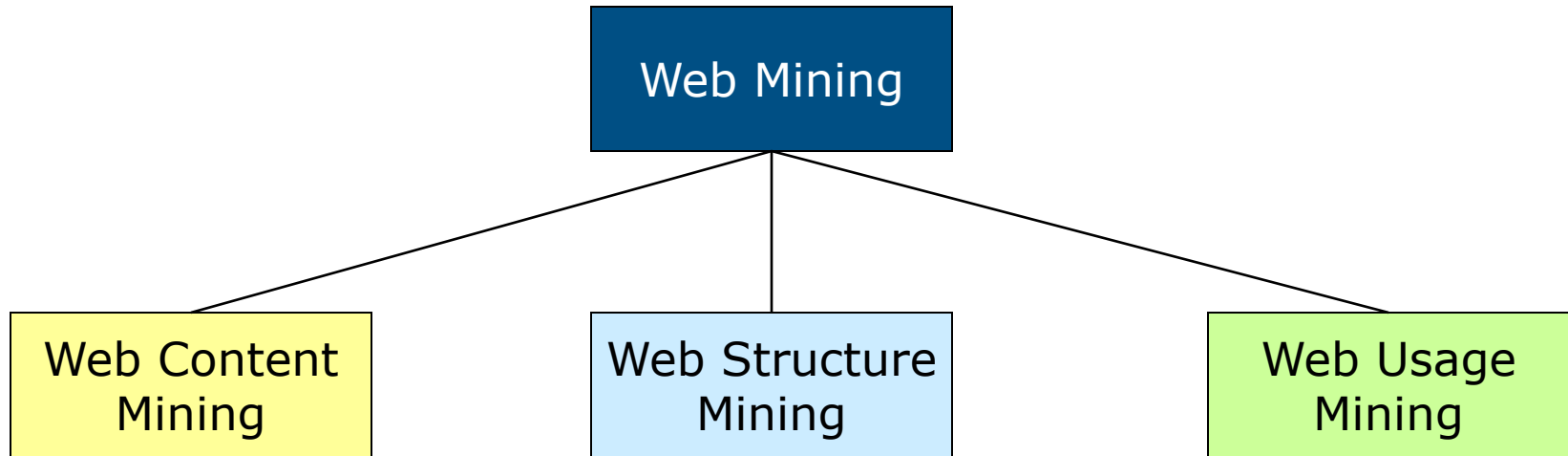
- Summarization of Web pages
- Summarization of Web searches
- Mining multimedia Web content
- Web pages classification
- ...

Web Mining Taxonomy



- Mining linking structure
- Discover authoritative pages
 - ▶ PageRank
- Discover hub

Web Mining Taxonomy



- ❑ Mining weblogs to discover usage patterns
- ❑ Applications:
 - ▶ Personalization of Web content
 - ▶ Improve Web design

- ❑ Web page is more than plain text
- ❑ Web page structure is defined by the **DOM** (Document Object Model) tree, where nodes are the **HTML tags**
- ❑ **Issues**
 - ▶ Not all the pages follows the standards
 - ▶ DOM tree does not always reflect the page semantic

