

Web Madenciliği

26 Eylül 2017

Dr. Mehmet Sıddık Aktaş

Yıldız Teknik Üniversitesi

Bilgisayar Mühendisliği Bölümü



World Wide Web (Web, WWW, W3)

WWW Gelişim Süreci

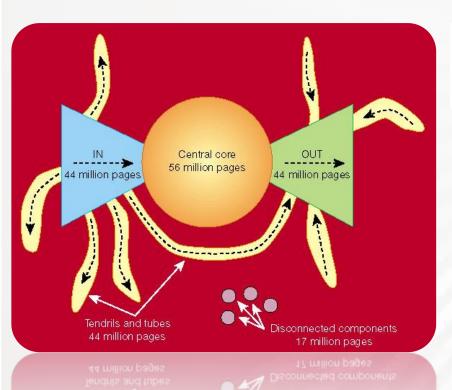




- CERN (the European Organization for Nuclear Research)
 - Tim Berners-Lee
- GUIs
 - Berners-Lee (1990) (HTML, HyperText Transfer Protocol HTTP, Web Browser), HTTP Web Server
 - Erwise and Viola(1992), Midas (1993) (Initial GUI based Browsers)
- Mosaic (1993)
 - National Center for Supercomputing Applications (NCSA)
 - a hypertext GUI for the X-window system
 - HTML: markup language for rendering hypertext
 - HTTP: hypertext transport protocol for sending HTML and other data over the Internet

World Wide Web



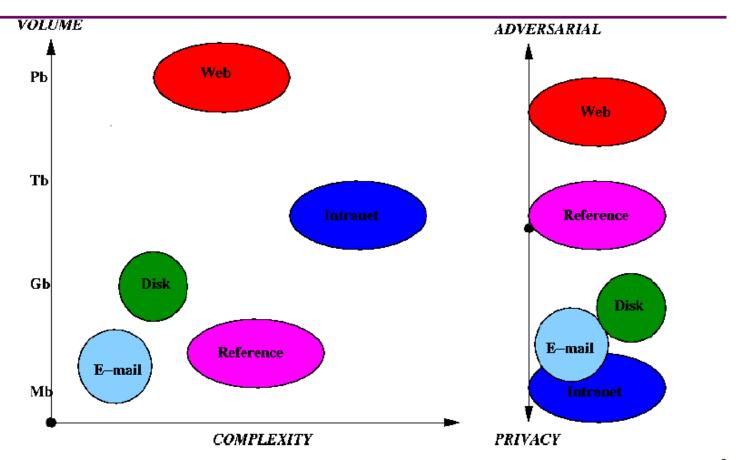


The Web is a bow tie

- Hypertext dökümanları
 - Text, Multimedia
 - Links
- •Farklı doküman tipleri
 - •PDF, Word, Excel ...
- Web
 - Milyarlar mertebesinde döküman,
 - •Milyonlar mertebesinde farklı yazar var.
 - Herhangi bir merkezi editor yok.
 - Milyonlarca dağıtık
 bilgisayar üzerinde çalışıyor.
 - Farklı iletişim araçları kullanılarak ulaşılabiliyor.
 - •Internet üzerinde çalışıyor.

