



visionnext

誠信、創新、服務、價值

Information Retrieval & Digital Archive Management

陳俊良/新視科技

2005 -12 - 20

不斷創 **新視** 野無限

簡報大綱



- Company and self-introduction
 - Digital Trend
 - Search Technology Introduction
 - Google Advanced Search Technology
 - Map & Reduce for scalability
 - File System
 - Page Ranking
 - Search Result Clustering
 - P2P Search
 - Multimedia Search (option for image & video)
 - Digital Archive (數位典藏)
 - Content Value Chain Concept
 - Q&A
-
-

Speaker Profile



● 陳俊良

- 1972 生於台灣高雄
- 1994 台大資訊工程系畢
- 1996 役畢
- 1997 中研院-台大語音實驗助理: 金聲三號系統介面開發
- 1998 Best Student Paper Award from International Symposium on Chinese Spoken Language Processing (ISCSLP)
- 1998 微軟MCSE (Microsoft Certificate System Engineer)認證
- 1999 台大資訊系研究所畢
- 1999 新視科技Co-founder & Technical VP
- 2000 新視科技推出全球華人第一套網路圖片搜尋引擎
- 2002 新視數位典藏事業部負責人
- 2002 政大第一屆MBA網路碩士學分班
- 2003 參與故宮數位典藏系統建置計畫
- 參與中華電信訓練所數位學習網建置計畫
- 擔任經濟部工業局文化數位內容人才養成班”數位學習”講師
- 2004 參與臺博館數位典藏管理系統
- 文化創意協會監事
- 參與師大歷史博物館新視三方創意加值計畫
- 參與中華電信資訊安全稽核系統
- 2005 參與達志電子商務圖庫管理系統、電影資料館典藏管理系統建置

● 成立時間

- 西元 1999年

● 主要投資者

- 宏碁創投、PC Home、及創業技術團隊等

● 營運據點

- 台北 / 北京

● 核心技術與產品

- 多媒體搜尋引擎技術服務入口網站
- 延生出數位典藏管理系統及網路不當資訊管理產品

● 企業文化精神

- 以「誠信、創新、服務」的企業文化精神，做為對客戶的承諾與產品發展方向，進而為客戶創造「價值」



技術服務



● 多媒體搜尋

- 巨量搜尋引擎
- 內涵檢索
 - 以圖找圖
 - 語音辨識

● 多媒體內容管理

- 網頁網站分類
 - 動態快速
- 個人相簿典藏
- 典藏品管理



● 提供建置數位典藏的經驗與技術

- 客戶：故宮、文建會、科博館、歷史博物館、臺灣博物館、桃園縣文化局、台南縣文化局、達志正片....

What are You Searching for ?



數位時代來臨



- 數位內容產業
 - 未來台灣在華人世界的機會
 - Content is the king
 - 內容輸出=~文化輸出
- 多媒體時代來臨
 - 網路頻寬逐漸提昇, 通訊成本逐漸下降
 - skype
 - 數位家電日漸普及, 數位內容快速成長
 - Xbox360、PDA phone、home media center
 - 網路多媒體素材比重逐步增加
 - 寬頻+多媒體=多媒體網路時代
 - 網路應用正面臨286->386新一波的轉變
 - 量變--> 質變



新一波的資訊革命



● 資訊傳播的途徑

- 電話、書報雜誌、電視媒體、書信、口語、.....
- 網路 (新興媒體): 資訊檢索從網路開始

● 數位化的概念 - -> 數位時代的來臨

- 電腦普及、數位設備漸興(掃描器、數位相機、DV...等)
- 電子檔案的存在形式易於傳播交換
- 體積漸小、利用頻率增加

● 網路化

- 各種異質設備的相互連結 (手機、PDA、IA家電...)

● 多媒體化

- 以多媒體的文件格式存在,包含文字, 聲音 影像, Video視訊等
 - 資訊爆炸
-
-

國家重點發展計畫

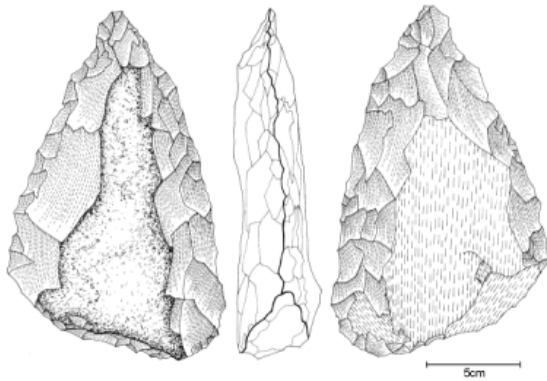


- 產業升級：
 - 勞力密集產業
 - 技術密集產業
 - 數位內容產業 (知識經濟產業)
 - 數位典藏，數位學習，KM……
 - 數位落差
 - 數位台灣 (e-Taiwan)
 - 寬頻到府600萬用戶計畫
 - <http://www.nici.nat.gov.tw/>
 - 數位典藏國家型科技計畫
 - <http://www.ndap.org.tw/>
 - 數位學習內容國家型計畫
 - <http://elnp.ncu.edu.tw/>
 - 數位內容產業計畫
 - E世代人才培育計畫
-
-

資訊檢索

Information Retrieval

Information Age



Google™



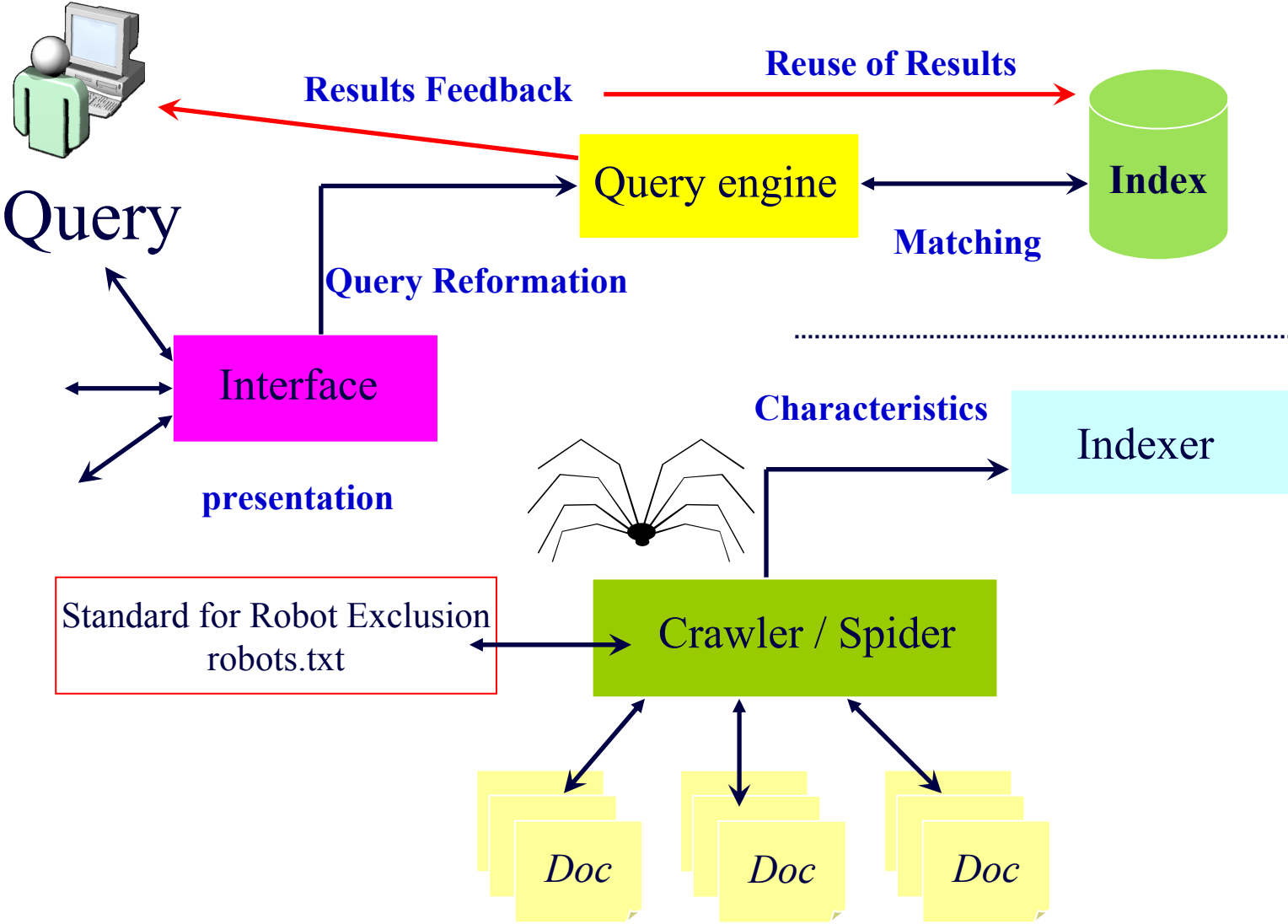
Information age: there is a special system between man and information, how should we call this epoche?

檢索環境特性比較



環境	Internet	Intranet
內容資訊量	大（巨量）	相對小
資料結構	較鬆散，不易管理 Non-structural	較複雜，具結構性 Structural
使用語言	HTML	XML/ DTD/ RDF
應用	搜尋引擎/ 入口網站	數位典藏 / 數位學習
檢索技術	跨語檢索 / 語意檢索 / Clustering	多媒體檢索 / 以圖找圖 Content-based Search
資訊組織	網站網頁分類 目錄服務	聯合目錄 / OAI

Search Engine Basic Concept



What is "Google"?



- Definition: Googol <http://en.wikipedia.org/wiki/Googol>

Pronunciation: 'gü-"gol

Function: *noun*

Google is a play on the word **googol**, which was coined by Milton Sirotta, nephew of American mathematician Edward Kasner, and was popularized in the book, "Mathematics and the Imagination" by Kasner and James Newman. It refers to the number represented by the numeral 1 followed by 100 zeros.

- Originally called "Backrub", the logic behind the Google search engine was developed by graduate students **Larry Page** and **Sergey Brin** at Stanford University in 1995. Their first place of business was literally a garage. The garage location was chosen because it had a washer/dryer and a hot tub out back, they were already serving 10,000 searches a day.

- <http://www-db.stanford.edu/~backrub/google.html>

- Established in 1998
- Issued an IPO in April, 2004

<http://www.google.com/corporate/history.html>

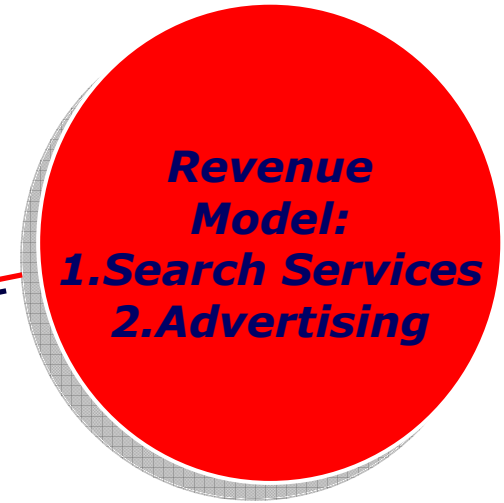
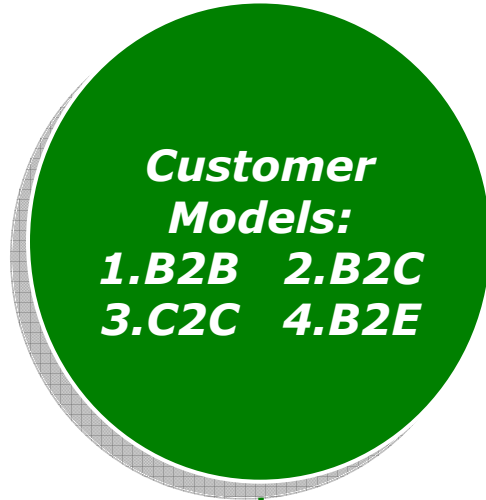


Google's Business Model



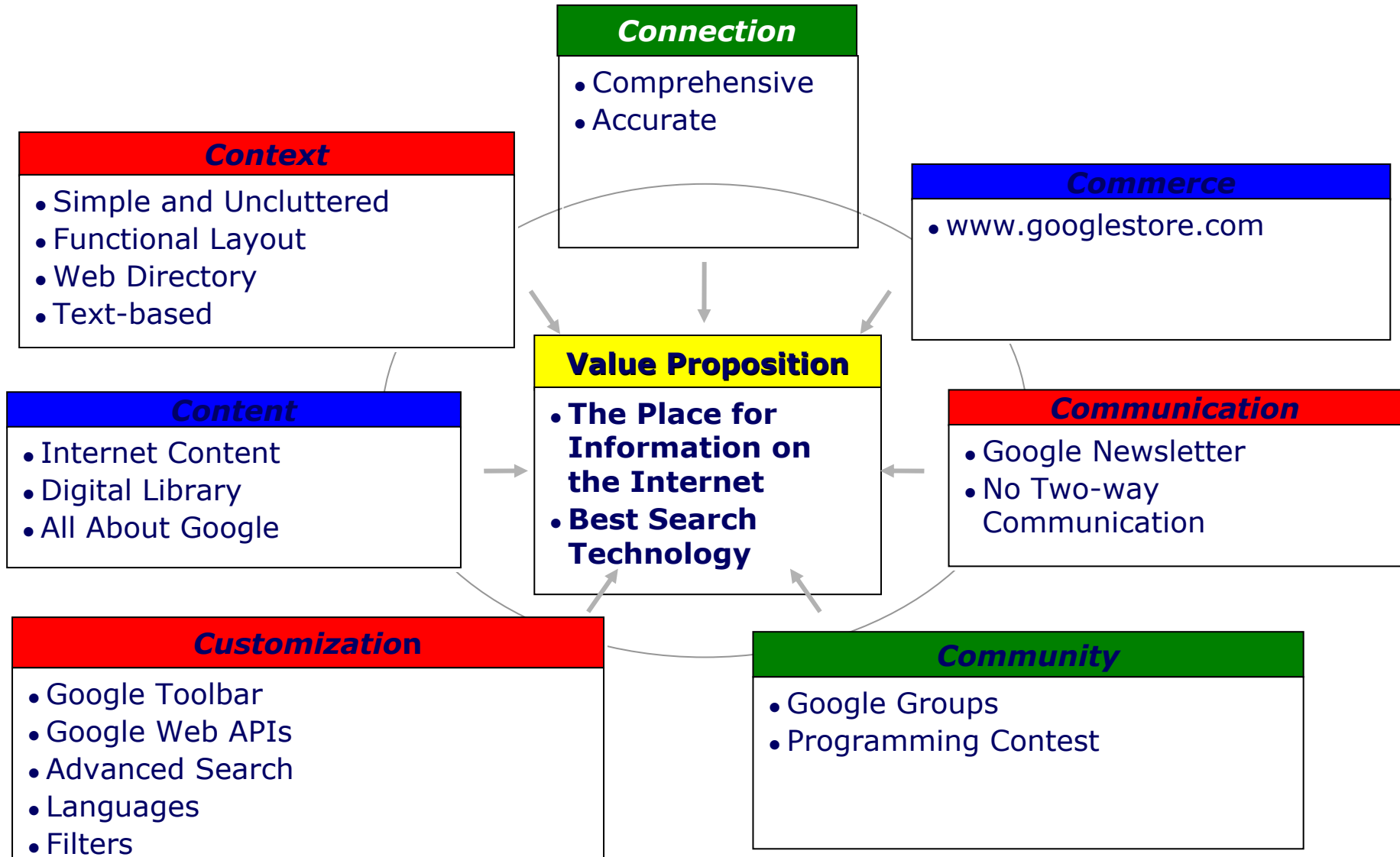
● Search Savior

● Ad-monger



Google™

Google's 7Cs Reinforcement

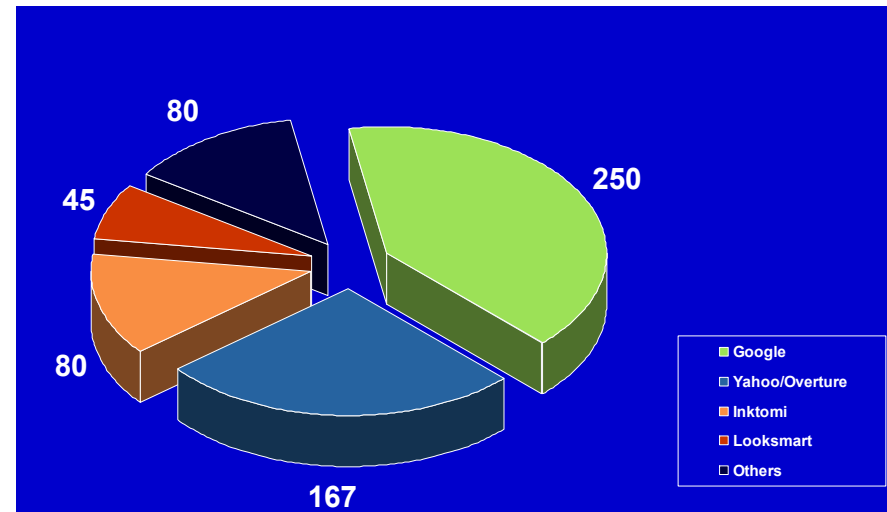
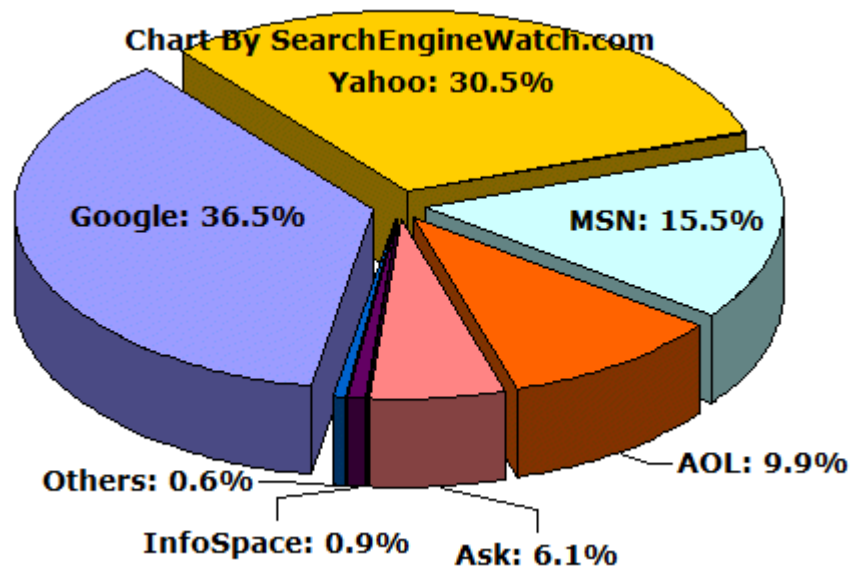


Market Shares / Traffic



- Share Of Searches: July 2005

- <http://searchenginewatch.com/reports/article.php/2156431>

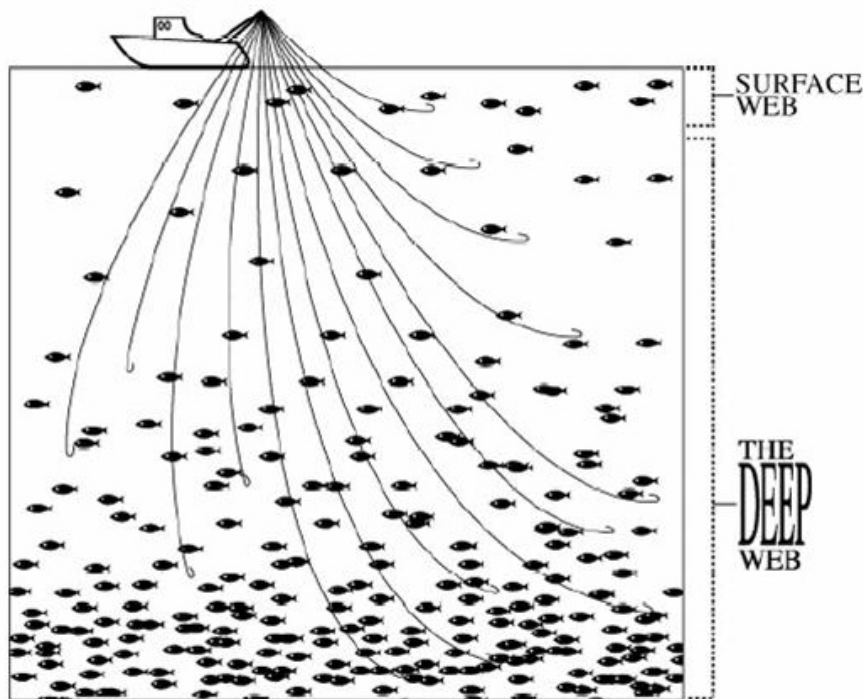


Searches Per Day in Millions

How big Internet is ?