Mining Web Page Layout Structure

- ❑ Web
- ❑ Web
- ❑ Mode
- ❑ Issue
- ▶ N
- ▶ D

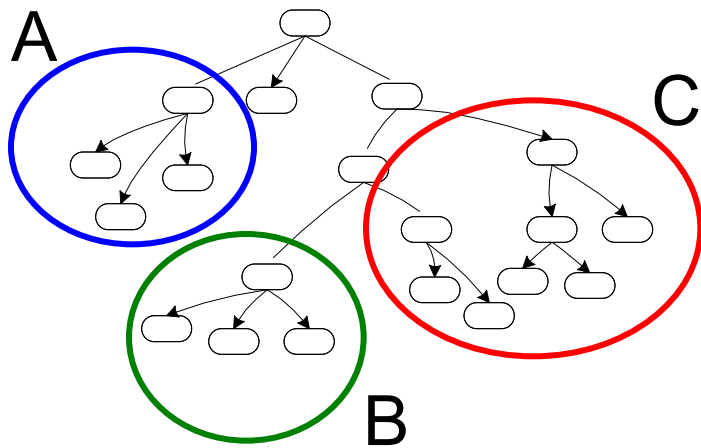
The screenshot shows a web browser window titled "Page Analysis - Yahoo!igans! E-Cards" with the URL `http://ecards.yahooigans.com/content/ecards/category?c=133&g=16`. The page content includes the "Yahooligans! E-Cards" logo, a navigation bar, and a grid of animal-themed e-cards. On the right side, a DOM tree and an attribute table are visible. Two yellow arrows originate from the DOM tree: one points to a `TR` element at index 195, and another points to a `TR` element at index 200. These elements correspond to the "Elephant Sunrise" and "Prowling Fox" e-cards, respectively, which are highlighted with red boxes in the browser view.

Attribute	Value
tagName	TR
sourceIndex	195
outerHTML	<TR style="..."
innerText	
innerTextLen	9
Left	10
Top	692
offsetLeft	0
offsetTop	440
offsetWidth	620
offsetHeight	84
currentStyle...	transparent
currentStyle.f...	12pt
currentStyle.f...	normal
currentStyle.f...	400
currentStyle.z	0

ent Object

ntic

Vision-based Page Segmentation

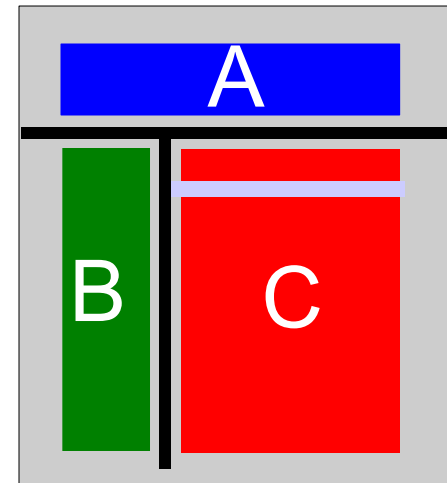


DOM tree

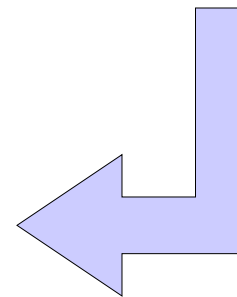
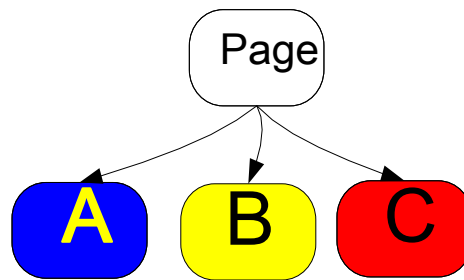
Visual Block
Extraction



Visual Separator
Detection



Page Layout



Example of Web Page Segmentation

Page Analysis - Yahoo!igans! E-Cards
http://ecards.yahooigans.com/content/ecards/category?c=133&w=16

YAHOO!IGANS! E-Cards
Home > Yahoo!igans! E-Cards > Send an E-Card

Animals

1 Choose a Card 2 Address the Card 3 Choose a Message 4 Preview/Send Card

Just a Hello From My Dearhouse? Woohoo! A Bit Calm

Frank Situation Lurch Animals? Cubs How's Your Day

Leon King Cheekah Family Leopard Mabe

Timber Wolf Giraffe Elephant Subote Prowling Fox

Attribute	Value
tagName	TR
sourceIndex	195
outerHTML	<TR style="..."
innerText	
innerTextLen	9
Left	10
Top	692
offsetLeft	0
offsetTop	440
offsetWidth	620
offsetHeight	84
currentStyle...	transparent
currentStyle f...	12pt
currentStyle f...	normal
currentStyle f...	400
currentStyle...	0

(DOM Structure)

Page Analysis - Yahoo!igans! E-Cards
http://ecards.yahooigans.com/content/ecards/category?c=133&w=16