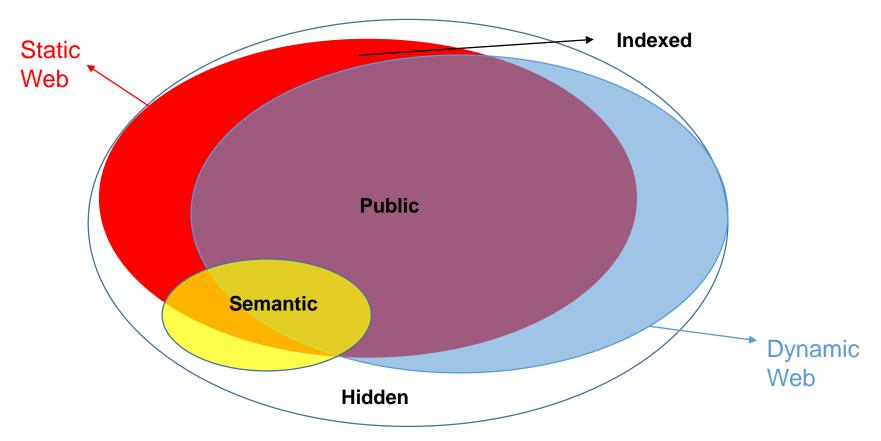


Different Views on Data



http://ijcai-11.iiia.csic.es/files/proceedings/T16-Web-Mining.pdf

The Different Facets of the Web



http://ijcai-11.iiia.csic.es/files/proceedings/T16-Web-Mining.pdf

http://hal.archives-ouvertes.fr/docs/00/05/44/58/PDF/webgraph.pdf

Jean-Loup Guillaume and Matthieu Latapy

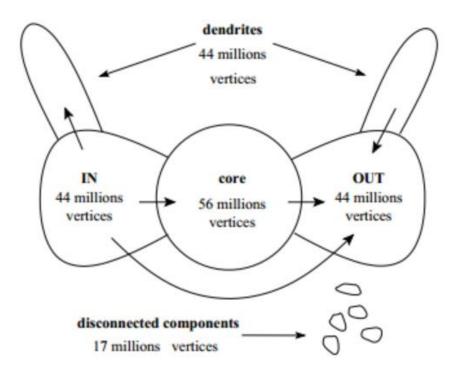
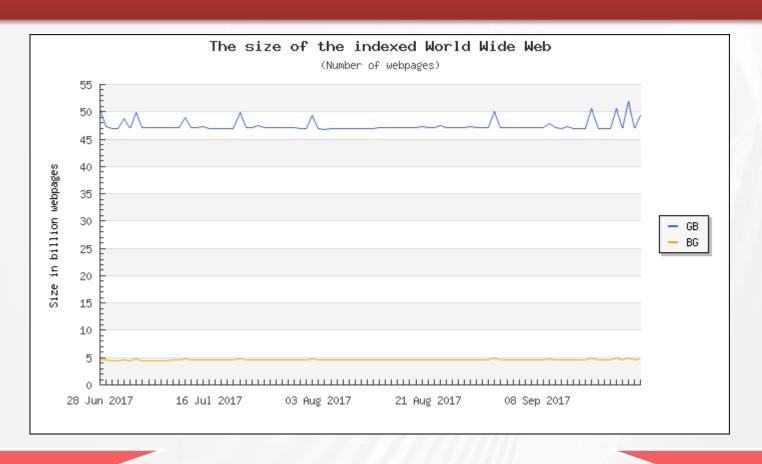


Fig. 3: The bow-tie macroscopic structure of the Web graph [BKM+00]: the core, the IN component, the OUT component and the dendrites. Each of these parts contains around one quarter of the pages, the disconnected part being reduced to less than 10% of the whole.

Web Ne Kadar Büyük?





Web Ne Kadar Büyük?

www.worldwidewebsize.com 26 Eylül 2017



Web Madenciliği

Internet Üzerindeki Veri Yığınları



- Web sayfaları
- ☐ Erişim Log dosyaları
- Kullanıcı kayıt bilgileri
- Oturum ve hareket bilgileri
- ☐ Site yapısı ve içeriği

Veri ve Web Madenciliği



Web Madenciliği

☐ Çeşitli yapıdaki web sayfası dokümanlarını, içeriklerini, link yapılarını ve kullanım bilgilerini incelemek, bunlardaki anlamlı bilgileri keşfetmek için veri madenciliği tekniklerinin kullanılmasıdır.

Web Madenciliği Teknikleri



- Web madenciliği kullanılan verilerin yapısına göre 3 gruba ayrılır.
 - Web içerik madenciliği
 - □ Text ve multimedia madenciliği
 - Web yapı madenciliği
 - Link analizi, grafik madenciliği
 - Web kullanım madenciliği
 - Log analizi, sorgu madenciliği
- Web madenciliği teknikleri beraber kullanılabilir.
 - Web' in karakteristik özellikleri
 - Uygulamalar (Ör: Arama motoru)



- The Web as an object
- User-driven Web design
- Improving Web applications
- Social mining
-



- James Surowiecki, a New Yorker columnist, published this book in 2004
 - "Under the right circumstances, groups are remarkably intelligent"
- Importance of diversity, independence and decentralization
 Aggregating data

"large groups of people are smarter than an elite few, no matter how brilliant—they are better at solving problems, fostering innovation, coming to wise decisions, even predicting the future".