- Total registered domains: over 63 million
- Total number of live domains: over 25 million
- Total number of eCommerce domains: over 1.2 million
- Total live domains containing more than 50 pages of content: over 2.5 million
- Web sites: over 8 billion pages



大海撈針

- 深海網比淺海網大500 倍
- 95% 深海網是公開的且可免費獲取的
- 深海網內容質量高1000+ 倍
- 估計有7,500 TB 訊息量
- 目前有45,000個淺海網搜尋引擎

挑戰極限



● 巨量、天量、無可限量

- No single search engine indexes more than 16% of web sites
- All search engines combined covering only 42%

● 極端異質

- Variable information value
- Variable length
- Often containing grammatical mistakes and typos
- Content may be outdated, false, or unreliable
- Multiple data formats
- Multiple languages and alphabets
- Structure VS non-structure data (XML, SGML, EDI....)

● 效率問題

- 15,000 ~ 20,000 search queries requested per minute
 - Passive update
-
-

Average size and type of web doc.

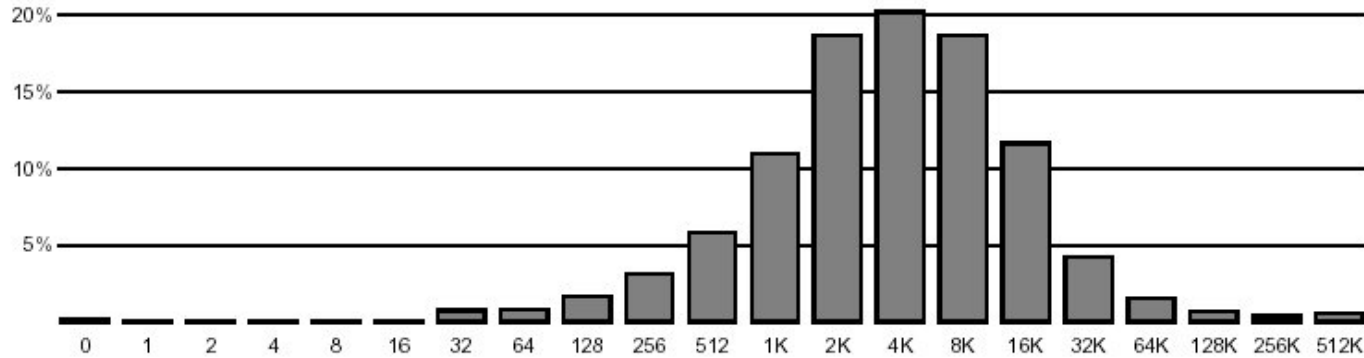
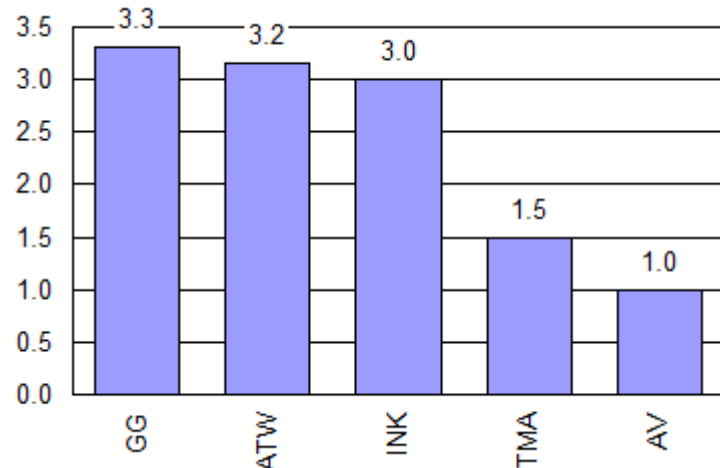


Figure 2: Histogram of the sizes of successfully downloaded documents.

MIME type	Number	Percent
text/html	41,490,044	69.2%
image/gif	10,729,326	17.9%
image/jpeg	4,846,257	8.1%
text/plain	869,911	1.5%
application/pdf	540,656	0.9%
audio/x-pn-realaudio	269,384	0.4%
application/zip	213,089	0.4%
application/postscript	159,869	0.3%
other	829,410	1.4%
Total	59,947,946	100.0%

Table 3: Distribution of MIME types.

Search Engine Index Size



Billions Of Textual Documents Indexed

As of Sept 2, 2003

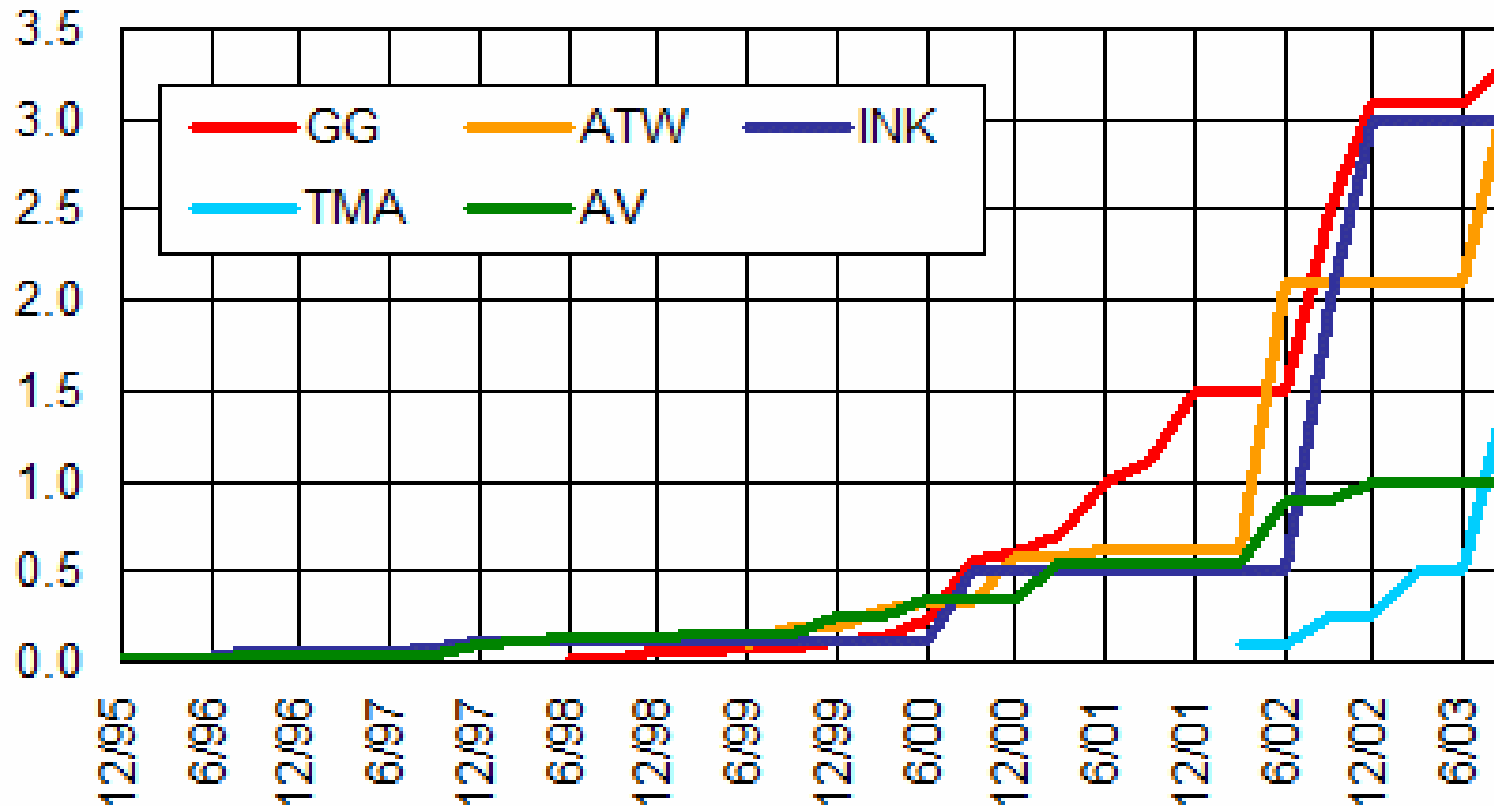
KEY: GG=Google, ATW=AllTheWeb, INK=Inktomi, TMA=Teoma, AV=AltaVista.

Search Engine	Showdown Estimate (millions)	Claim (millions)
Openfind		3,500
Google	968	2,000
WiseNut	579	1,500
AllTheWeb	580	507
Northern Light	417	358
AltaVista	397	500
Hotbot	332	500
MSN Search	292	500

Search Engine Supremacy



Billions Of Textual Documents Indexed
December 1995-September 2003



KEY: See [above](#).

<http://searchenginewatch.com/reports/article.php/2156481>

KEY: GG=Google, ATW=AllTheWeb, INK=Inktomi, TMA=Teoma, AV=AltaVista.

What google search?



[Web](#) [Images](#) [Groups](#) [News](#) [Local](#) [more »](#)

More, more, more ...










[Google Home](#)










[About Google](#)

[Help Center](#)








[Google Downloads](#)

Google Services

-  [Alerts](#)
Receive news and search results via email
-  [Answers](#)
Ask a question, set a price, get an answer
-  [Blog Search](#)
Find blogs on your favorite topics
-  [Book Search](#)
Search the full text of books
-  [Catalogs](#)
Search and browse mail-order catalogs
-  [Directory](#)
Browse the web by topic
-  [Groups](#)
Create mailing lists and discussion groups
-  [Images](#)
Search for images on the web
-  [Labs](#)
Try out new Google products

-  [Local](#)
Find local businesses and services
-  [Maps](#)
View maps and get directions
-  [Mobile](#)
Use Google on your mobile phone
-  [News](#)
Search thousands of news stories
-  [Scholar](#)
Search scholarly papers
-  [Special Searches](#)
Search within specific topics
-  [University Search](#)
Search a specific school's website
-  [Web Search](#)
Search over 8 billion web pages
-  [Web Search Features](#)
Do more with search

Google Tools

-  [Blogger](#)
Express yourself online
-  [Code](#)
Download APIs and open source code
-  [Desktop](#)
Search your own computer
-  [Google Earth](#)
Explore the world from your PC
-  [Picasa](#)
Find, edit and share your photos
-  [Toolbar](#)
Add a search box to your browser
-  [Translate](#)
View web pages in other languages

Google Technology
=
Search Technology

MapReduce



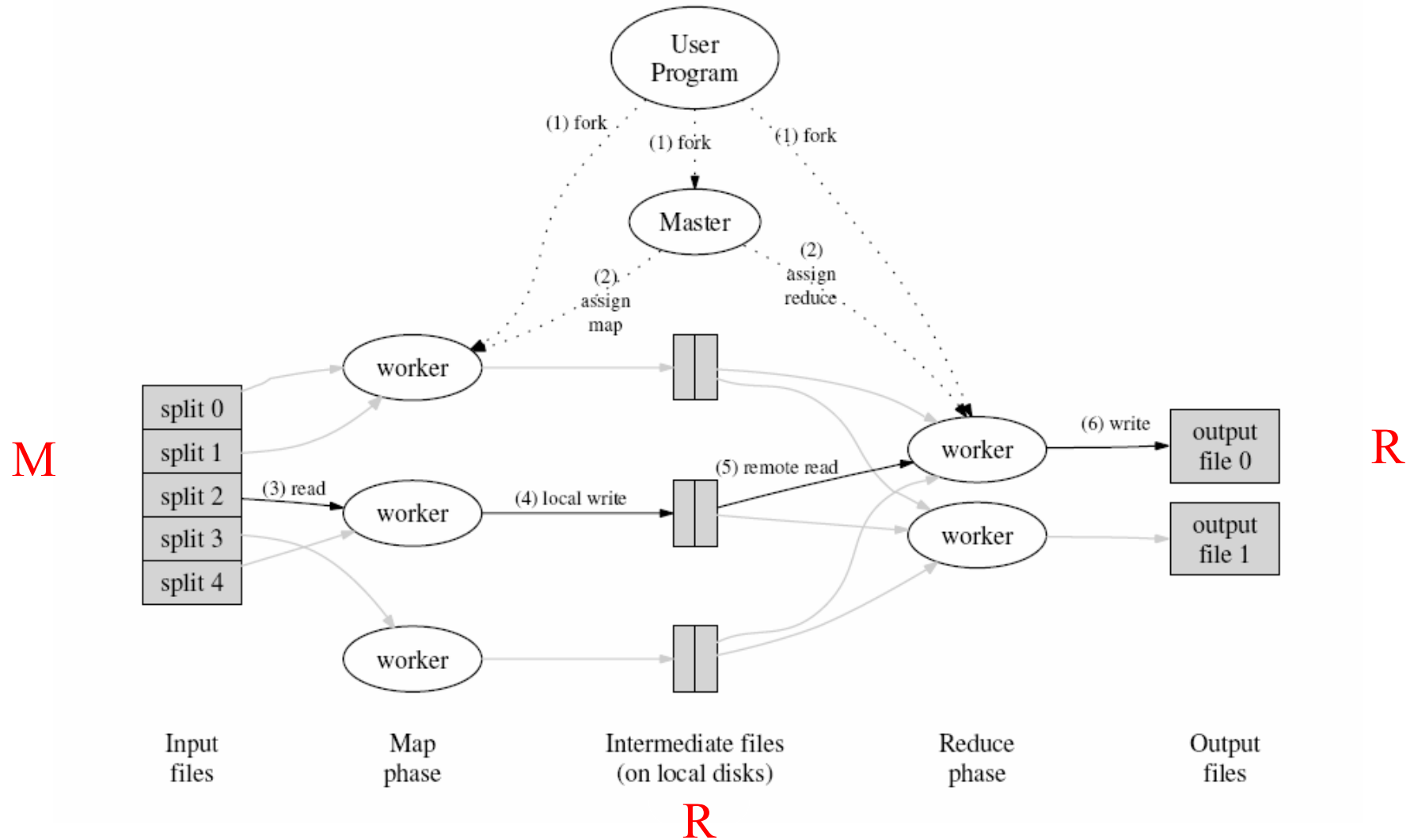
- Simplified Data Processing on Large Clusters
 - A programming model and an associated implementation (library) for processing and generating large data sets (on large clusters).
 - A new abstraction allowing us to express the simple computations we were trying to perform but hides the messy details of parallelization, fault-tolerance, data distribution and load balancing in a library.
 - Distributed grep:
 - Map: (*key*, whole doc/a line) → (the matched line, *key*)
 - Reduce: identity function
-
-

Programming Model



- A *map* function that process a **key/value** pair to generate a set of intermediate **key/value** pairs, and
 - Map: $(k1, v1) \rightarrow \text{list}(k2, v2)$
 - A reduce function that merges all intermediate values associated with the *same* intermediate key.
 - Reduce: $(k2, \text{list}(v2)) \rightarrow \text{list}(v2)$
 - Many real world tasks are expressible in this model.
 - Assumption: data has no correlation, or it is small.
-
-

Execution Overview



SPAM Problem



- Repeat keywords over and over just to get better rankings
- Target web pages to keywords that are unrelated to the exact web page
- Use hidden text or links on your web pages
- Submit web pages that are already indexed
- Participate in link-exchange programs to boost link-popularity
- Link to criminal, illegal, or very offensive websites

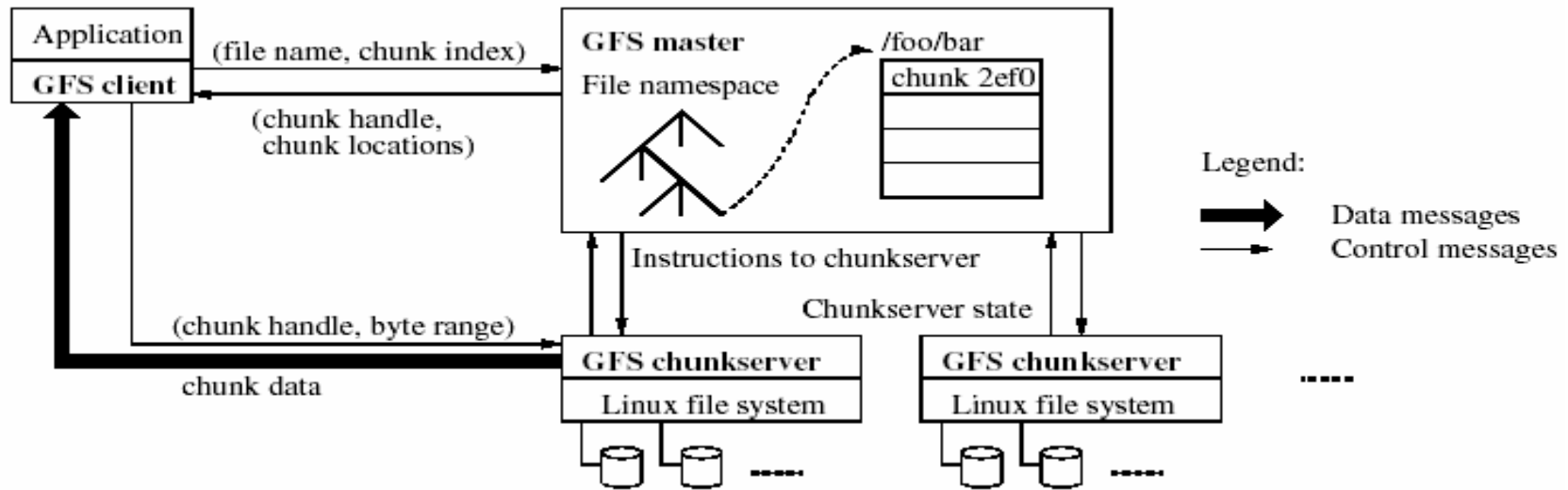


Google File System Architecture



- Files are divided into chunks
 - Fixed-size chunks (64MB)
 - Replicated over chunkservers, called replicas
 - Unique 64-bit chunk handles
 - Chunks as Linux files
 - Single master
 - Multiple chunkservers
 - Grouped into Racks
 - Connected through switches
 - Multiple clients
 - Master/chunkserver coordination
 - HeartBeat messages
-
-

GFS Architecture



- Contact single *master*
- Obtain chunk locations
- Contact one of chunkservers
- Obtain data

Link-based ranking techniques



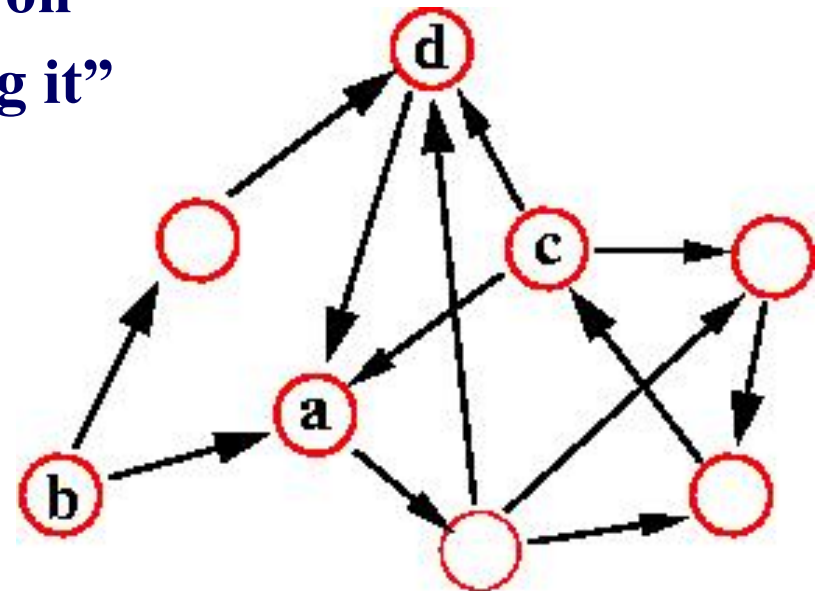
- **Ragerank (Brin&Page/Google)**

“significance of a page depends on significance of those referencing it”

- **HITS (Kleinberg/IBM)**

“Hubs and Authorities”

... more later



$$s(a) \sim s(b) + s(c) + s(d) ?$$

What is a Markov Chain?