YAHOO!IGANS! E-Cards
Home > Yahoo!igans! E-Cards > Send an E-Card

Animals

1 Choose a Card 2 Address the Card 3 Choose a Message 4 Preview/Send Card

Just a Hello From My Dearhouse? Woohoo! A Bit Calm

Frank Situation Lurch Animals? Cubs How's Your Day

Leon King Cheekah Family Leopard Mabe

Timber Wolf Giraffe Elephant Subote Prowling Fox



Attribute	Value
tagName1	TD
innerHTML1	<A href="ad..."
innerText1	
textLength1	1
tagName2	TD
innerHTML2	<FONT face="..."
innerText2	Prowling Fox
textLength2	13
FrameSource...	0
SourceIndex	209,227
TightDegree	9
Containing	-1
FontSize	12
FontWeight	400
ObjectRectLeft	486
ObjectRectTop	692

(VIPS Structure)

Mining Web's Link Structure

- ❑ How to identify **authoritative** page?
- ❑ The answer is in the **Web linkage structure**
- ❑ Issues in Web linkage mining
 - ▶ Links do not always represent endorsements (e.g., adv)
 - ▶ Important competitors do not usually link each other
 - ▶ Authoritative pages are generally not self-descriptive
- ❑ To discover authorities we should also look for **hub pages**
 - ▶ Hub are pages that provide **collections of links to authorities**
 - ▶ Hub pages are not necessary highly linked
 - ▶ Hub pages implicitly confer authorities on focused topics
- ❑ **Hub and authoritative pages have a mutual reinforcement relationship**
 - ▶ A good hub page points to many good authorities, a good authority is a page pointed by many good hub pages

Examples

Log In · Register · Change Country 
OPEN HAPPINESS ▶ My Coke Rewards Shop Our Store Coke in the USA ▼

Language

- Local
- Azərbaycan
- বাংলা
- Български
- Бразилски
- Бразилски (портгалски)
- Català
- Česky
- Cymraeg
- Dansk
- Deutsch
- Dänmarks
- Español
- Esperanto
- עברית
- Français
- Furlan
- Galego
- Hrvatski
- Bahasa Indonesia
- Italiano
- Italiano
- עברית
- Jawa Jend
- Azərbaycanca
- Lietuvių
- Македонски
- Nederlands
- Norsk
- Norsk (bokmål)
- Norsk (nynorsk)
- Polski
- Português
- Руски
- Pycckий
- Česky
- Simple English
- Srpskohrvatski / Српскохрватски
- Suomi
- Svenska
- Tagalog
- తెలుగు
- Türkçe
- Українська
- Tiếng Việt
- Winaray
- Zimwabiša

Brands

This section does not cite any references or sources.
Please help improve this article by adding citations to reliable sources. Unsourced material may be challenged and removed. (January 2006)

The cola brands with the greatest global volume are Coca-Cola and Pepsi.

Azla

- **Azla Cola** is a brand of cola in Israel, marketed specifically to Haredim. The kosher emblem is as large and prominent as the brand name.
- **Mexico Cola**, an Arab activist beverage, was sold in the Middle East, parts of Europe and North Africa. (Discontinued)
- **Pakola** is a popular beverage from Pakistan.
- **PC Cola** was popular in the Philippines with its franchisee **Asawa Beverages**. PC was introduced to Israel in 1965 with the slogan "PC: Just like in America!" It is now available in Bangladesh.
- **Zaino Cola** was not so popular in the Philippines with the slogan "Zep Cola is no zain!"
- **Star Cola** is a brand from Gaza/Palestine. However, there is also "Star Cola" in Myanmar. It is the most popular brand for cola in Myanmar since its introduction in 1997.
- **Super Drink**, is a popular cola in the Palestinian Territories and the State of Israel.
- **Thambi Upi** is a popular cola brand in India.
- **Compa Cola** was India's most popular brand prior to the introduction of Pepsi and Coca-Cola to the Indian market in 1991.
- **Zain Zain Cola**, popular in Iran and parts of the Arab world.
- **Parsi Cola**, popular in Iran.
- **Red Bull Cola**, popular in Thailand.
- **myCola**, popular in Sri Lanka, is distinctly sold in small plastic bottles (the major other colas most widely available in glass bottles)
- **myCola** was popular in South Korea in 1990s.