Sign In | Help

Home | Sign Up

Photos: Explore Flickr • Learn More



Tags / jaguar / clusters



SEARCH

(Or, try an advanced search.)











car, cars, auto, etype, automobile, classic, vintage, autoshow, red, show

See more in this cluster...











zoo, animal, cat, animals, bigcat, seattle, woodlandparkzoo, sleep, edinburgh, caged

See more in this cluster...











guitar, fender

See more in this cluster...









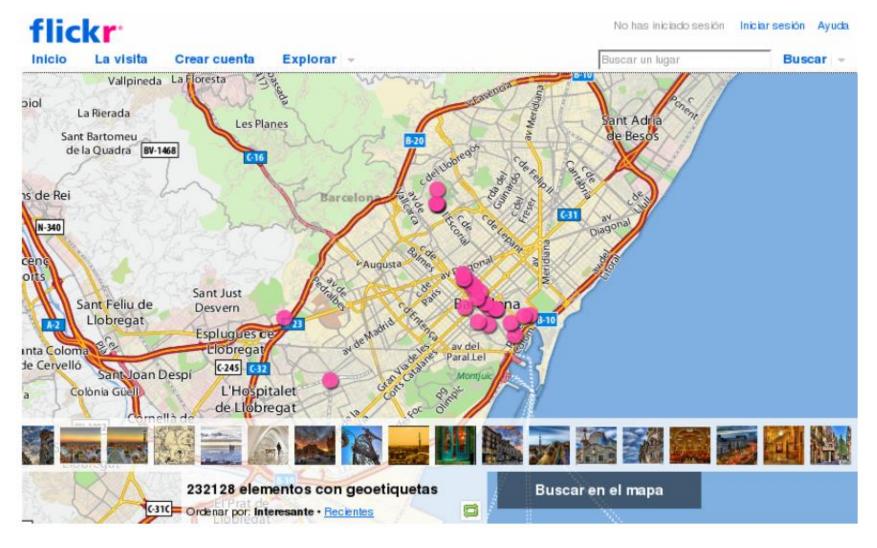


aircraft, raf

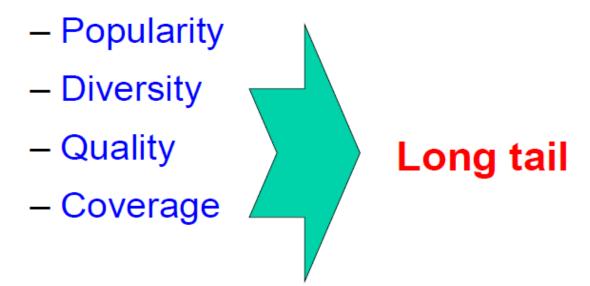
See more in this cluster...



Flickr: Geo-tagged pictures









Explore Flickr through tags

canon china christmas city concert england europe family festival flower flowers food france friends fun germany green italy japan london music nature new newyork night nikon nyc paris park party people portrait red sanfrancisco sky snow spain street summer sunset taiwan travel trip uk usa vacation water wedding white winter



The Wisdom of Crowds

- Crucial for Search Ranking
- Text: Web Writers & Editors
 - –not only for the Web!
- Links: Web Publishers
- Tags: Web Taggers
- Queries: All Web Users!
 - –Queries and actions (or no action!)

Web İçerik Madenciliği

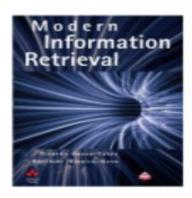


■ Web İçerik Madenciliği nedir?

- Web kaynaklarının içeriklerinden yararlı bilgiyi elde etmek olarak tanımlanabilir.
- Web sitelerinin dokümanlarındaki terimleri ve link' leri bularak sayfanın indekslenmesi ve web sitesinin yapısı ortaya çıkarılmaya çalışılır.



What is Information Retrieval (IR)?



IR: Part of computer science which studies the retrieval of information (not data) from a collection of written documents. The retrieved documents aim at satisfying a user information need usually expressed in natural language.

- Documents, unstructured, text, large
- Information need
- Store, search, find
- The World Wide Web?
- Relational databases?

DIKW

• Data: Raw web pages

• Information: Result of query

• Knowledge: Result of processing query result by user

• Wisdom: Synthesis of many such actions by a set of users

One possible classification of steps in process



Information Retrieval vs. Databases

Information retrieval	Data retrieval
Retrieve all objects relevant to some information need	Retrieve all objects satisfying some clearly defined conditions
Find all documents about the topic "semantic web"!	SELECT id FROM document WHERE title LIKE '%semantic web%'
Result list	Well-defined result set



[selke@tbdb ~]\$ db2 "SELECT id FROM document WHERE title L IKE "%semantic web%" FETCH FIRST 3 ROWS ONLY" ID

45489
9635899
98556
3 record(s) selected.