- A Markov chain has two components:
 - 1) A network structure much like a web site, where each node is called a state.
 - 2) A transition probability of traversing a link given that the chain is in a state.
 - For each state the sum of outgoing probabilities is one.
- A sequence of steps through the chain is called a *random walk*.



Markov chains



- Markov chains have been extensively studied by statisticians and have been applied in a wide variety of areas.

Markov Property:

$$P(S_t | S_{t-1}, S_{t-2}, \dots, S_2, S_1) = P(S_t | S_{t-1}) \quad (3.3)$$

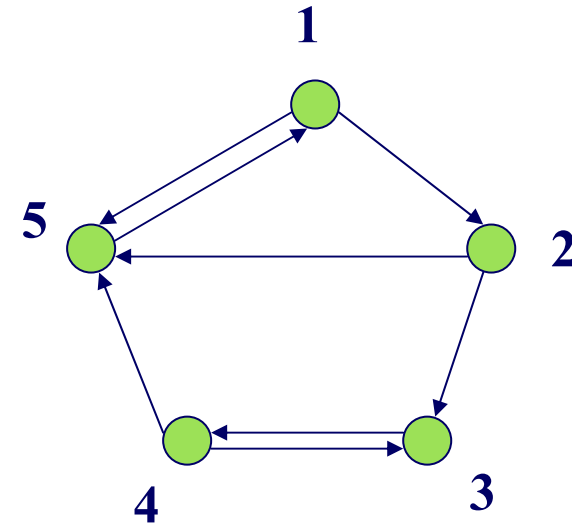
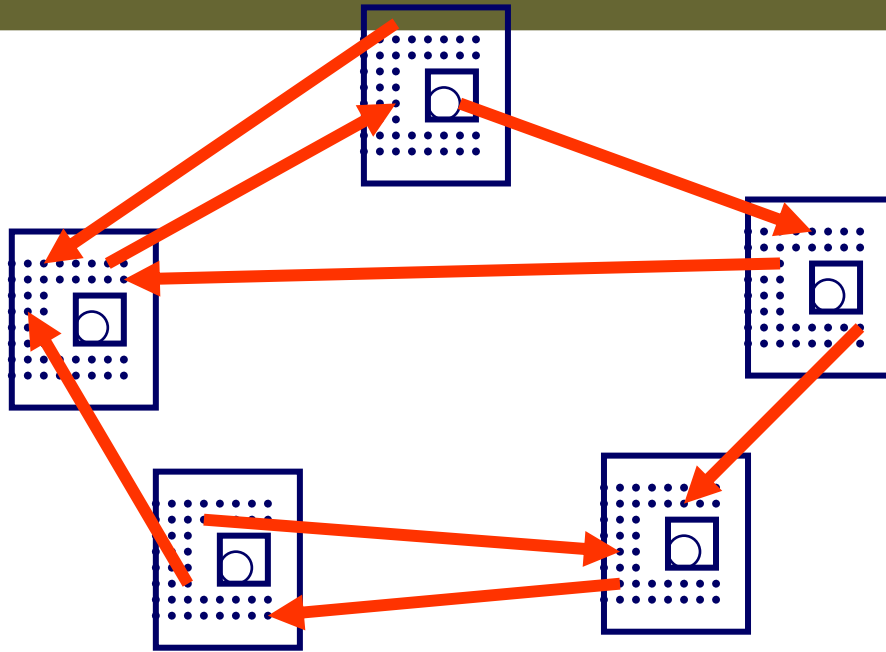
Page Rank (*PR*) - Definition



$$PR(W) = \frac{T}{N} + (1-T) \left(\frac{PR(W_1)}{O(W_1)} + \frac{PR(W_2)}{O(W_2)} + \dots + \frac{PR(W_n)}{O(W_n)} \right)$$

- W is a web page
- W_i are the web pages that have a link to W
- $O(W_i)$ is the number of outlinks from W_i
- T is the teleportation probability
- N is the size of the web

Mathematical Techniques



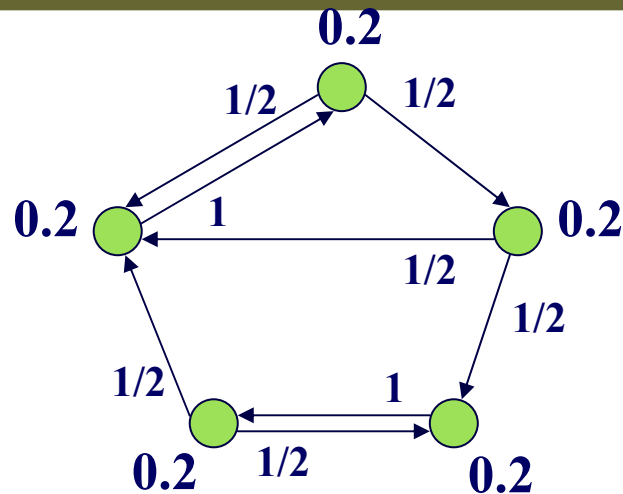
- **Ragerank (Brin&Page/Google)**

significance of a page depends on significance of pages referencing it

- **HITS (Kleinberg/IBM)**

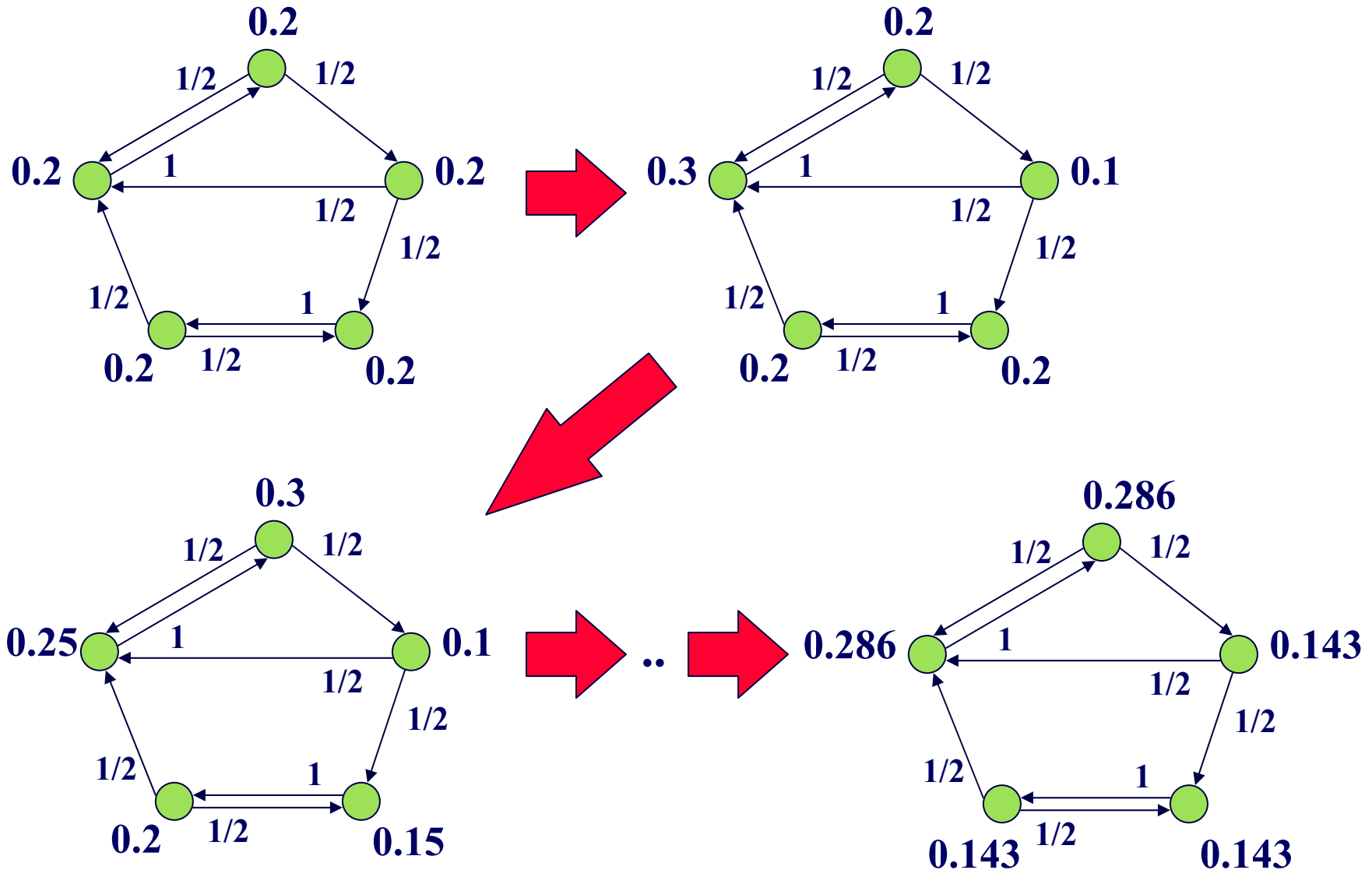
hubs and authorities: good hubs point to good authorities & vice versa

Pagerank



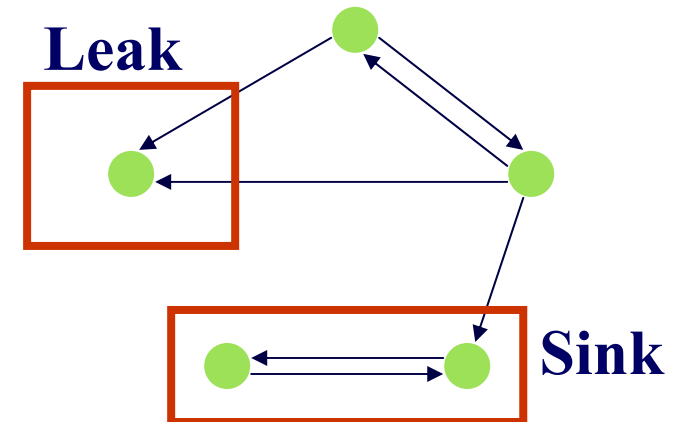
- initialize the *rank value* of each node to $1/n$ (0.2 for 5 nodes)
 - a node with k outgoing links transmits a $1/k$ fraction of its current *rank value* over that edge to its neighbor
 - iterate this process many times until it converges
 - **NOTE:** this is a random walk on the link graph
 - **Pagerank:** stationary distribution of this random walk
-
-

Mathematical Techniques



Convergence

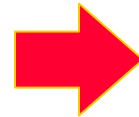
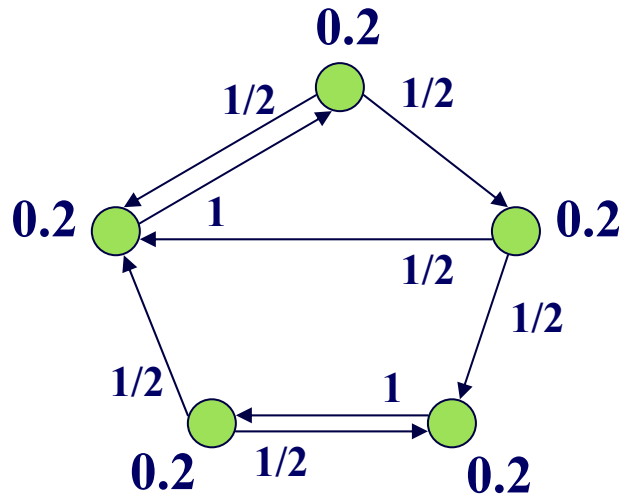
- **dealing with leaks:**
 - pruning, or
 - adding back links
- **dealing with sinks**
 - add a random jump to escape sink
 - with prob. $b = 0.15$ jump to random node
- **assume: if in a leak, always take random jump**
- **in this case, always converges to unique solution!**



Mathematical Techniques



Matrix notation



A

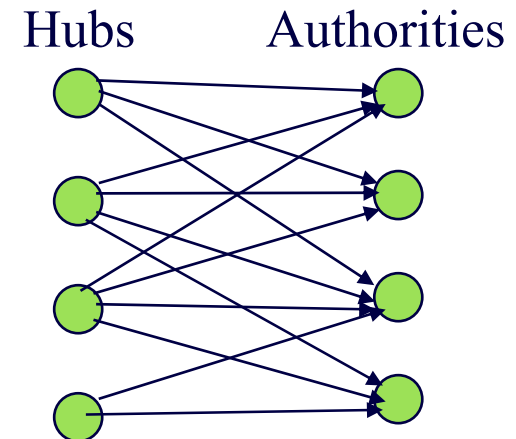
0	1/2	0	0	1/2
0	0	1/2	0	1/2
0	0	0	1	0
0	0	1/2	0	1/2
1	0	0	0	0

- stationary distribution: vector \mathbf{x} with $\mathbf{x}A = \mathbf{x}$
- A is primitive, and \mathbf{x} Eigenvector of A
- computed using Jacobi or Gauss Seidel iteration

HITS : Hubs and Authorities



- Algorithm developed by Kleinberg in 1998.
- query-dependent: first get 100 docs using term-based techniques
- Attempts to computationally determine hubs and authorities on a particular topic through analysis of a relevant subgraph of the web.
- Based on mutually recursive facts:
 - Hubs point to lots of authorities.
 - Authorities are pointed to by lots of hubs.
- Together they tend to form
- a bipartite graph:



HITS Algorithm

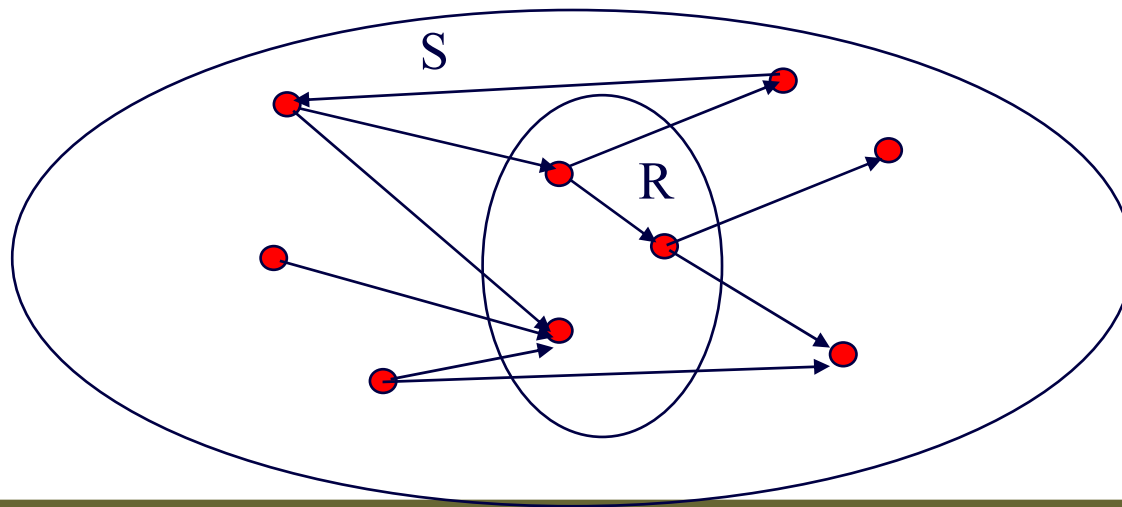


- Computes hubs and authorities for a particular topic specified by a normal query.
- First determines a set of relevant pages for the query called the *base set S*.
- Analyze the link structure of the web subgraph defined by *S* to find authority and hub pages in this set.

Constructing a Base Subgraph



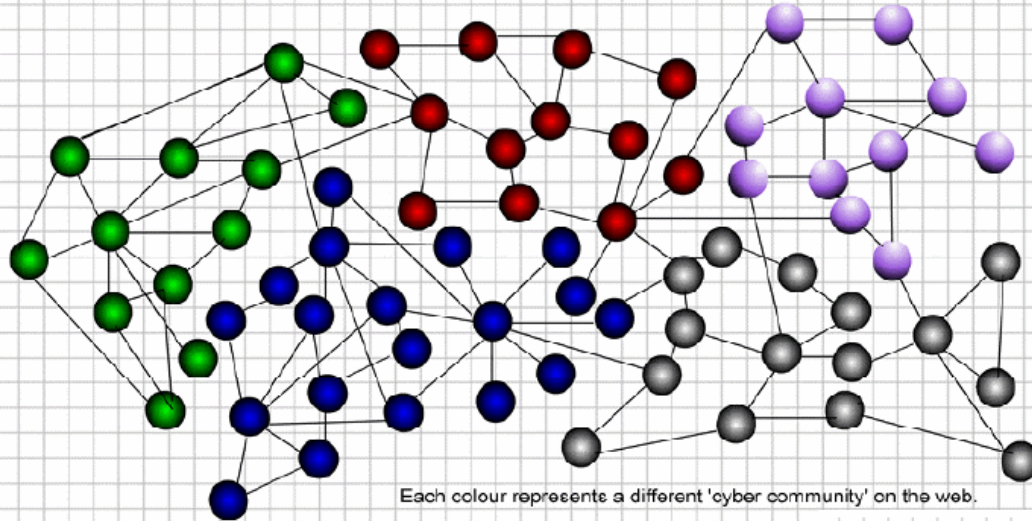
- For a specific query Q , let the set of documents returned by a standard search engine (e.g. VSR) be called the *root set* R .
- Initialize S to R .
- Add to S all pages pointed to by any page in R .
- Add to S all pages that point to any page in R .



Results



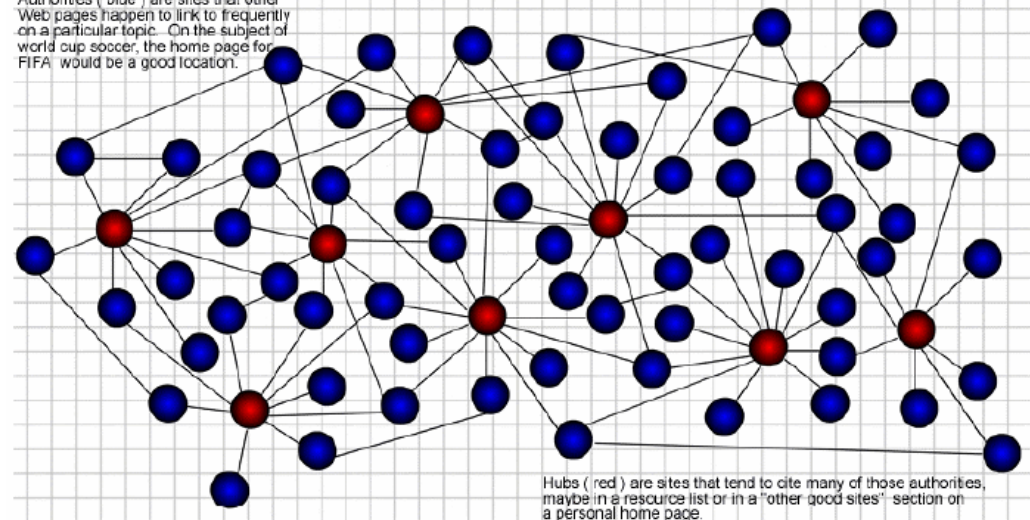
Link analysis can discover 'cyber communities', even pages with a very narrow interest, such as 'Japanese Kindergarten education in the US' (Broder)



Each colour represents a different 'cyber community' on the web.

Kleinberg's 'hubs & authorities'.