Europe

- **Afri-Cola**, a German brand, had a higher caffeine content (about 25 mg/L) until the product was relaunched with a new formulation in 1999, and has it again since a second relaunch with the original formulation in April 2005.
- **Afrit Cola** is the native cola in Catalonia, (Spain).
- **American Cola** and **Adra Cola** are the local drinks in Romania.
- **Bier Cola** made by A.G. Bier (the makers of the popular (or Bier) drink) in the United Kingdom.
- **Bright Cola** is a local brand from Brittany, (France) it offers different and unique flavors like a citrusy orange and an original cola taste.
- **Cooldia** is a local brand from former Yugoslav, originally produced by Slovenska company from Slovenia (then part of a Yugoslav). A couple of years ago it was bought by Orpa Kolinska. It is still popular in former Yugoslav republics, especially in Slovenia.
- **Cola Cola** in Albania.
- **Cola Turka** and **Lu Cola** are two local brands in Turkey.
- **Cola Cola** is the native cola of Sweden.
- In Denmark, the native **Jolly Cola** was more popular than Coca-Cola and Pepsi-Cola during the 1990s and 70s.
- **Inta-kola** is a cola soft drink from Hamburg, Germany. It uses the highest possible concentration of caffeine for beverages allowed by German law (25 mg / 100ml) and is available in most of Germany, as well as parts of western and central Europe.
- **Prosecco-cola** is a product of UOOL.
- **Yaroslava Cola** is a native cola of Denmark.
- **Kofola** is the third best selling soft-drink, in Czech and Slovak, behind Coca-Cola and Pepsi.
- **Open brand cola** is available in economy, standard and premium forms at all Tesco, Asda, Sainsbury's and Morrisons supermarkets.
- **Red Bull Cola** has been available throughout Europe since 2003.
- **Ukanta Cola** is a flavoured cola from the United Kingdom available in parts of Western Europe.
- **Virgin Cola** was popular in the South Africa and Western Europe in the 1990s but has waned in availability.
- **Viva-Cola** is a German cola brand with a distinct citrus flavor; nowadays it is mostly sold in eastern Germany.

North America

- **Royal Crown (RC Cola)** is widely available in the United States, Canada, Mexico, and Bangladesh.
- **Big Cola (Big Cola)** is sold in the northern parts of Mexico.
- There is also an open source recipe for a cola drink, **OpenCola**.
- **LuCola** and **Triangolo** are brands from Cuba (also sold widely in Italy).
- **Janos Soda** also makes a cola, using cane sugar.
- **Jal Cola** is sold by **Vital Plant Beverages**, of Rochester, New York. Originally, the slogan was "All the sugar and twice the caffeine." They dropped the slogan when they switched from cane sugar to high fructose corn syrup.
- **Johnnie Ryan** is a regional cola bottled in Niagara Falls, New York. Established in 1929, they make it with 100% cane sugar and also sell 22 other flavors.

South America

- **Inca Kola** is another brand that is now marketed in many countries by the Coca-Cola group; it is the major cola in some South American countries. This bright yellow carbonated beverage is especially popular in Peru, which was once the heartland of the Inca (or Inka) Empire. Inca Kola was only recently bought by Coca-Cola.
- **Schen Cola** is a variety of cola produced in Brazil by **Primo Schenck**.

Etymology



A can of generic brand Cola



A green glass bottle with Anadolulda Coca-Cola, seen in the Palestinian market.