- Very similar to information retrieval
- Main differences:
 - Links between Web pages can be exploited
 - Collecting, storing, and updating documents is more difficult
 - Usually, the number of users is very large
 - Spam is a problem



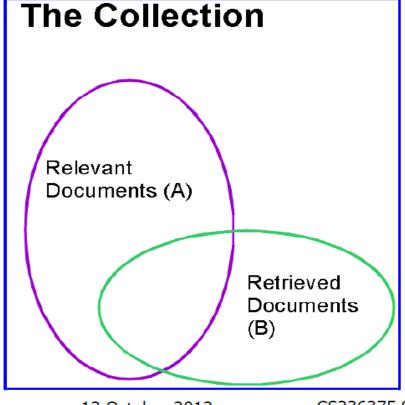




- Any IR system is based on an IR model
- The model defines ...
 - ... a query language,
 - ... an internal representation of queries,
 - ... an internal representation of documents,
 - a ranking function which associates
 a real number with each query-document pair.
- Optional: A mechanism for relevance feedback







- Given a query, the system retrieves a set B of documents
- Every retrieved document is either relevant or irrelevant to the query

Quality metrics:

- <u>Recall</u>: (A∩B) / A
- <u>Precision</u>: (A∩B) / B

13 October 2013

CS236375 Search Engine Technology



Recall and Precision on the Web

- Relevance of document to queries is not binary there are many shades of gray
- Broad-topic queries:
 - abundance problem
 - Precision is the dominating factor: users mostly satisfied with a few good results (a few authoritative pages)
- Narrow-topic queries:
 - Find a needle in an enormous haystack
 - Recall demands engines cover significant portions of the Web
- Common measure: precision@10
- Nowadays larger emphasis on diversity
 - Positive recall for many aspects of the query

13 October 2013

CS236375 Search Engine Technology

New Google Update "Improves Diversity" Of Domains In Results

Sep 14, 2012 at 5:03pm ET by Danny Sullivan

There's been chatter about Google having done some type of algorithm update this week that's having an impact on rankings. Now it's finally confirmed, a new change meant to allow more domains to appear in search results.

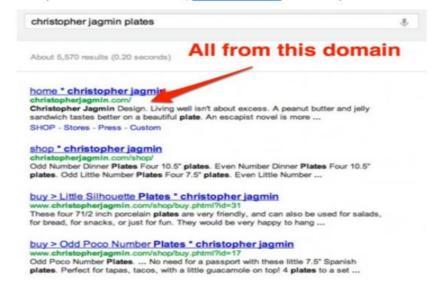


The head of Google's spam fighting team Matt Cutts tweeted this about it:

Just fyi, we rolled out a small algo change this week that improves the diversity of search results in terms of different domains returned.

Lately, more and more people have been noting that Google's search results can sometimes be dominated by pages that all come from the same domain. In other words, do a search, and all the listings seem to come from the same web site.

Here's an example of this from last month, when we wrote about the problem:



Diversity of Search Results



The Bag of Words Representation

- A very popular representation of documents is the bag of words model
- Each document is represented by a bag (= multiset) of terms from a predefined vocabulary
- Standard case:
 - Vocabulary
 - = set of all the words occurring in the collection's documents
 - Each document is represented by the words it contains

That's one small step for a man, a giant leap for mankind

http://www.ifis.cs.tu-bs.de/teaching/ss-11/irws



The Bag of Words Model (2)

Cons:

- Word order gets lost
- Very different documents could have similar representations
- Document structure (e.g. headings)
 and metadata is ignored

Pros:

- Simple set-theoretic representation of documents
- Efficient storage and retrieval of individual terms
- IR models using the bag of words representation work well!







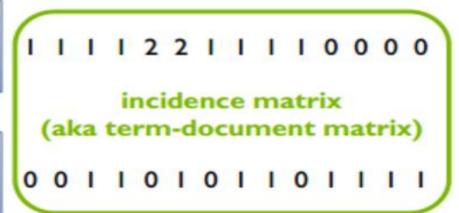
The Bag of Words Model (3)

 Any document can be represented by an incidence vector:

Step one step one step one step one step one step one step one tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tailkonau tai

That's one small step for a man, a giant leap for mankind

Taikonaut Zhai's small step is a giant leap for China



Web Yapı Madenciliği



■ Web Yapı Madenciliği nedir?