Authorities (blue) are sites that other Web pages happen to link to frequently on a particular topic. On the subject of world cup soccer, the home page for FIFA would be a good location.



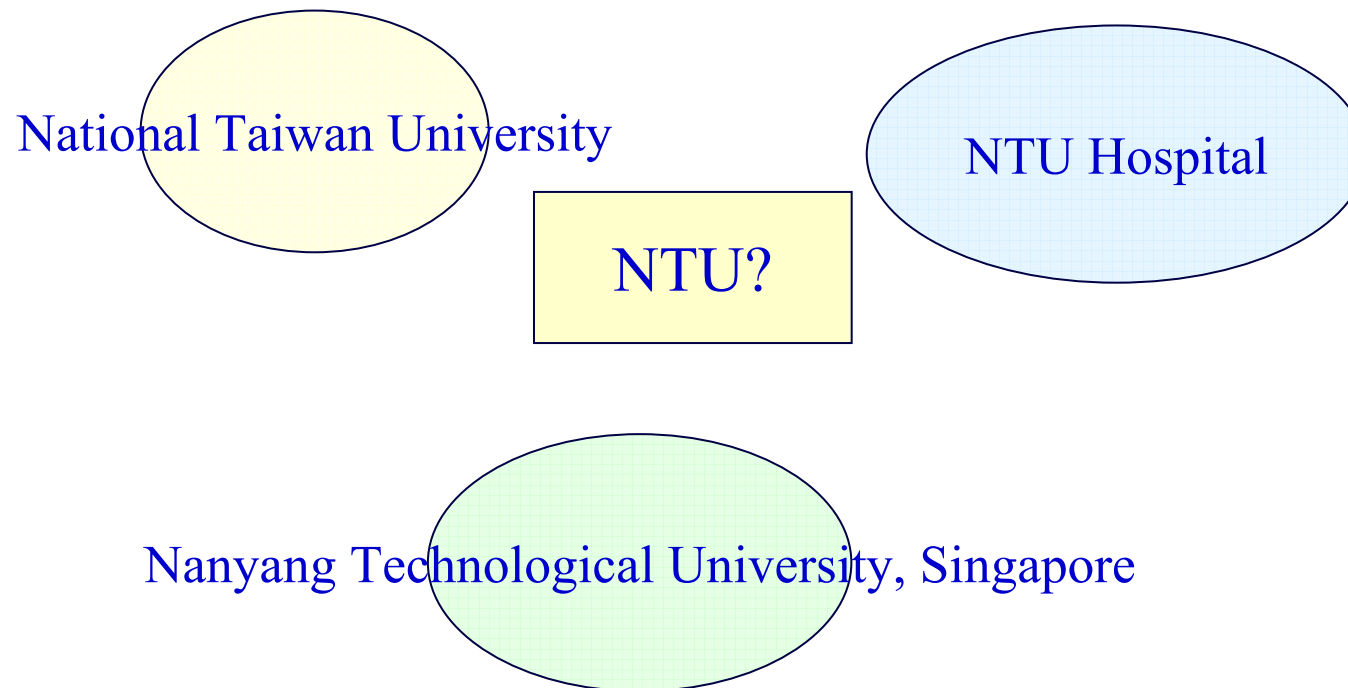
Hubs (red) are sites that tend to cite many of those authorities, maybe in a resource list or in a "other good sites" section on a personal home page.

Different UI experience



- Clusty.com (by Vivisimo)
 - Clustering engine
 - Snap.com (by Idealab)
 - Sorting by popularity, satisfaction, Web popularity, Web satisfaction, domain, ...
 - Alexa.com (by Amazon)
 - Average user review ratings, ...
 - TouchGraph Google Browser: visualization
 - <http://www.touchgraph.com/TGGoogleBrowser.html>
 - Kartoo.com
 - A visual meta search engine
-
-

Example of Search Result Clustering



Clustering Search Engines



- Vivisimo.com
 - Clusty.com
- KillerInfo.com
- InfoNetWare.com
- **iBoogie.com**
- SnakeT (Snippet Aggregation for Knowledge ExTraction):
<http://roquefort.unipi.it/>
 - A hierarchical clustering engine for snippets
- ...

Example on Vivisimo



company | products | solutions | customers | demos | partners | press

NTU the Web Search Advanced Help!

NEW search auctions at Clusty.com

Clustered Results Top 210 results retrieved for the query NTU (Details)

- NTU (210)
 - Technological University (33)
 - Nottingham Trent (17)
 - Search (12)
 - Taiwan (12)
 - Networks (11)
 - Union (8)
 - Northern Territory University (8)
 - Club (9)
 - Management (8)
 - Digital, Library (6)
 - More

Find in clusters:
Enter Keywords

- Nanyang Technological University** [new window] [frame] [preview]
Provides resources for prospective students interested in undergraduate and graduate studies, lists schools, offices, and events.
URL: www.ntu.edu.sg - show in clusters
Sources: Looksmart 1, Wisenut 1, Lycos 3, MSN 3, Open Directory 5
- National Technological University** [new window] [frame] [preview]
Academic All Academic Courses Search Course Catalog
URL: www.ntu.edu - show in clusters
Sources: Lycos 1, MSN 1, Wisenut 3, Open Directory 25
- Nottingham Trent University** [new window] [frame] [preview]
Provides course overviews by faculties and departments, presents main research areas and features a prospectus.
URL: www.ntu.ac.uk - show in clusters
Sources: Open Directory 2, Lycos 4, MSN 4, Looksmart 9

Clusty.com



A screenshot of a Microsoft Internet Explorer browser window displaying the Clusty.com search results for the query "NTU". The browser's address bar shows "http://clusty.com/search?query=NTU". The Clusty.com interface includes a search bar with "NTU" entered, a "Cluster" button, and navigation tabs for "Web+", "News", "Images", "Shopping", "Encyclopedia", "Gossip", and "Customize!". The search results are categorized by "Cluster by: Topics" and show "Top 218 results of at least 83,258 retrieved for the query NTU". On the left, a sidebar lists clusters such as "Technological, University (45)", "School (19)", "Nottingham Trent University (17)", "Northern Territory University (12)", "Search (12)", "Union (9)", "National Taiwan University (11)", "Languages (8)", "Club (9)", and "NTU Forum (4)". The main content area displays "Sponsored Results" for "National Technological University Online" and "NTU Online Engineering Programs", followed by "Search Results" listing "Nanyang Technological University", "National Technological University", and "Nottingham Trent University". The Windows taskbar at the bottom shows the system tray with the time "上午 11:24" and a 94% battery level.

InfoNetWare.com



The screenshot shows a Microsoft Internet Explorer browser window with the title "RealTerm search: NTU - Microsoft Internet Explorer". The address bar contains a search query: "http://www.infonetware.com/?query=NTU&ndocs=700&timeout=8&base=infonetware&xbase=inw&html_base=inw&page=infonetware.html&izgui=nav_frame&pref=std&power+search.x=23&power+search.y=6". The search engine used is Google. The search results are for the term "ntu" and show a list of 95 documents, with the first 10 displayed. The left sidebar shows a "TOPICS" list with various categories and their respective document counts. The main content area displays the following search results:

- 0 National Technological University**
 - Academic. All Academic Courses. Search Course Catalog ... Introducing the NEW NTU MBA Program! NTU Certificate Programs Now Available! Fast Track in Computer Science from NTU. Keep on the leading-edge of the IT and Engineering ...
www.ntu.edu
Add to shortlist bbc:1 msn:0 altavista:2
- 1 Nottingham Trent University**
 - ... Postgraduate and Professional. About NTU. News & Events. Contacts ... Latest news from Nottingham Trent University. NTU Open Days ...
www.ntu.ac.uk
Add to shortlist bbc:0 msn:3 altavista:0
- 2 National Taxpayers Union & National Taxpayers Union Foundation**
 - ACTION CENTER. Sign up with NTU's Taxpayer Action Center. Search our Articles! What's New. In each Congressional election cycle, NTU surveys candidates about their positions on important tax and budget issues. ... NTU Rates Congress. Each year NTU rates each Member of Congress on ...
www.ntu.org
Add to shortlist bbc:4 msn:1 altavista:4
- 3 Nanyang Technological University**
 - This site requires Java Script to be enabled. Please enable Java Script on your browser, or use a Java Script capable browser to proceed.
www.ntu.edu.sg
Add to shortlist bbc:2 msn:2 altavista:5
- 4 Theory, Culture and Society Centre**
 - ... Fax: +44 (0)115 848 6331. E-mail: tcs@ntu.ac.uk ...
tcs.ntu.ac.uk

The browser's status bar at the bottom shows "網際網路" (Internet).

iBoogie



Home Contact Us Technology Make iBoogie Homepage

Web MP3/Audio Images Video Directory Advanced

Any language education in australia Let's Boogie

Expand Web Tips™ Collapse 165 results out of 165 4,068,001 web pages found

- All results
 - Study in australia
 - Tertiary education
 - World
 - Aspects
 - Health schools australia - national
 - Study abroad
 - Study options - access
 - Australia and australian
 - School education
 - Education and training
 - University in australia
 - University of technology
 - Queensland university - 2005 high
 - Charles sturt university
 - Home education in australia
 - Students
 - Academic - challenge
 - International students
 - Online
 - Department of education
 - Department of education and traini
 - Australian government
 - Children services - south australia
 - Queensland department
 - Higher education
 - Learning
 - Search
 - Education services
 - Education south australia
 - International education

- Welcome to EdNA Online**
HOME. Login to MyEdNA. Username: Password: Remember me on this computer. Jump to my MyEdNA page. New Today! School Education . New Today! VET. New Today! Higher Education . Education News
<http://www.edna.edu.au/> - Similar pages
msn WiseNut
- Department of Education and Training**
... is widely acknowledged that formal education is the key to a successful fut...
Government Schools in Western Australia ...
<http://www.edna.edu.au/> - Similar pages
msn WiseNut
- <http://www.edna.edu.au/edna/page1.html>**
<http://www.edna.edu.au/edna/page1.html> - Similar pages
WiseNut
- Australia and Australian Business Directory**
Australia link resources of discount shopping, books and magazines, programs and home business opportunities and more.
<http://www.pinoyshowcase.com/search/Australia/> - Similar pages
finwhat
- IDP Portal**
Global Apply Online. IDP Media Centre. Search over 16,000 courses. Ac...
in Australia ... Legal Disclaimer | Feedback | Copyright IDP Education
<http://www.idp.com/> - Similar pages
msn WiseNut
- <http://www.edna.edu.au/EdNA>**
www.edna.edu.au/EdNA
<http://www.edna.edu.au/EdNA> - Similar pages
msn
- Government Education Portal - Australia**
Search. A-Z Topic List. Browse all Topics. Site Map
<http://www.education.gov.au/> - Similar pages
msn
- Engineering Education Australia**
Engineering Education Australia (EEA) is a wholly owned business unit of the Institution of Engineers, Australia , brokering professional development...
<http://www.eeaust.com.au/> - Similar pages
WiseNut
- Open Directory - Regional: Oceania: Australia - Education**
the open directory only in Australia (Education). See also: This category in other languages: Educator. Information on all aspects of education in

Related searches can be expanded in the side frame in iBoogie, so you can quickly see if they are relevant

Way Back Machine



INTERNET ARCHIVE
WayBackMachine

Enter Web Address: All [Adv. Search](#) [Compare Archive Pages](#)

Searched for <http://www.theatrelinks.com> 32 Results

Note some duplicates are not shown. [See all](#).
* denotes when site was updated.

Search Results for Jan 01, 1996 - Nov 02, 2004								
1996	1997	1998	1999	2000	2001	2002	2003	2004
0 pages	0 pages	0 pages	0 pages	0 pages	1 pages	8 pages	20 pages	2 pages
					Nov 27, 2001 *	Jan 22, 2002 * Jan 24, 2002 * May 28, 2002 * May 31, 2002 * Jul 18, 2002 * Sep 28, 2002 * Nov 26, 2002 * Nov 27, 2002	Feb 10, 2003 * Mar 20, 2003 * Mar 28, 2003 * Mar 31, 2003 * Apr 09, 2003 * Apr 11, 2003 * May 25, 2003 * May 31, 2003 * Jun 12, 2003 * Jun 23, 2003 * Aug 06, 2003 * Aug 07, 2003 * Sep 27, 2003 * Sep 28, 2003 * Oct 17, 2003 * Nov 26, 2003 * Dec 04, 2003 * Dec 12, 2003 * Dec 15, 2003 * Dec 29, 2003 *	Jan 29, 2004 * Feb 05, 2004 *

[Home](#) | [Help](#)

Copyright © 2001, [Internet Archive](#) | [Terms of Use](#) | [Privacy Policy](#)

搜尋引擎種類 P2P Search



Kuro



ezPeer

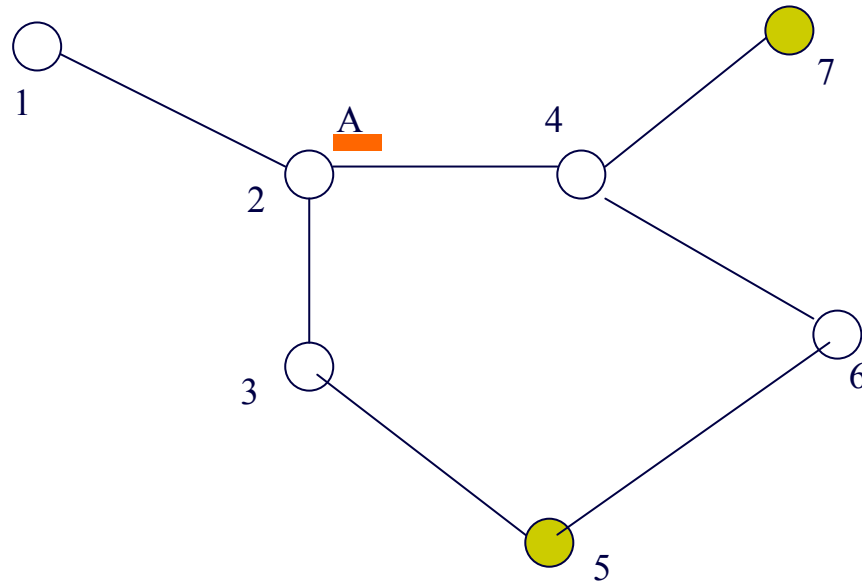
類別	歌曲歌詞圖片	大小	擁有者	速度	資訊
曲	國歌(a-mei).mp3	1.36MB	11122...	未知	1'27"/1
曲	張惠妹張惠妹-國歌.MP3	1.36MB	melohen	64K	1'27"/1
曲	女(國歌(張惠妹珍藏版)).mp3	1.36MB	twoviva	T1	1'27"/1
曲	張惠妹_國歌.mp3	1.36MB	sadie...	未知	1'27"/1
曲	阿妹-國歌.mp3	1.36MB	modulus	CABLE	1'27"/1
曲	好歌推薦張惠妹-國歌.mp3	1.36MB	gule123	T3+	1'27"/1
曲	張惠妹-國歌.mp3	1.36MB	chwdrw...	DSL	1'27"/1
曲	張惠妹-國歌.mp3	1.36MB	jainu1...	DSL	1'27"/1
曲	國語歌手\女歌手張惠妹張惠妹-國歌.MP3	1.36MB	j4406	未知	1'27"/1
曲	Chinese Pop\FEMALE張惠妹\k-mei(國歌).mp3	1.36MB	gunma	T1	1'27"/1
曲	張惠妹_國歌.MP3	1.36MB	maha6...	未知	1'27"/1
曲	民歌+老歌(國歌_A-mei版).mp3	1.36MB	x_s_w...	DSL	1'27"/1
曲	Chinese\k-mei(國歌).mp3	1.36MB	jsh67821	CABLE	1'27"/1
曲	MP3中文(女歌手張惠妹\k-mei(國歌).mp3	1.36MB	xinja	DSL	1'27"/1
曲	MP3\k-mei(國歌).mp3	1.36MB	xinja	DSL	1'27"/1
曲	應景歌曲張惠妹-國歌.mp3	1.36MB	mouth	DSL	1'27"/1

共找到32首歌曲, 0個歌詞, 0張圖片

下載

因為音樂 我們分享

Gnutella search mechanism

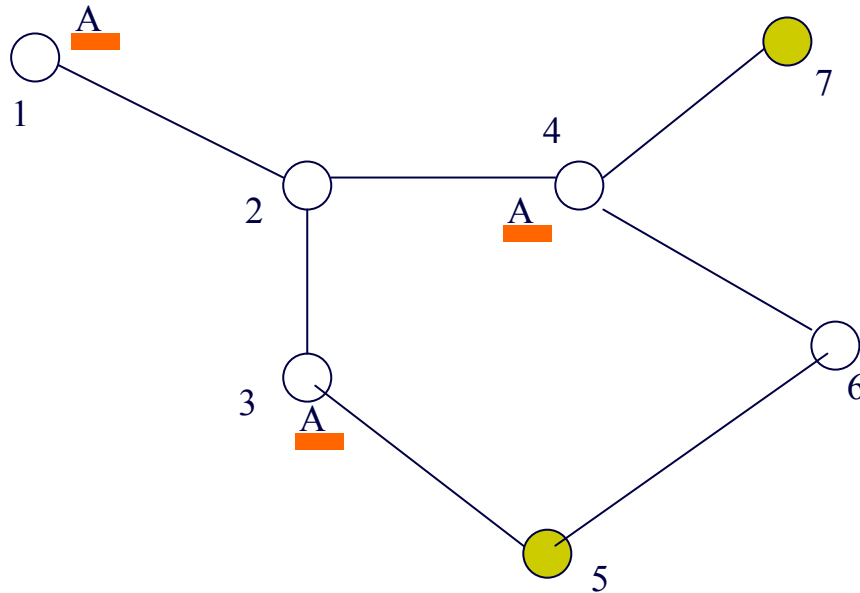


Steps:

- Node 2 initiates search for file A



Gnutella search mechanism

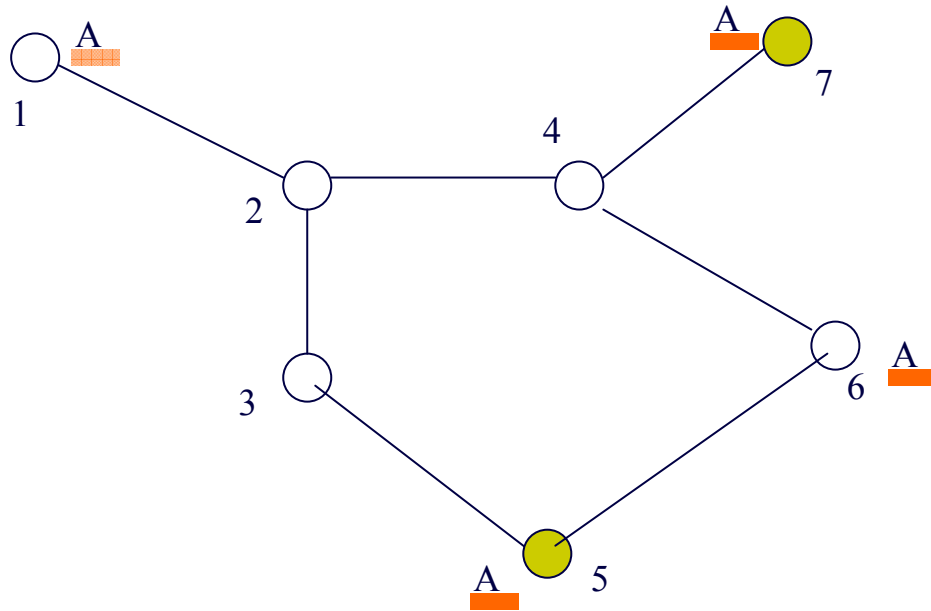


Steps:

- Node 2 initiates search for file A
- Sends message to all neighbors



Gnutella search mechanism

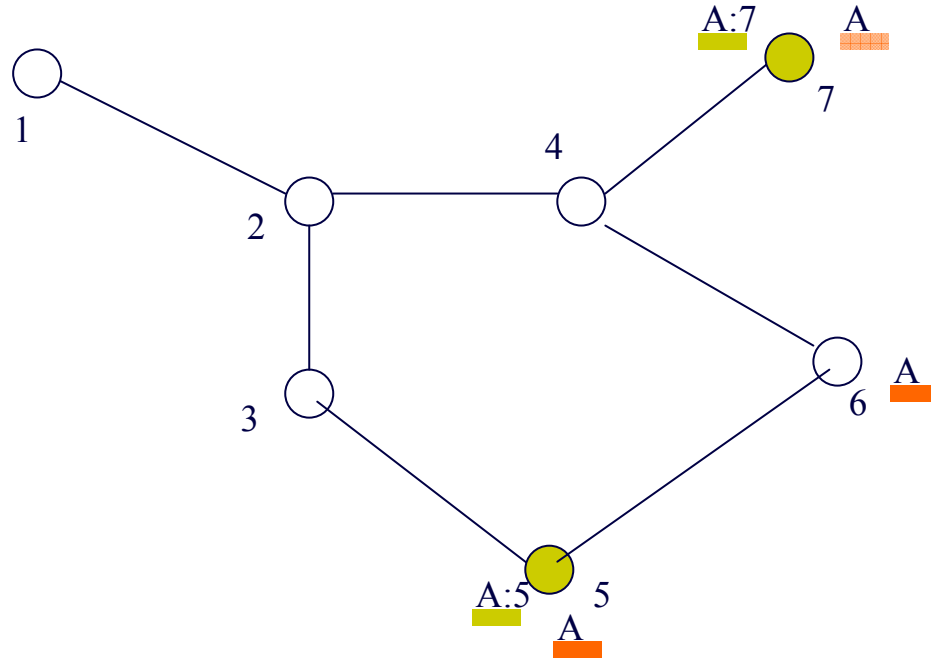


Steps:

- Node 2 initiates search for file A
- Sends message to all neighbors
- Neighbors forward message



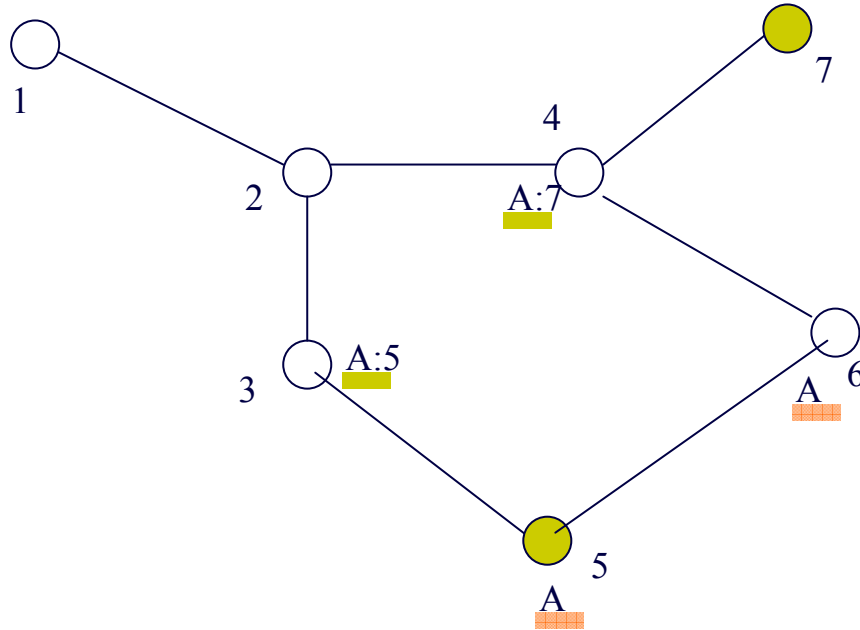
Gnutella search mechanism



Steps:

- Node 2 initiates search for file A
- Sends message to all neighbors
- Neighbors forward message
- Nodes that have file A initiate a reply message

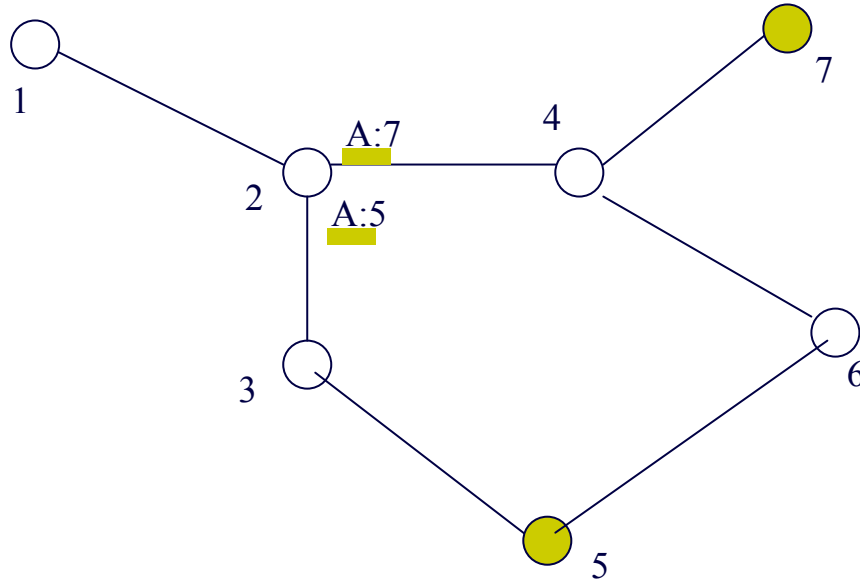
Gnutella search mechanism



Steps:

- Node 2 initiates search for file A
- Sends message to all neighbors
- Neighbors forward message
- Nodes that have file A initiate a reply message
- Query reply message is back-propagated

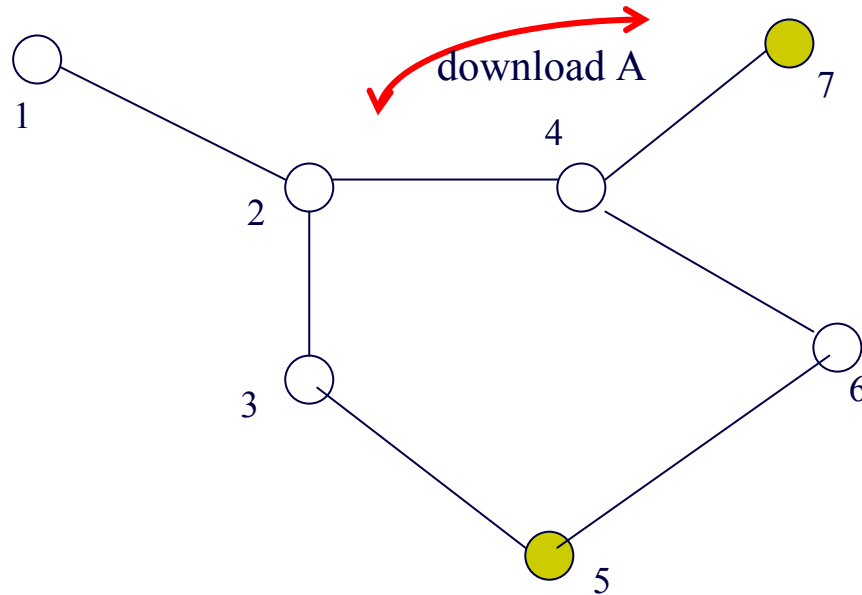
Gnutella search mechanism



Steps:

- Node 2 initiates search for file A
- Sends message to all neighbors
- Neighbors forward message
- Nodes that have file A initiate a reply message
- Query reply message is back-propagated

Gnutella search mechanism



Steps:

- Node 2 initiates search for file A
- Sends message to all neighbors
- Neighbors forward message
- Nodes that have file A initiate a reply message
- Query reply message is back-propagated
- File download



visionnext

誠信、創新、服務、價值

Image Search

不斷創新視野無限

Image Search Approaches



● Context-based approach

- text-based 、 language/knowledge-based
- Caption-based approach and hypermedia
- Google 、 openfind 、 eefind 、 ditto

● Content-based approach

- image-based 、 vision-knowledge-based
- QBIC (IBM) 、 PhotoBook (MIT) 、 VisualSEEk (Columbia)



一圖勝過千言萬語



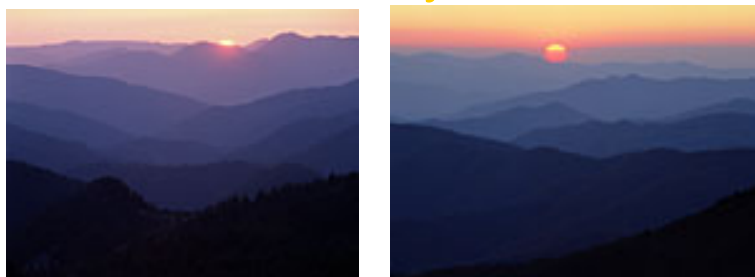
Semantic Similarity



Shift-Invariant Similarity



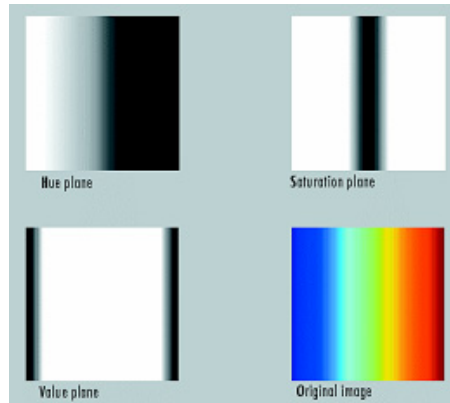
Color Similarity



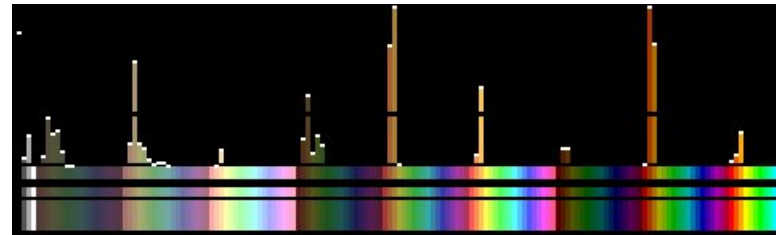
Indexing



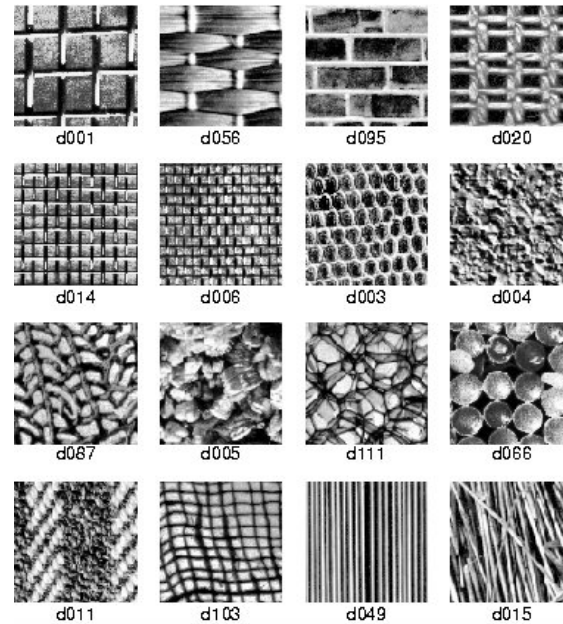
Color



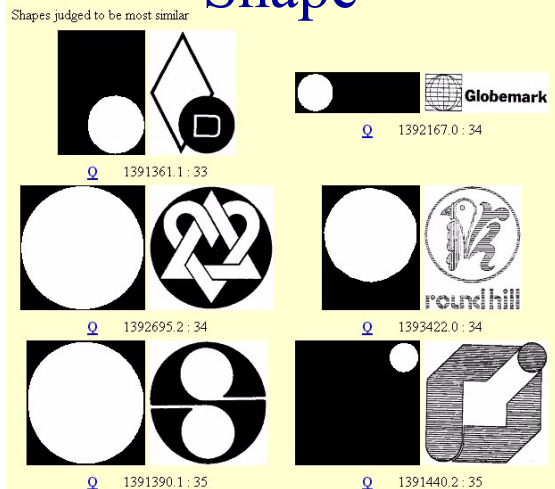
Color Histogram



Texture



Shape

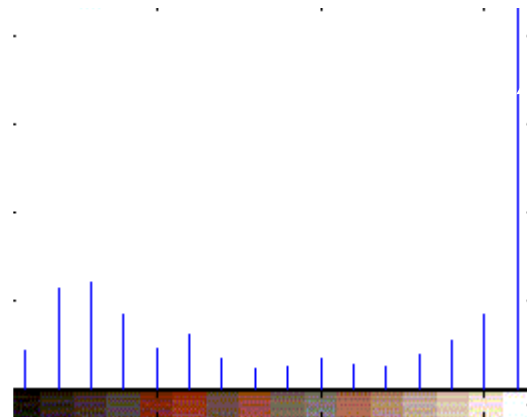


Colour Histogram



- Describe the colors and its percentages in an image.

$$f_c = \left\{ (I_j, P_j) \mid I_j \in \text{ColorValue}, 0 \leq P_j \leq 1, \sum_{1 \leq j \leq N} P_j = 1, \text{ and } 1 \leq j \leq N \right\}$$



Similarity by what?

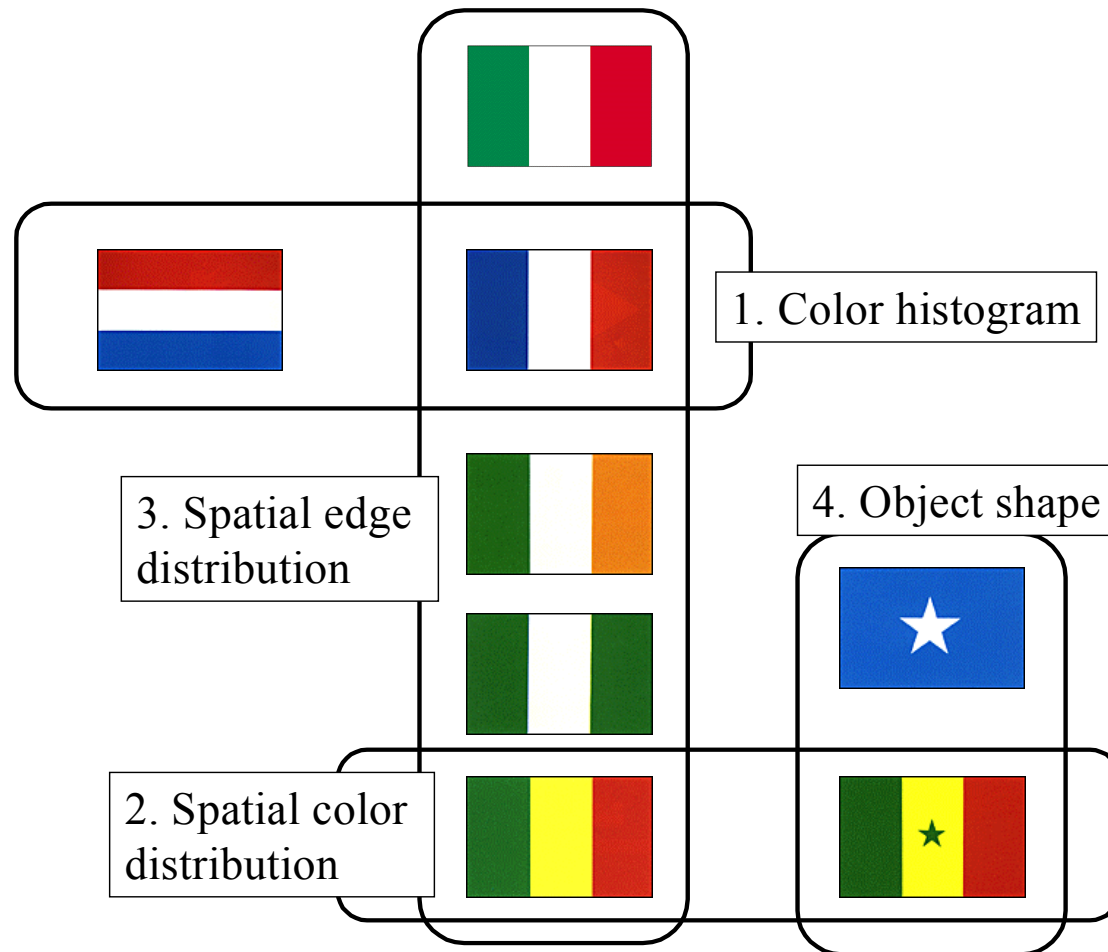


Image Directory for Browsing



PC home Online 網路家庭 圖片搜尋 | PC home | 圖片搜尋 | 網站搜尋 | 意見 |

PC home > 圖片搜尋

全部 只找大圖 (200x200以上)

"本週熱門100"

- 冠軍: [可愛的圖片](#)
- 亞軍: [藍色生死戀](#)
- 季軍: [流星花園](#)
- 殿軍: [數碼寶貝](#)
- 第5: [小叮噠](#)

[更多...](#)

"圖利他人"

- [推薦有好圖的網站](#)

"找圖交流區"

- [找圖交流區](#)
- [意見信箱](#)
- [使用說明](#)

影視娛樂

[連續劇](#) [台灣男藝人](#) [台灣女藝人](#)
[日劇](#) [日本男藝人](#) [日本女藝人](#) [更多...](#)

動漫電玩

[電玩人物](#) [漫畫人物](#) [卡通人物](#)
[中文漫畫](#) [日本漫畫](#) [更多...](#)

休閒消費

[手機](#) [家電](#) [鐘錶](#) [玩偶](#) [鞋襪](#) [收藏](#)
[星座](#) [化妝品](#) [酒](#) [服飾](#) [珠寶](#) [更多...](#)

資訊通訊

[面版](#) [桌面](#) [網頁素材](#) [電子賀卡](#)
[網路硬體](#) [電腦硬體](#) [更多...](#)

動物

[哺乳類](#) [鳥類](#) [爬蟲類](#) [兩棲類](#) [魚類](#)
[節肢昆蟲類](#) [腔腸軟體棘皮類](#) [更多...](#)

植物

[被子植物](#) [裸子植物](#) [蕨類菌類](#)
[水果](#) [蔬菜](#) [花草](#) [樹](#) [更多...](#)

自然天文

[天氣](#) [地形](#) [風景](#) [天文](#) [生物](#)
[海洋景觀](#) [超自然現象](#) [更多...](#)

景點古蹟

[台灣景點](#) [旅遊場所](#) [台灣古蹟](#)
[美加景點](#) [歐洲景點](#) [日韓景點](#) [更多...](#)

運動

[中華職棒](#) [美國職籃](#) [美國職棒](#)
[日本職棒](#) [撞球](#) [足球](#) [網球](#) [更多...](#)

人物


[作家](#) [漫畫家](#) [畫家](#) [音樂家](#) [配音員](#)
[運動選手](#) [政治人物](#) [科技人](#) [更多...](#)

Search by Text: Metadata (1)



CalPhotos (Berkeley digital library project)

CalPhotos
California Department of Water Resources



This form can be used to look for photographs from more than 17,000 images of California's natural resources from the California Department of Water Resources. [Click here](#) for more information about the DWR photos.

Instructions: To look for photos, choose **one or more** of the options below and click "Search".

Photo Caption	contains	<input type="text"/>
Subject	begins with	Computers (5)
Category	equals	any
Location	begins with	California (1)
Date of Photo	equals	<input type="text"/> Example: 1995-01-31
Perspective	equals	any
Photo ID		<input type="text"/>

Search by Text: Metadata (2)



visionnext

熊貓: Koala

WTC

The screenshot shows the Ditto search engine interface. At the top, a blue banner contains the number '5' and the text 'The attention span of the average user is 7 seconds.' Below this is the search bar with 'koala' entered. The search results are displayed as a grid of 12 image thumbnails, each with a small caption and file size. The thumbnails show various images of koalas. At the bottom of the page, there are navigation links and a copyright notice for 2001.