Hyperlink-Induce Topic Search (1)

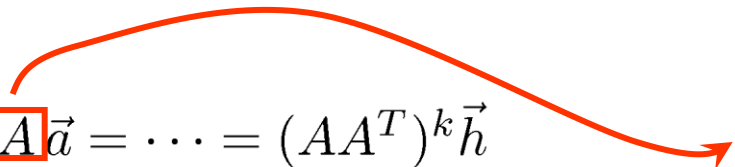
□ Startup

- ▶ **Root set** built from results from an index-based search engine
- ▶ **Base set** built including pages linked by and linking to the root set pages

□ Authority weight, a_p , and hub weight, h_p , are iteratively computed

$$a_p = \sum_{\forall q:q \rightarrow p} h_q \qquad h_p = \sum_{\forall q:q \leftarrow p} a_q$$

□ In matrix form

$$\begin{cases} \vec{h} = \mathbf{A} \vec{a} = \dots = (\mathbf{A} \mathbf{A}^T)^k \vec{h} \\ \vec{a} = \mathbf{A}^T \vec{h} = \dots = (\mathbf{A}^T \mathbf{A})^k \vec{a} \end{cases}$$


Adjacency Matrix

□ The **authority weight vector** and the **hub weight vector** if normalized converge to the eigenvectors of $\mathbf{A} \mathbf{A}^T$ and $\mathbf{A}^T \mathbf{A}$

Hyperlink-Induce Topic Search (2)

- Underlying assumptions:
 - ▶ Links convey endorsement
 - ▶ Pages co-linked by a certain page are likely to be related to the same topic
- VIPS-based approach
 - ▶ **Block-to-page** relationship

$$Z_{ij} = \begin{cases} 1/s_i, & \text{if block } i \text{ point to page } j \\ 0, & \text{otherwise} \end{cases}$$

where s_i is the number of pages linked by block i

- ▶ **Page-to-block** relationship

$$X_{ij} = \begin{cases} f_{p_i}(b_j), & \text{if } b_j \in p_i \\ 0, & \text{otherwise} \end{cases}$$

where $f_p(b)$ represents how b is important in page p

- ▶ Adjacency matrix can be defined as

$$W_P = XZ$$

Hyperlink-Induce Topic Search (3)

The screenshot shows the CNN International website interface in Microsoft Internet Explorer. The browser address bar shows 'http://edition.cnn.com/'. The page features a navigation menu on the left, a main content area with several news stories, and a 'WORLD BUSINESS' section at the bottom. A search bar is located at the top right of the page content.

Home Page
November 12, 2003 -- Updated 0136 GMT (0936 HKT)

IAEA: Iran had secret nuke agenda
The International Atomic Energy Agency has concluded that Iran has secretly produced small amounts of nuclear materials, including low-enriched uranium and plutonium that could be used to develop nuclear weapons, according to a confidential report obtained by CNN.

EXPLOSIONS ROCK BAGHDAD
Mortars strike the heavily fortified site of the coalition HQ in Iraq. [Full Story](#) | [Video](#) | [Coalition casualties](#) | [Bush hails sacrifice](#)

MORE TOP STORIES

- [Al Qaeda strategy shift: Experts](#) | [London 'target'](#)
- [Saudi bomb suspects questioned](#) | [Video](#)
- [Tension ahead of Bush's UK visit](#) | [Poll criticizes president](#)
- [Millionaire not guilty of murder](#) | [Video](#)
- [Berlusconi heads for soccer clash](#)
- [Move to expel anti-Semitic slur MP](#)
- [Japan leads Asian recovery](#) | [Small losses on Wall St.](#)
- [Vietnam uncovers 7th century ruins](#)
- [Rock star Van has to pay the man](#)

WORLD BUSINESS

MARKETS: updated 0140 GMT

MARKET	CHANGE	VALUE	%
NIKKEI	+76	10283	+0.8%
H.SEN	-153	12003	-1.3%
FTSE	+3	4345	+0.1%
DAX	-16	3729	-0.4%

ROYAL SPOOF | **HIGH ANXIETY** | **EYE ON CHINA**

Importance = Low

Importance = Med

Importance = High

- ❑ Is different from general-purpose multimedia data mining
 - ▶ Multimedia data is embedded in Web pages
 - ▶ Links and surrounding text might help the data mining process
- ❑ VIPS algorithm is the basis to extract knowledge
 - ▶ A **block-to-image** relationship can be built
 - ▶ The block-to-image relationship can be integrated with a block-level link analysis
 - ▶ The resulting **image graph** reflect the semantic relationship between the images
- ❑ The image graph can be used for classification and clustering purposes