- Web yapı madenciliğin amacı web sayfaları arasındaki linkleri takip ederek bilgi üretmektir.
- Web yapı madenciliği, yapısal veri tiplerine göre 2 ye ayrılır.
 - □ Hyperlink bir web sayfasını faklı bir lokasyona yönlendiren yapısal eleman olduğu için webdeki hyperlinklerin modelinin çıkarılmasıdır,
 - Web sayfası dokümanlarındaki HTML ya da XML etiketleri analiz ve tanımlarında ağaç (tree) benzeri yapıların kullanılmasıdır.

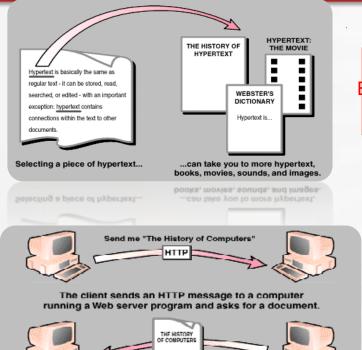
Karıştırılan Web Kavramları



- ☐ Internet Web
- ☐ Link Hyperlink
- □ Hypertext Hypermedia
- HTTP HTML
- Anchor Text URL

Karıştırılan Web Kavramları





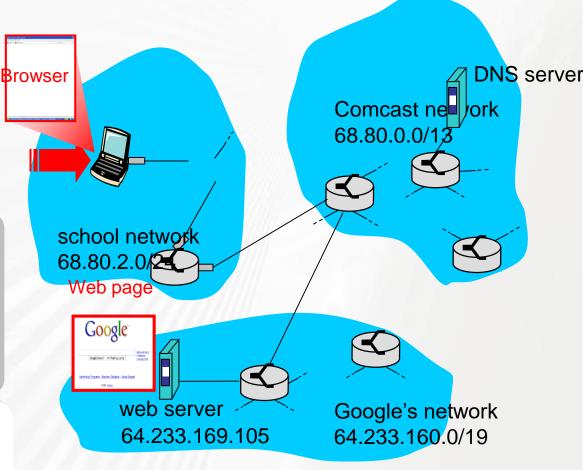
A hyperlink is a link you can click on or activate with the keyboard or other device in order to go somewhere else. The hyperlinks are used on the pages and they navigate from one page to another between same or different Sites/servers.

The Web server sends the hypermedia HTML document to the client.

You end up seeing the document on your screen.

A URL/Link can be thought of as the "address" of a web page and is sometimes referred to informally as a "web address."

Like if we write "http://google.com" in the address bar, it is a link and if we apply it on the page so that, anyone can navigate from the site to Google, it is called hyperlink.



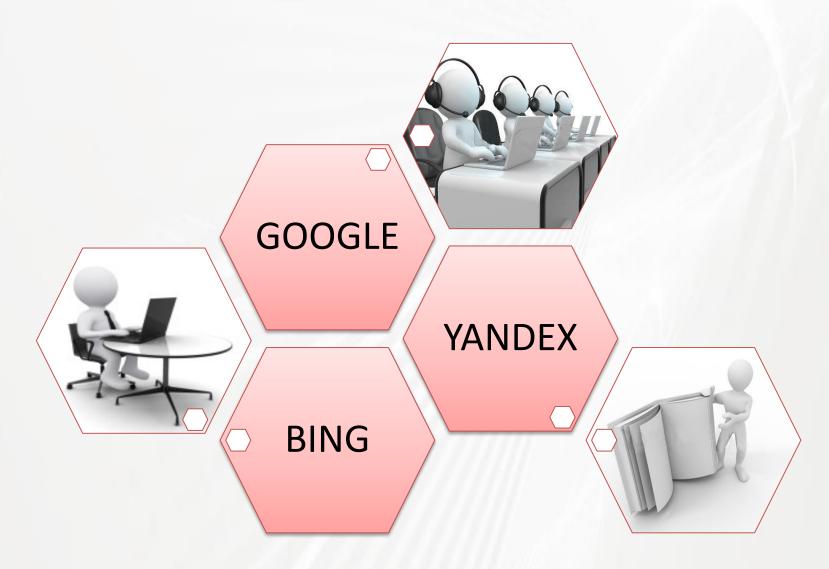
Anchor text usually gives the user relevant descriptive or contextual information about the content of the link's destination. The anchor text may or may not be related to the actual text of the URL of the link the English-language Wikipedia's homepage might take this form:

Wikipedia

The anchor text in this example is "Wikipedia"; the unwieldy URL http://en.wikipedia.org/wiki/Main_Page displays on the web page as Wikipedia, contributing to clean, easy-to-read text.