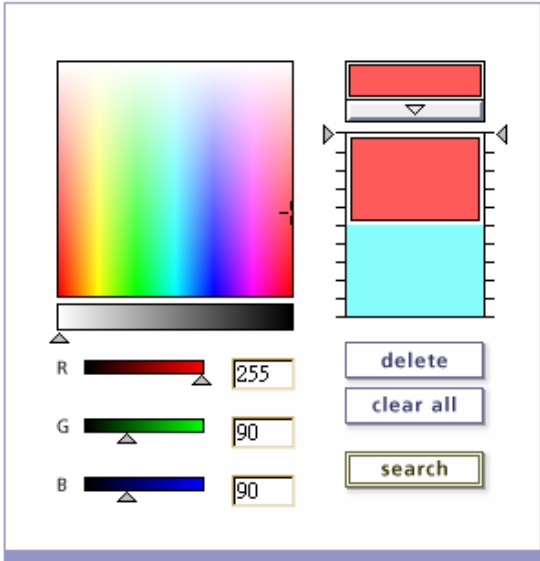
The screenshot shows the Google search engine interface. The search bar contains 'WTC'. The search results are displayed as a grid of 12 image thumbnails, each with a small caption and file size. The thumbnails show various images related to the World Trade Center (WTC), including maps and architectural drawings. A red circle highlights the Google logo and search bar area. At the bottom of the page, there are navigation links and a copyright notice for 2001.

About Us | Add Ditto Search to Your Site | Contact Us | Copyright | Disclaimer | Family Filters
Help | Home | Partners | Privacy Policy | Products | Remove a URL | Request a Review | Submit a Site

Ditto.com provides a visual mechanism to search the web using pictures instead of text. Users are directed to the originating web site on which the pictures are located. Should you wish to use any picture, photo or audio you see during the search process, you must obtain the appropriate permission from the owner of the material.







©2001 Google

IBM QBIC – Query by Color



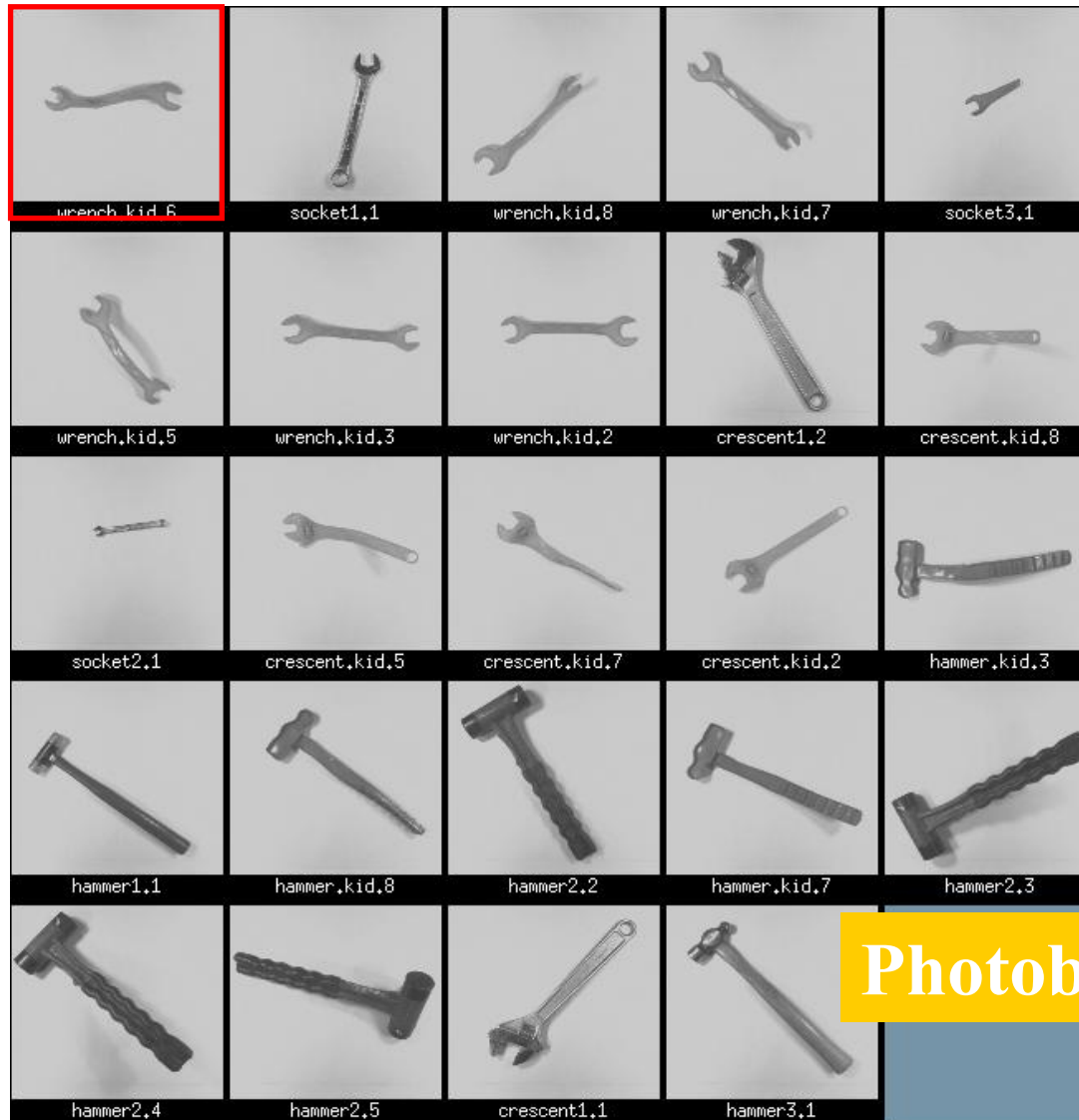
1. Use your mouse to select a colour from the palette.
2. Click the arrow button to add the colour to the bucket.
3. Slide the triangular handles on the bucket to adjust the percentage of this colour.
4. You may repeat this process until the bucket is full. When you are ready, click Search.

SEARCH RESULTS

	1) Bouquet (Vase with Two Handles)		2) View of Collioure
	Matisse, Henri 1907		Matisse, Henri Circa 1905
	3) Seville Still Life		4) Zorah Standing
	Matisse, Henri Circa 1910/1911		Matisse, Henri 1912
	5) Portrait of Suzanne Dufy, the		6) Composition

State Hermitage Museum

Search by Shape



Photobook search










Query by Example



SQUID (Surrey Univ.)

Click here to have random contours

Click on your *desired image* to see its similar images.

 <p>match value= 0</p>	 <p>match value= 3</p>	 <p>match value= 4</p>
 <p>match value= 4</p>	 <p>match value= 4</p>	 <p>match value= 4</p>
 <p>match value= 4</p>	 <p>match value= 4</p>	 <p>match value= 4</p>

Query by Spatial Relation



- VisualSEEk system (Columbia Univ.)

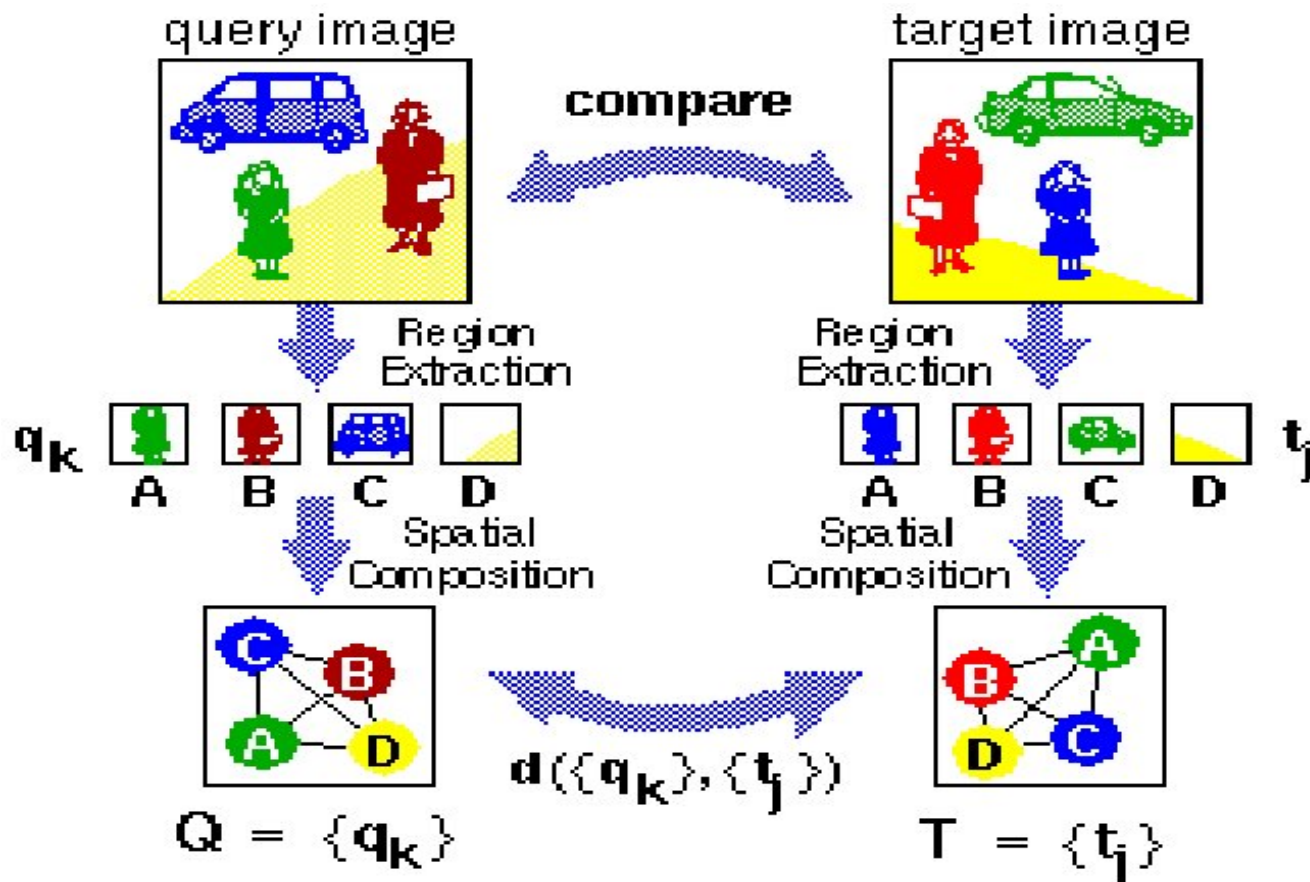


Image Search Using Regions



- Blobworld: Image retrieval using regions

- <http://elib.cs.berkeley.edu/photos/blobworld/>

系統功能示範。以內容為主的圖形檢索系統，可用形體搜尋圖檔中類似的物件。例如用老虎的形體檢索含類似形狀的圖檔。

Keywords (optional):

Step 2:
Adjust the weights below if you'd like, then click "Submit."

How important is the selected region?	<input type="radio"/> Not	<input type="radio"/> Somewhat	<input checked="" type="radio"/> Very
How important are the features of this region?			
Color	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>
Texture	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>
Location	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>
Shape/Size	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>

How important is the background (everything outside the region)?	<input type="radio"/> Not	<input checked="" type="radio"/> Somewhat	<input type="radio"/> Very
--	---------------------------	---	----------------------------

Relevant Feedback



圖片內容相似檢索 (30000張推薦圖庫)

 <input type="radio"/> 相似 <input type="radio"/> 不相似	 <input type="radio"/> 相似 <input type="radio"/> 不相似	 <input type="radio"/> 相似 <input type="radio"/> 不相似	 <input type="radio"/> 相似 <input type="radio"/> 不相似	 <input type="radio"/> 相似 <input type="radio"/> 不相似	 <input type="radio"/> 相似 <input type="radio"/> 不相似	 不相似
 <input type="radio"/> 相似 <input type="radio"/> 不相似	 <input type="radio"/> 相似 <input type="radio"/> 不相似	 <input type="radio"/> 相似 <input type="radio"/> 不相似	 <input type="radio"/> 相似 <input type="radio"/> 不相似	 <input type="radio"/> 相似 <input type="radio"/> 不相似	 <input type="radio"/> 相似 <input type="radio"/> 不相似	 不相似
 <input type="radio"/> 相似 <input type="radio"/> 不相似	 <input type="radio"/> 相似 <input type="radio"/> 不相似	 <input type="radio"/> 相似 <input type="radio"/> 不相似	 <input type="radio"/> 相似 <input type="radio"/> 不相似	 <input type="radio"/> 相似 <input type="radio"/> 不相似	 <input type="radio"/> 相似 <input type="radio"/> 不相似	 不相似

共找到 100 圖片 (1-18) 頁: [1](#) [2](#) [3](#) [4](#) [5](#) [6](#)

相似程度 傳回圖數

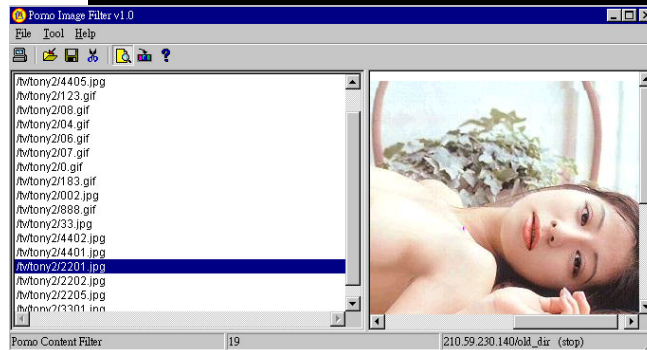
請至少選擇一個相似圖片, 按 開始查詢, 或按 重新選擇

• VisionNEXT 新視科技

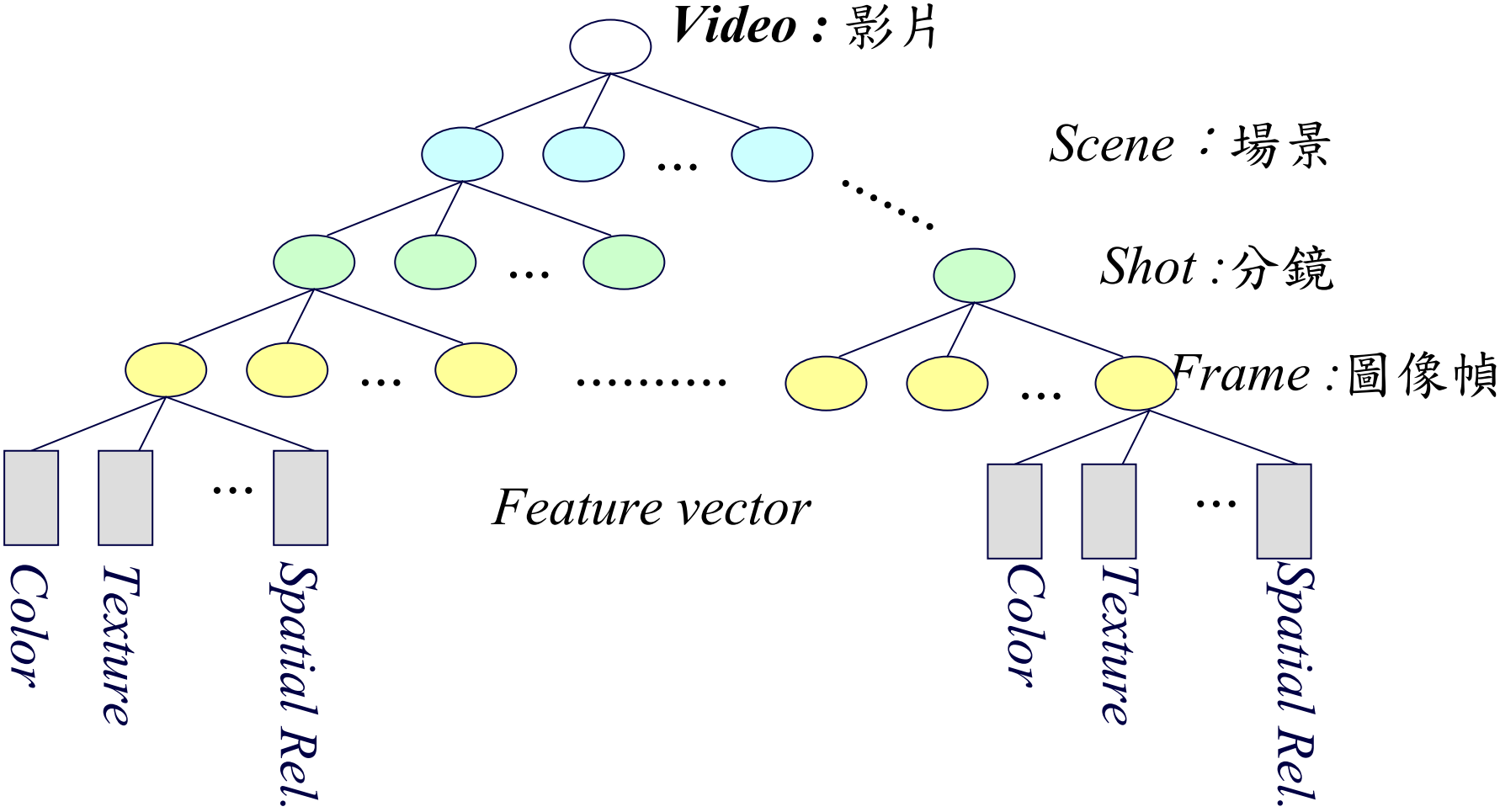
以圖找圖(CBIR)應用



內容相似圖片搜尋系統



Video Hierarchical Structure



以圖找圖結合視訊搜尋



annext

主題	中華民國國軍招募短片
描述	實現理想·從軍開始
關鍵詞	結婚黨◆國軍招募中心
作者	國軍人才結構中心製
長度	00分38.6秒
場景數	15
檔案大小	686k Bytes
檔案格式	Real Audio
製片時間	3/18/2001 22:44:59
播放品質	Corporate LAN (150 Kbps)

共有 15 場景

場景 1 場景 2 場景 3 場景 4 場景 5

場景 6 場景 7 場景 8 場景 9 場景 10

場景 11 場景 12 場景 13 場景 14 場景 15

實現理想·從軍開始
080-000-050
國軍人才結構中心

排列方式 4 行 4 列 選擇 共 100 張圖 7 頁 回網頁

頁次: 1/7 1 2 3 4 5 6 7

頁次: 1/7 1 2 3 4 5 6 7

網站聲明: real提供網際多媒體搜尋服務,在茫茫的網海中,協助網友搜尋相關圖像、音樂、動畫電影、網頁、網站等各種相關物件資訊。為維護原創作者之智慧財產權,所有圖像僅供網友搜索瀏覽使用,不得他用。若從事任何商業並利行為,須事先徵求原所有權人的同意。若有多媒體物件所有權人對我們以網路方式來推廣您的網頁或網站,認路有不安或侵犯之意,我們願意盡心求教,並將所有連結移除,以示尊重之意! 有任何問題或意見,歡迎與我們聯繫!