Web usage mining is the extraction of interesting knowledge from server log files

□ Applications

- ▶ Mining logs of a single user
 - Web content personalization
- ▶ Mining logs of groups of users
 - Supporting Web design

□ Issues

- ▶ Where is the data?
- ▶ How to preprocess the data?
- ▶ Which mining techniques?

- ❑ Logs can be collected at different levels
 - ▶ Server side
 - ▶ Proxy side
 - ▶ Client side

Data sources: server side

- ❑ Web server log
 - ▶ Standard format (e.g., LogML)
 - ▶ Large amount of information (IP, request info, etc.)
 - ▶ User session can be difficult to identify
 - ▶ Special buttons (e.g., *Back*, *Stop*) cannot be tracked
- ❑ TCP/IP packet sniffer
 - ▶ Data collected in real-time
 - ▶ Data from different web servers can be merged easily
 - ▶ Some special buttons can be tracked (e.g. *Stop*)
 - ▶ Does not scale very well
- ❑ Exploiting the server application layer
 - ▶ Very effective
 - ▶ Not always possible
 - ▶ Requires ad-hoc solutions for each web server

Data sources: proxy side

- ❑ Almost the same information available on server side
- ❑ Data of **groups of users** accessing to **huge groups of web servers**
- ❑ Sessions can be anyway identified

Data sources: client side

- Collecting data with JavaScript or Java applets
- Exploiting a modified Web browser
- Perfect identification of the user session
- Requires user collaboration

- ❑ Data cleaning consists of removing from Web logs useless data for mining purposes
- ❑ Content requests (e.g. images) are usually easily removed
- ❑ Robots and Web spiders should be removed on the basis of
 - ▶ Remote hostname
 - ▶ Access to robots.txt
 - ▶ Navigation pattern

Preprocessing: session identification and reconstruction

□ Goals

- ▶ Identifying the session of different users
- ▶ Reconstruction the navigation path in identified session

□ Challenges

- ▶ Proxy
- ▶ Browser caching and special buttons

□ Solutions

- ▶ Cookies
- ▶ URL rewriting
- ▶ JavaScript (e.g. SurfAid)
- ▶ Consistency of navigation path
- ▶ Timeout heuristic for session termination

- ❑ Personalization of Web content
 - ▶ Behavior anticipation
 - ▶ Recommendation of interesting links
 - ▶ Content reorganizations
- ❑ Pre-fetching and caching
 - ▶ Caching and pre-fetching of content to reduce the server response time
- ❑ Support to Web design
 - ▶ Analysis of frequent patterns to improve the usability of Web sites
- ❑ E-commerce
 - ▶ Analysis of customer behaviors (attrition, fidelity, etc.)

- ❑ Generally URLs are the only information available on pages
- ❑ A richer information about visited pages may help the discovering of interesting Web usage patterns
- ❑ Main approaches
 - ▶ Pages categorization
 - Pre-defined
 - Automatically discovered with Web mining techniques
 - ▶ Semantic Web for Web Usage Mining
 - Ontology mapping
 - Learning of ontology from data
 - Extraction of concept-based navigation paths

- The main techniques used for the analysis of collected data are

- ▶ Association rules

$A.html, B.html \Rightarrow C.html$

- ▶ Sequential patterns extraction

- General purpose algorithm (e.g., AprioriAll)
- Ad hoc solution for Web logs (WAP-mine)

- ▶ Clustering of sessions

- Based on sequence alignment
- *Association rule hypergraph partitioning*
 - build a graph representing frequent patterns
 - Edges weighting based on pattern relevance
 - Partitioning of graph to extract users' behaviors