ARAMA MOTORLARI







Mission Impossible?

Search engines:

- Crawl and index tens of billions of documents
- Answer hundreds of millions of queries per day
- Devote less than 1 second of processing time per query execution

OSMELIES & SPIROLES WITH STORY

Users:

- Submit very short queries (averaging about 2.6 terms)
- Expect to receive the most relevant results on the Web
- In a blink of an eye

In terms of 1990 Information Retrieval research - almost unimaginable!



Web Searchers - Observations

- Make ill defined queries
 - Short (2.54 terms average, 80% contain less than 3 words)
 - Use imprecise (and often misspelled) terms
 - Unfamiliar with query syntax (80% queries without operator)
- Wide variance in information needs, expectations, education/knowledge, screen sizes, IP bandwidth, patience
 - Different modalities (mobile, desktop) = different needs and expectations, even with same person
- Specific behavior (desktop and laptop)
 - 85% look over one result screen only (mostly "above the fold")
 - 78% of queries are not modified
- Overall, we as users are investing low cognitive effort per query (formulating and looking at results)



Google Search: spears



Web Images Groups News Froogle more »

spears

Search

Advanced Search Preferences

Web

Results 1 - 10 of

News results for spears - View today's top stories



Knee Injury Closes Spears' Onyx Hotel - Billboard - 1 hour ago

Britney Spears' tour is canceled - San Diego Union Tribune - 7 hours ago
As fall approaches, Spears may start to smell Curious - Houston Chronicle - Jun 14, 2004

Britney Spears :: The Official Web Site

The Official Web Site of Britney Spears. Your official source for all things Britney. ...

Remember, proceeds benefit the Britney Spears Foundation. ... www.britneyspears.com/ - 41k - Jun 14, 2004 - Cached - Similar pages

Britney Spears - britney.com - Jive Records

iTunes. Real/Rhapsody. Napster. Under 11. www.britney.com/ - 10k - Cached - Similar pages

Britney Spears Portal - pics, lyrics, MP3s and more!

Britney Spears guide to Semiconductor Physics: semiconductor ...

Britney Spears lectures on semiconductor physics, radiative and non-radiative transitions, edge emitting lasers and VCSELs. ...

britneyspears.ac/lasers.htm - 13k - Cached - Similar pages

BritneySpears.org: Your online guide to Britney!

A comprehensive Britney Spears fansite which pays tribute to Britney with the most active message board, daily news, many pictures, desktop media and more. ... www.britneyspears.org/ - 78k - Jun 14, 2004 - Cached - Similar pages

Britney-Spears. To You! - The Britney Spears Community

Britney Spears: biography, discography, musics, real, mp3, videos, pictures, clips, guestbook, www board, free page, search engine, links and more. ... www.britney-spears.to/ - 9k - Cached - Similar pages

The Mystery of Britney's Breasts

www.liquidgeneration.com/poptoons/britneys breasts.asp - 2k - Cached - Similar pages

Britney Spears spelling correction

The data below shows some of the misspellings detected by our spelling correction system for the query [britney spears], and the count of how many different ... www.google.com/jobs/britney.html - 40k - Cached - Similar pages