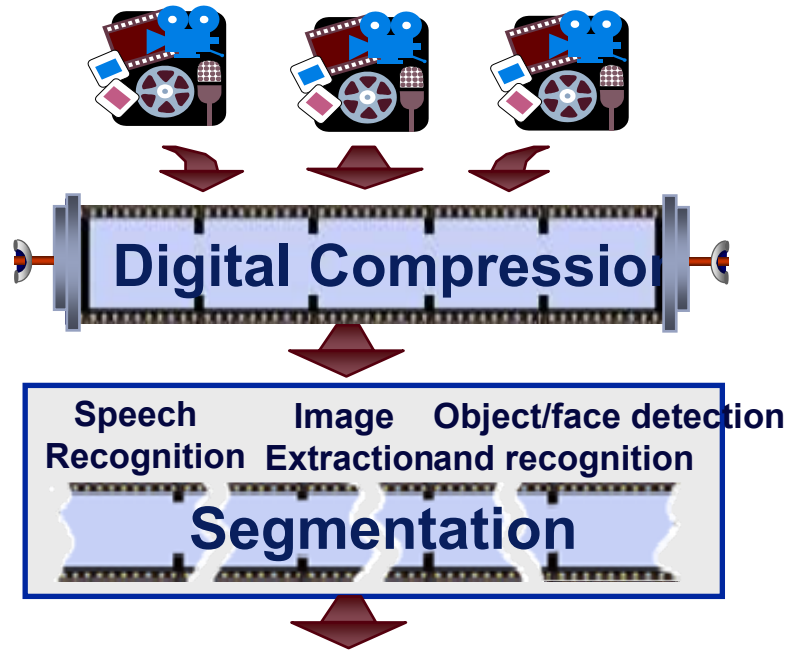
Copyright (C)2000 VisionNEXT Limited. All rights reserved.
最佳瀏覽環境: 800X600

ECHO System Overview

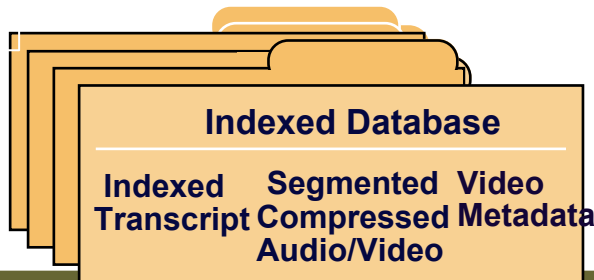


Library Creation

Offline

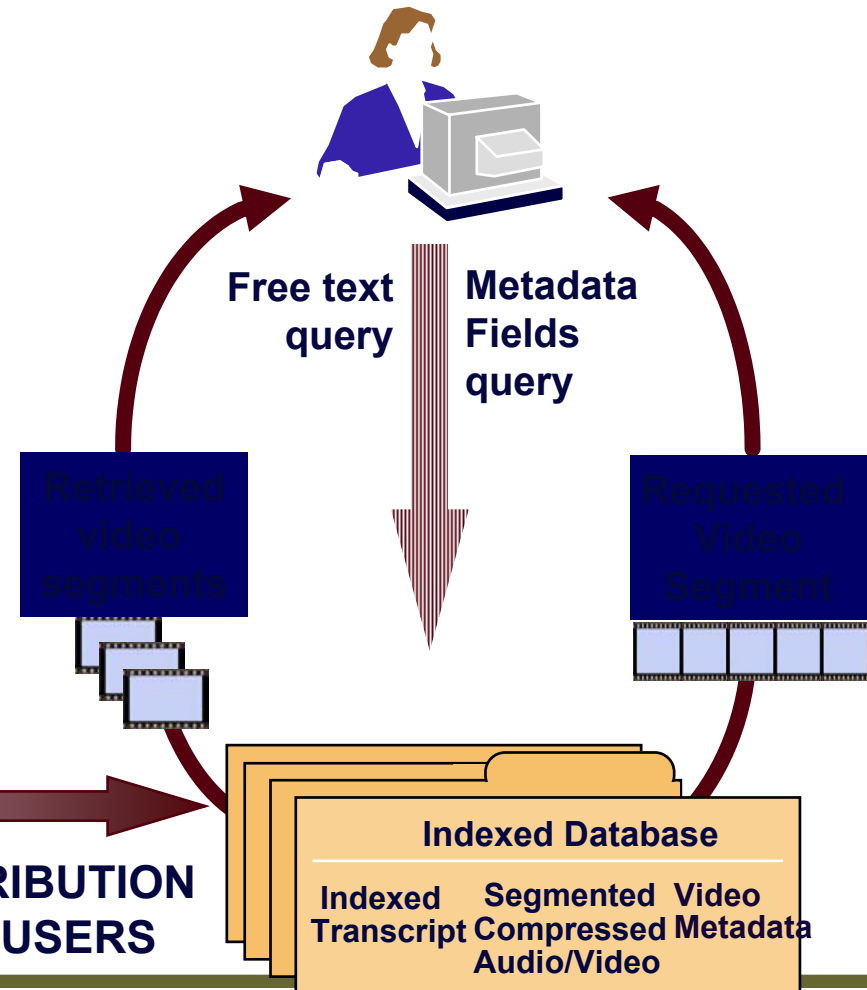


Insertion of video metadata



Library Exploration

Online



相關網站



● 中研院Sovideo 新聞搜尋系統

- <http://sovideo.iis.sinica.edu.tw/>

● TV Search Engine

- <http://www.blinkx.com/overview.php>



● Virage video search solution provider

- <http://www.virage.com/>



● Google Video Search

- <http://video.google.com/>

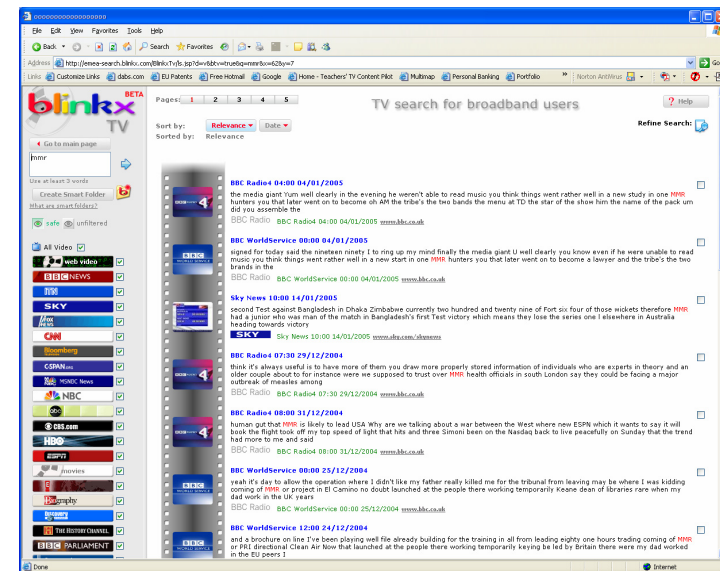
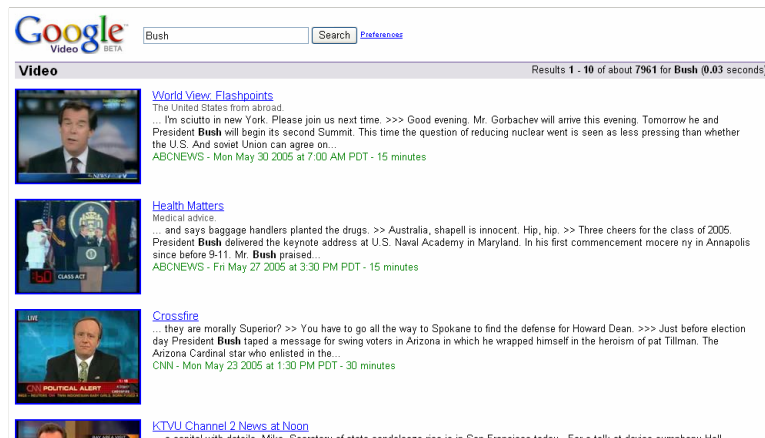


Image Search Difficulties



● Content-based approaches

- features not always useful (e.g., a collection of tree images)
- segmentation (semi-automatic)
- narrow domain

● Context-based approaches

- information loss (e.g., colour)
- which text relates to which image
- Limited, hard to express

● Demo

- CIRES (Content-based Image Retrieval System):
<http://amazon.ece.utexas.edu/~qasim/research.htm>
- SIMPLIcity: <http://www-db.stanford.edu/IMAGE/>

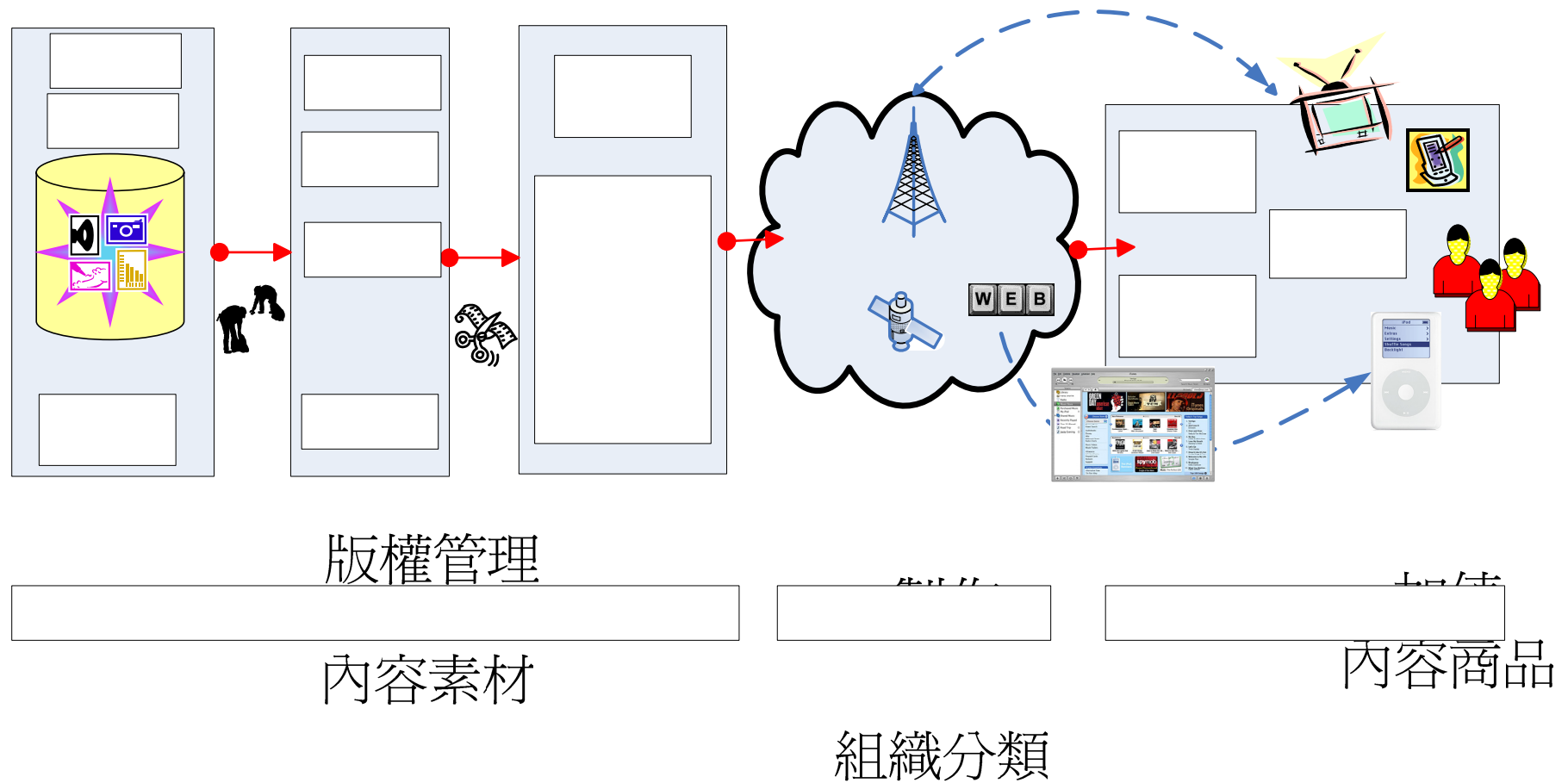


數位典藏

Digital Archive Management



Digital Value Chain



數位典藏的基本概念



- 可數位化的內容
 - 平面 / 立體 實體...
- 數位化目的
 - 資料保存 / 資訊傳播 / 增值應用與公共服務
- 數位化後階段
 - 數位化成果的展示與利用
 - 建立典藏品標準化管理(SOP)作業流程
 - 建立數位典藏內容管理機制
 - 管控目前數位化的進度與數量
 - 多媒體物件與metadata 資料的整合管理與搜尋
 - 建立未來多媒體資料交換的基礎
 - 邁向數位內容創意增值應用與服務的目標

典藏管理技術包羅萬象(1)



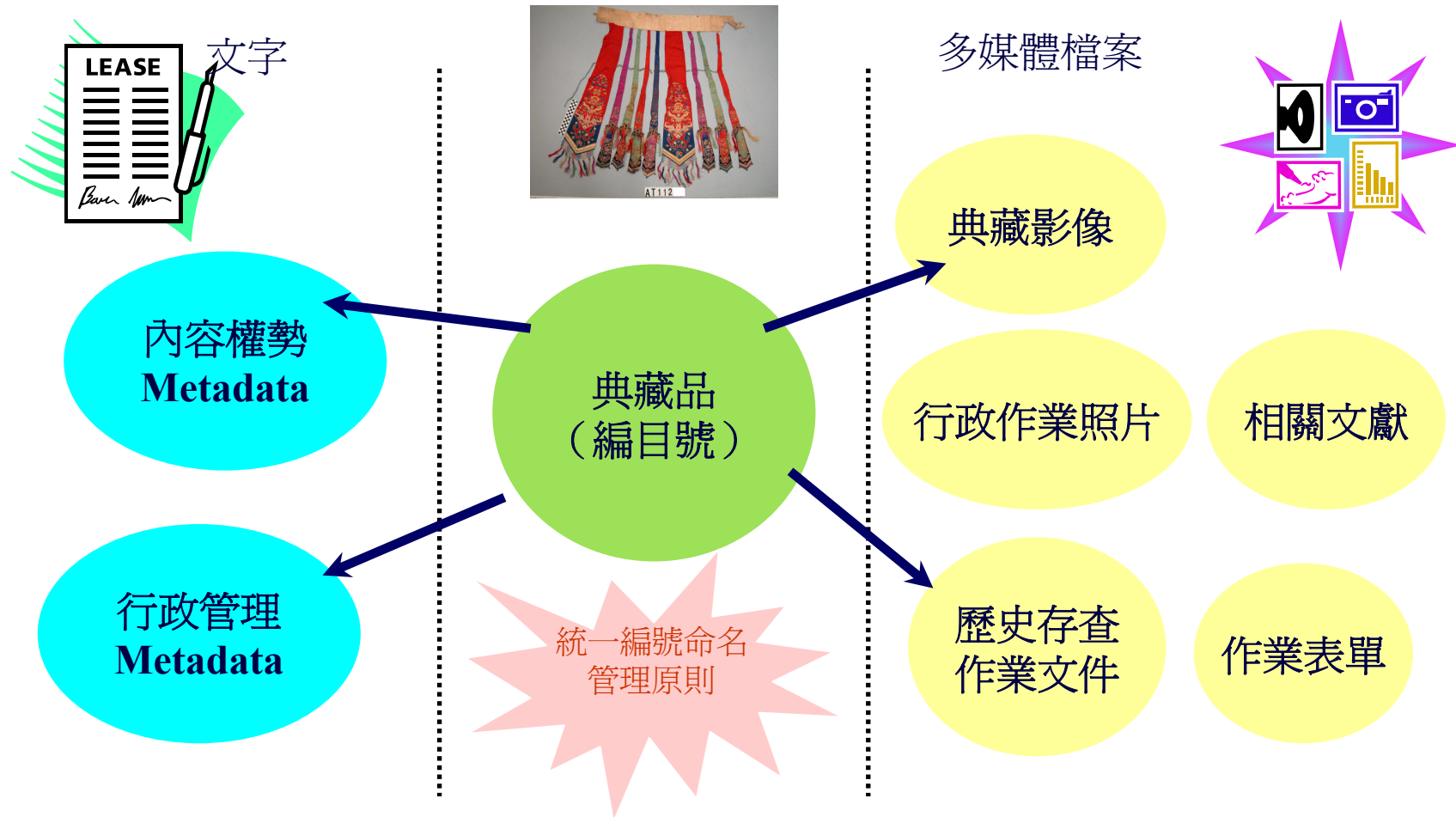
- 圖書館編目管理
 - Metadata DTD Schema制訂
 - 權威詞彙控制
 - XML 與資料庫處理技術
 - 目錄分類
 - 國際數位物件管理交換標準
 - OAI (Open Archive Initiative) Protocol
 - CNRI Handle System
 - 多媒體搜尋技術
 - **Time Based Media Support**
 - Audio, Video, Streaming
 - 多媒體檔案處理
 - 數位物件檔案保存 / 轉檔技術 (transcoding)
 - 浮水印 / 大檔瀏覽 /
-
-

典藏管理技術包羅萬象(2)

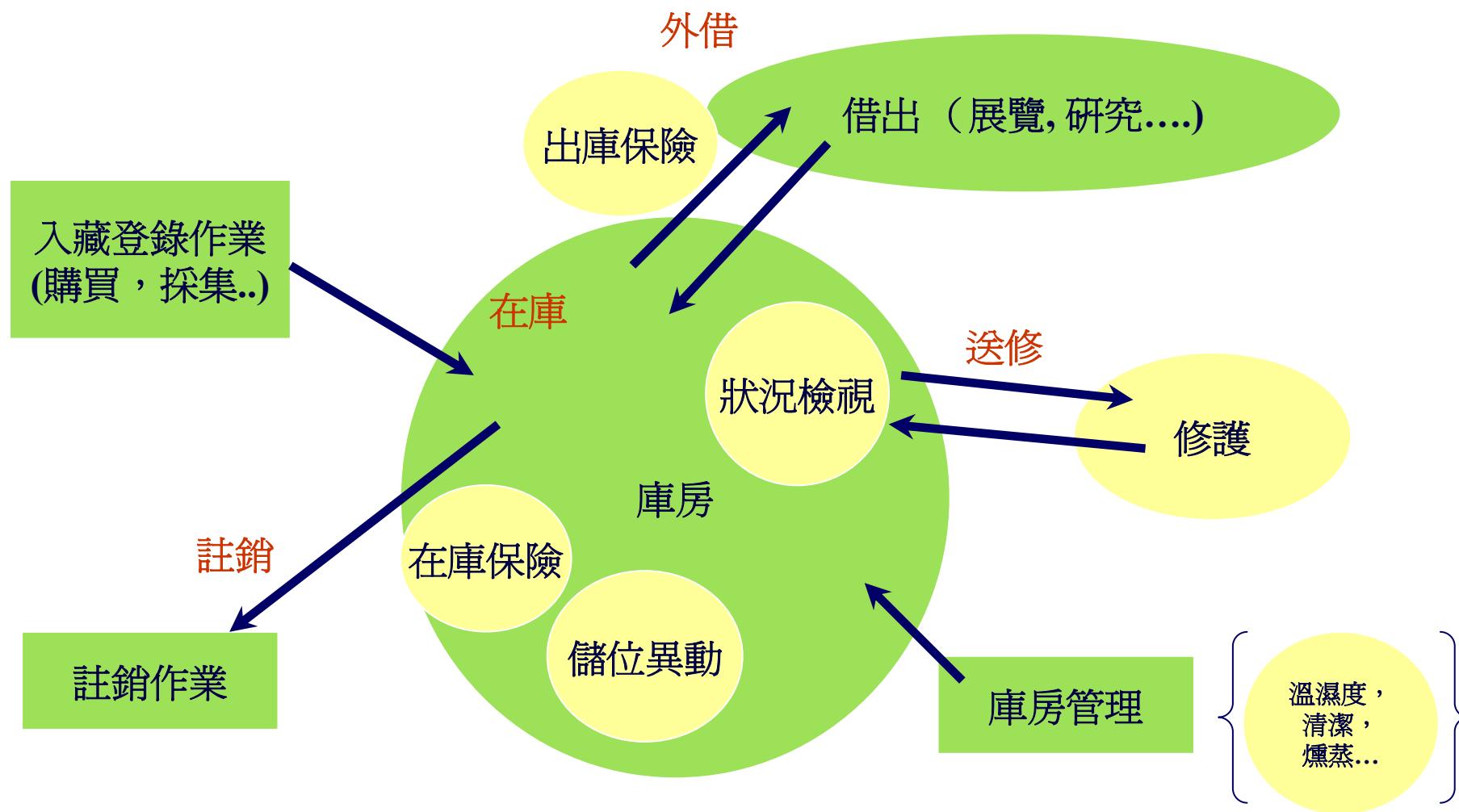


- 語言處理
 - UNICODE、UTF8、BIG5編碼
 - 語言處理（跨語、段詞、詞彙處理...）
 - 資料處理
 - 格式交換
 - 聯合目錄 DC (Dublin Core) 匯出匯入格式轉換
 - 資料備份, Migration
 - 資訊安全
 - 特殊工具整合
 - Ep: GPS
 - Scalability
 - DRM (Digital Right Management)
-
-