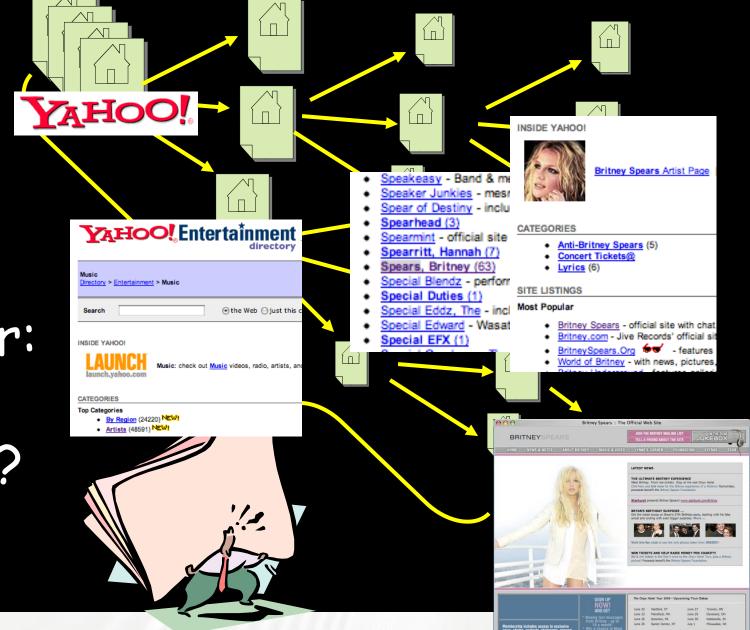
Britney Spears pictures news music Britney Spears lyrics

Britney Spears pictures mp3 sites gallery photos images music fun games chat lyrics. ... Britney Spears Forum Come see what is inside the Britney Spears forum! ... www.britney-spears.com/ - 42k - Jun 14, 2004 - Cached - Similar pages

Britney Spears Zone - Your Guide to Britney Pictures and News

Britney Spears, Britney Spears, Britney Spears, ... Britney Spears, ... www.britneyzone.com/ - 101k - Jun 14, 2004 - Cached - Similar pages

Soru: Arama motoru, sonuçlarda listelenen tüm sayfalarda, bu kelimenin olduğunu nasıl biliyor?
Cevap: Tüm bu sayfalar daha önceden indirilip, indeksleniyor.



TELLA FREND | 2001 THE LIST | TH THE ZONE ANGEROX | PREVIOUS POLICY

© 2004 BRITNEY SPEARS, ALL RIGHTS RESERVED

Crawler:
Nasıl
çalışır?

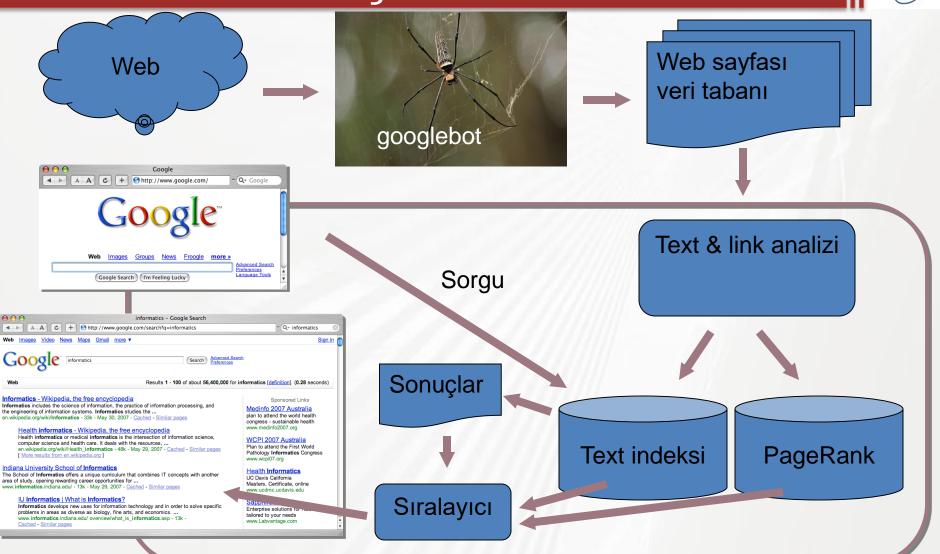
Crawler neden kullanılır?



- Arama motorlarını desteklemek (Google, Bing, Yahoo, v.b.)
- □ Özelleştirilmiş arama motorları, örnekler: news, shopping, papers, recipes, v.b.
- ☐İş mantığı: Potansiyel rakiplerin ya da iş ortaklarını takip edebilmek
- ☐ İlgilenilen Web sayfalarını izleyebilmek
- □ Kötü amaçlı kullanımları: e-posta adreslerinin "spamming", "phishing" gibi kötü amaçlar için bulunması

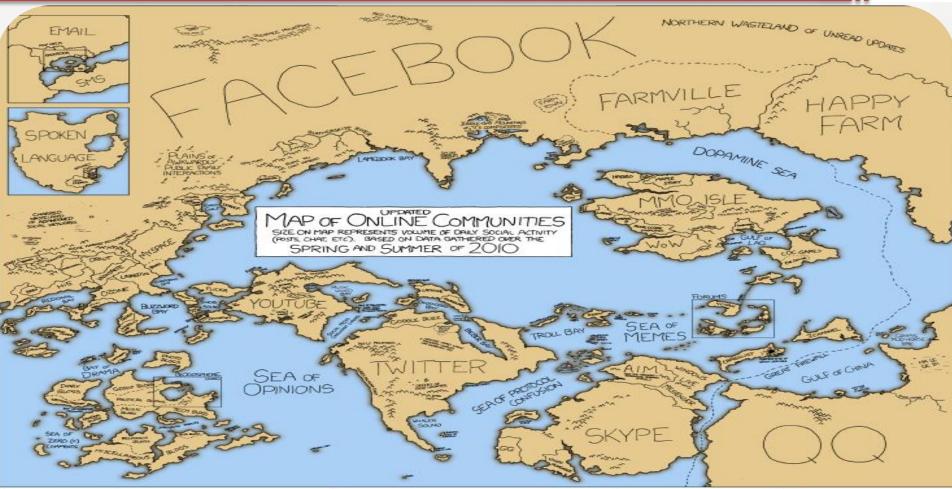
Arama motoru içindeki Crawler





Online Communities







ABOUT THIS MAP

COMMUNITIES RISE AND TALL, AND TOTAL MEMBERSHIP NUMBERS ARE NOLDINGER A GOOD MEASURE OF A COMMUNITY'S CURRENT TOTAL SOCIAL ACTIVITY IN A COMMUNITY — THAT IS, HOW MUCH TRUNKS, PLAYING, SHARING, OR OTHER SOCIALIZING HAPPENS THERE. THIS MEMORY OF APPLIES AND ORNAMES, BUT I DOTHY BEST AND TREPT DE CONSTRAINT.

ESTHAMES ARE BYSED ON THE BEST NUMBERS I COUD FIND SUF-MANUFED A GREAT DEAL OF GUESSAGRIC SPRINTON. INFORMACE, RANDOM SPRINTING, NORMANDOM SPRINTING, A 20,000-CELL SPREADSHEET, SPRINTING, CADUMS, TEA-USIF REDDING, GAST SACRICES, AND GUT INSTITUT. (I.E. PAURING THRISS UP).

SOURCES OF DATA INCLUDE GOOGLE AND BING, WINDEDA, ALEIA, 681-BONDS, CONT. STURBLE UPON WISDDRESS, AMMET BURN' WEBSTLE STRETCHE PROSE I LOUD FIND, PRESS RELIEFS, INVIDIAL SIZE EMPLOYEES. THANKS IN PRINCIPLAR TO PLAKS AT JEST-FM. LIVEDOURNAM, REDOUT, AND THE NEW YORK THINES, AS WELL AS SYSPOMINS AT A NUMBER OF SITTES INFO STRETCHES, CONTINUED OF ANDINAMENT.



Web Kullanım Madenciliği



■ Web Kullanım Madenciliği nedir?

- Web Kullanım Madenciliğinde kullanılan veriler, web üzerindeki çeşitli sunucularda tutulan kullanıcı erişim hareketlerinin yer aldığı çeşitli log dosyalarından elde edilir.
 - İstemcilerden gelen her istek, bir kayıt olarak, metin tabanlı log dosyalarına eklenir.
 - Log dosyasındaki kayıt formatı verilen servis çeşidine ve kullanılan işletim sistemine göre faklılıklar gösterir.
 - Bu log dosyalarından bazıları, access log (erişim), mail log, error log, ftp log şeklindedir.

Web Madenciliği



■ Sonuç olarak;

- Web madenciliği günümüzde internetin yoğun bir şekilde kullanımının artması nedeniyle üzerinde önemli ölçüde araştırma yapılan bir alan haline gelmiştir.
- Web madenciliği, kullanıcıların web sitesindeki davranışlarını çeşitli kayıtlar üzerinden inceleyerek web sitelerinin yeniden tasarım yada geliştirilmesi konusunda ipucları sunar.



Why Should I Know about All This?

Gartner

- "80% of business is conducted on unstructured information"
- "85% of all data stored is held in an unstructured format"
- "7 million Web pages are being added every day"



 "Unstructured data doubles every three months"