系統管理哪些資源？



藏品管理生命週期



技術加值鏈



Delivery / Display

e-Commerce

價值
創新

Watermark

2D/3D
Browsing

e-learning

Multi-casting

OAI

DOI

VisionNEXT
Value-added
Service
Platform

XML

Content Creation
Authoring

Digital Objects

Management

利用

Multimodal
Search

Speech
Retrieval

CBIR

Category

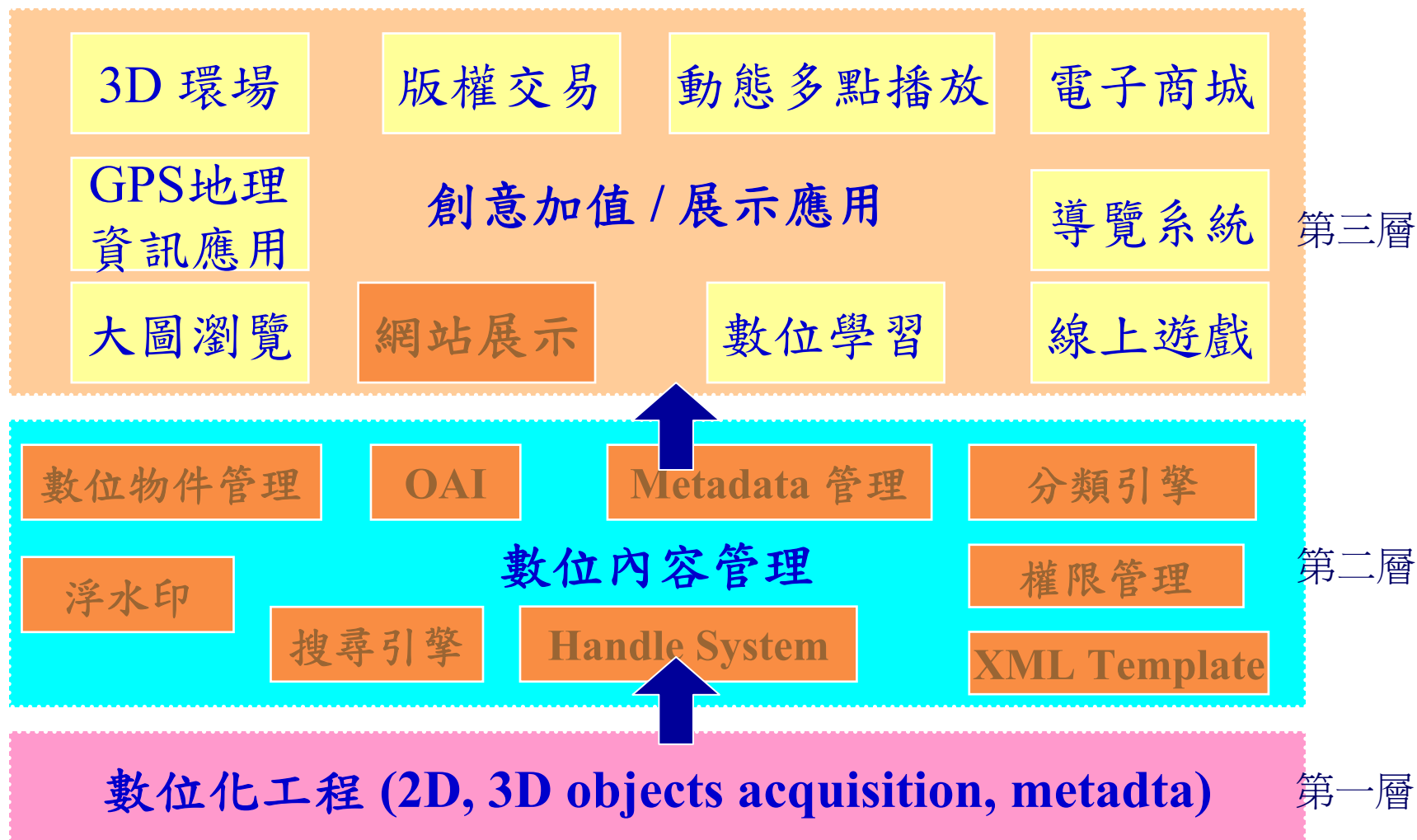
Metadata

保存

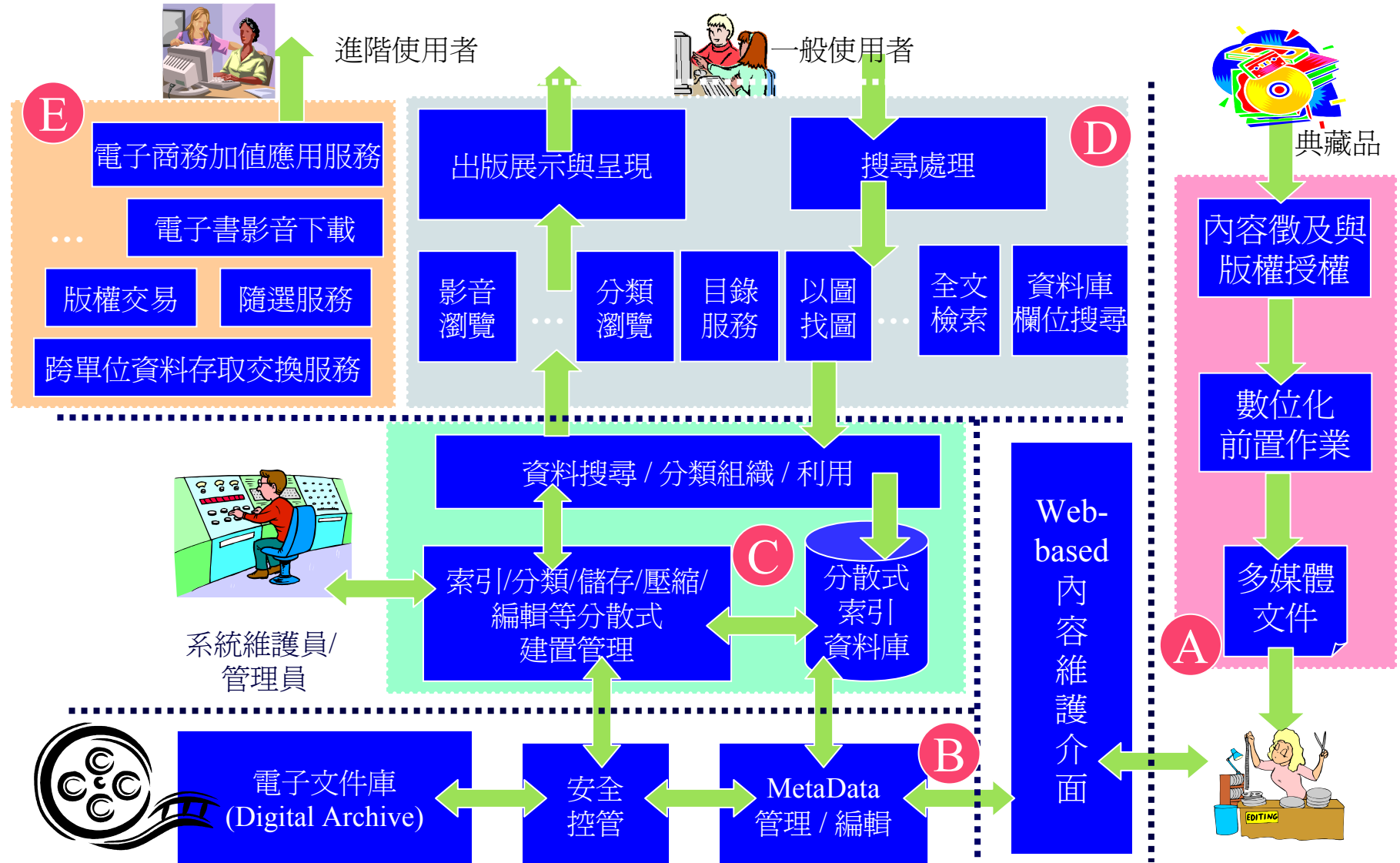
Search



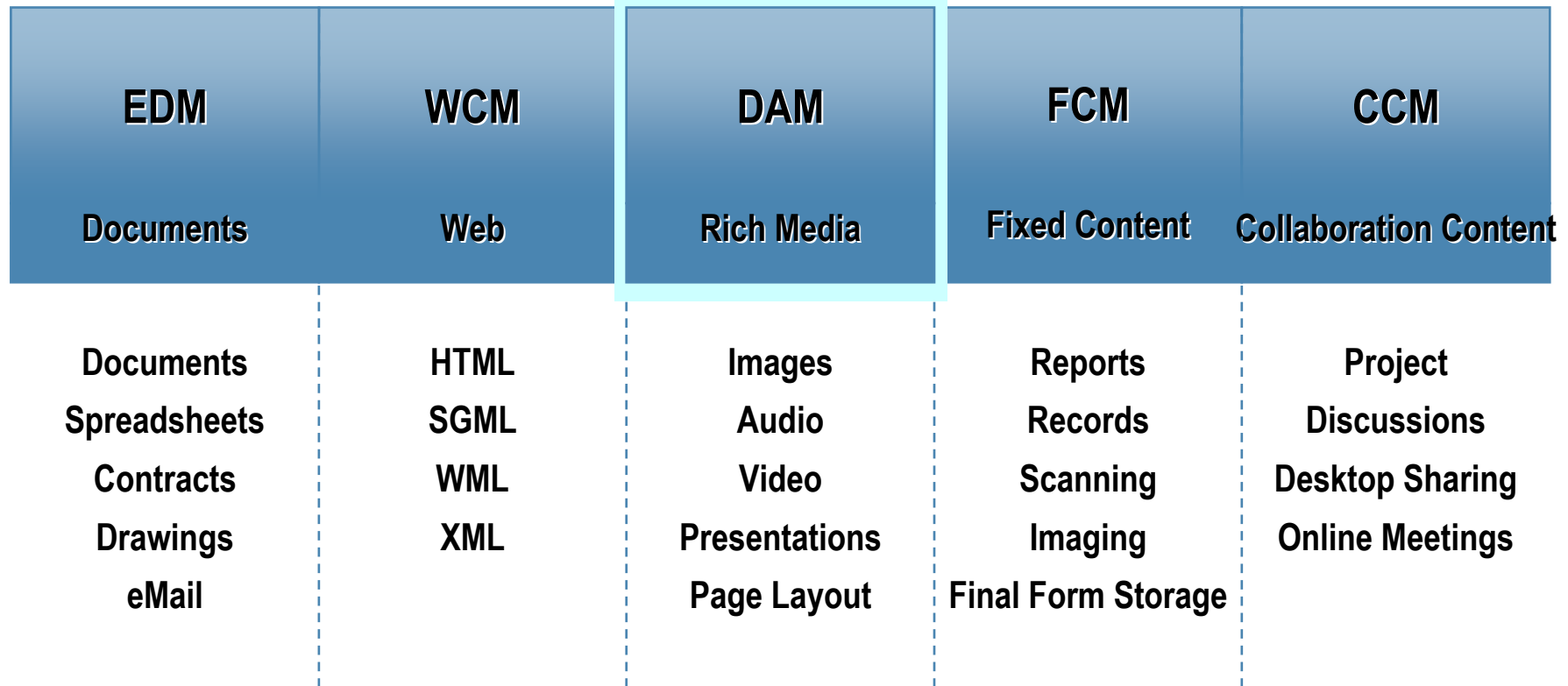
數位典藏延伸應用資訊技術

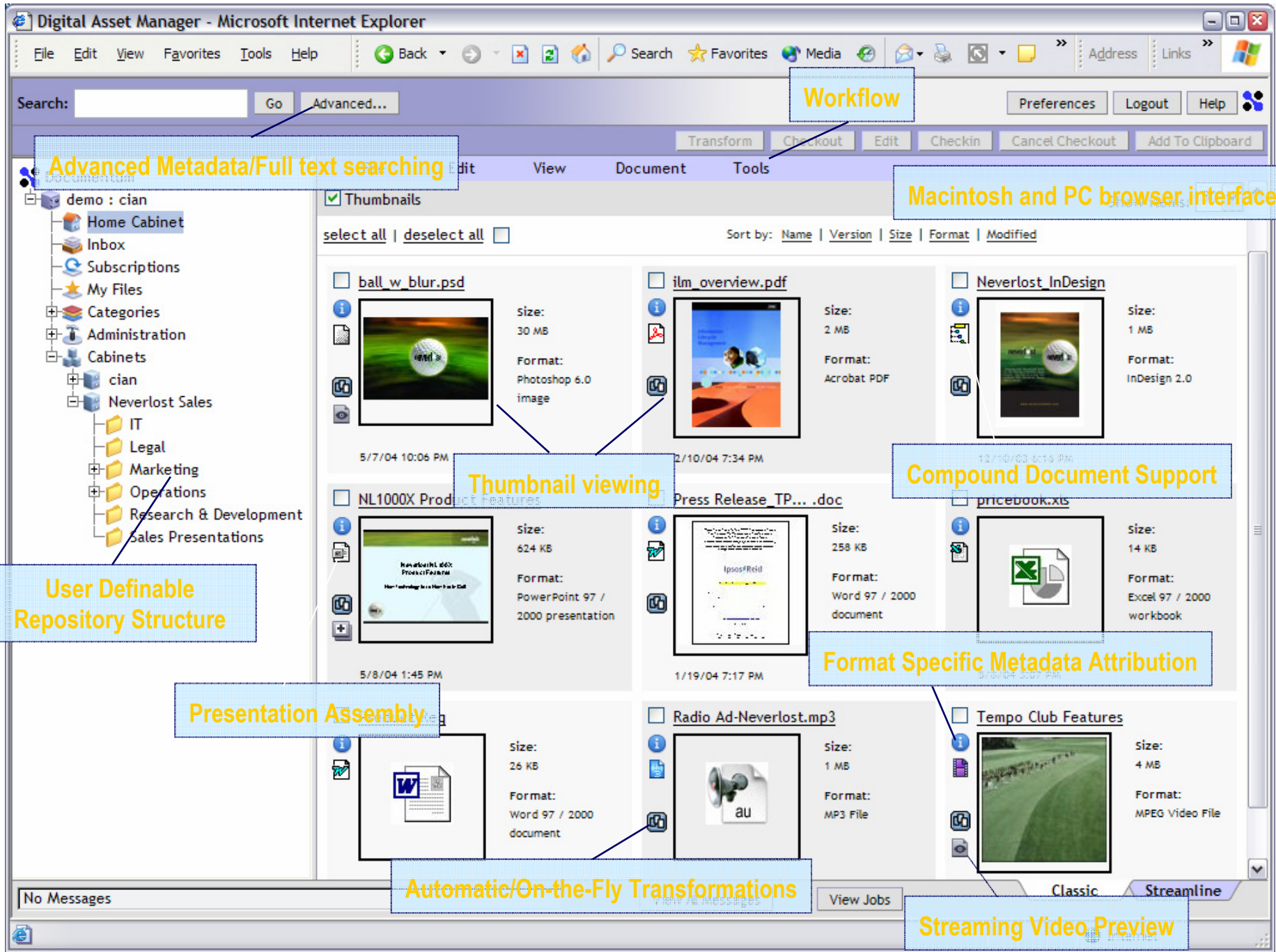


管理與增值服務系統



Enterprise Content Management Content Types





Workflow

Advanced Metadata/Full text searching

Macintosh and PC browser interface

User Definable Repository Structure

Thumbnail viewing

Compound Document Support

Format Specific Metadata Attribution

Presentation Assembly

Automatic/On-the-Fly Transformations

Streaming Video Preview

Video Support



onnext

video_details_04.jpg - Windows Picture and Fax Viewer

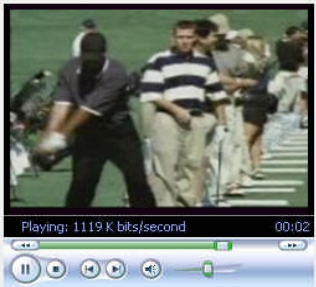
Search:

File Edit View Document Tools

Documentum

- Board board
- dam
- eLearning_525
- inlder
- matrix
- ml04
- root
- roja
- roy_wk3 : dmadm
- Home Cabinet
- Inbox
- Subscription
- Av File:
- Categories:
- Administration
- Cabinets
- walkstreet
- WinFun_50Sep1
- v2

Video Details



Playing: 1119 K bits/second 00:02









Mark In:

Mark Out:

Name: Neverlost_golf_commercial.mpeg
Subject: Tiger Woods at the Driving Range
Title: NeverLost Golf Commercial
Format: MPEG Video File
Size: 4 MB
Modified: 4/14/04 4:36 PM

Info Storyboard Captions Notes Sub-Clips

Thumbnails Show items: 10

Start	End	Caption
<input type="checkbox"/>  00:00:02:29	 00:00:04:29	Playing with NeverLost balls is like having a faithful, loyal friend - like having a dog, but without the concerns about one's carpet.
<input type="checkbox"/>  00:00:05:29	 00:00:08:29	When NeverLost launched its first ball in 1952, it was an immediate, international sensation. Golfers from Kelowna to Kabul delighted in SmartSound Technology, and its dimpled good looks.
<input type="checkbox"/>  00:00:09:29	 00:00:17:29	Now, NeverLost is helping golfers the world over take their game even further - with this new long-distance wonder, the 1000x. The core flexes on impact for optimum distance, and the dimple configuration allows the ball to cut through the air like a hot knife through butter.
<input type="checkbox"/>  00:00:19:29	 00:00:25:29	Once you've tried it, you'll never want to play with another ball again!



Web-based Video Management



EVIADA - Annotation - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Refresh Home Search Favorites Media Links

Event: Gbarbea Funeral 10

[00:08:46]

Mark Scene Mark Action Start

Action 2.2 [00:08:46 - 00:11:00]

Heading:

Language Ethnic Group

Geography Venue Genre Instrument **Participant** Theme Performance

Add Clear

Participant Role : Ritual Role :

Last Name : First Name :

Display Name :

Update

Timeline: Scene 1 (00:01:00 - 00:05:00), Scene 2 (00:10:00 - 00:15:00), Scene 3 (00:20:00 - 00:20:00)

Actions: Action 1.1, Action 1.2, Action 2.1, Action 2.2, Action 3.1, Action 3.2, Action 3.3

Scene	Action	Description	Start Time	End Time	Duration
Scene 1	Action 1.1	Band going down highway	00:00:00	00:01:43	1 min 43 sec
	Action 1.2	Bishop arrives	00:01:43	00:04:50	3 min 7 sec
Scene 2	Action 2.1	Music and Procession restart	00:06:00	00:08:46	2 min 46 sec
	Action 2.2	Choir begins	00:08:46	00:11:00	2 min 14 sec
Scene 3	Action 3.1	Coffin with followers	00:13:45	00:15:10	1 min 25 sec
	Action 3.2	Hymn: God be with you	00:16:00	00:17:22	1 min 20 sec
	Action 3.3	Benediction	00:17:22	00:19:20	1 min 38 sec

Save Cancel

加值服務:複合式多媒體檢索



數位典藏複合式多媒體檢索技術



本研究計畫目標為提供前端使用者一個複合式 (multi-model) 人性化的互動搜尋人機介面，透過語音檢索、文字檢索、基於內涵相似之圖像檢索等先進技術的整合，讓一般使用者可以在有線或無線網路環境下，方便與快速地瀏覽與檢索數位典藏文物，過程中可任意切換或選用一種多媒體資訊輸入來作為持續性互動的媒介，最後達成資訊檢索利用的目的。

合作單位：國立台灣師範大學資訊工程所、國立歷史博物館、新視科技股份有限公司

系統特色

- ★ 結合豐富博物館數位典藏多媒體內容
- ★ 適用於PDA無線網路與Web環境下
- ★ 提供語音辨識 / 以圖找圖之多媒體資訊檢索技術
- ★ 系統採分散式主從設計架構
- ★ 複合式使用者檢索介面

Q&A

數位年代是屬於善用資訊的人的時代！



Thanks