Search Engines

Prof. Poonam Zala,
Asst. Professor,
Computer Department
Umiya Arts and Commerce College.

Introduction

Web Search Engine is a software program that searches the Internet (bunch of websites) based on the words that you designate as search terms (query words).

Search engines look through their own databases of information in order to find what it is that you are looking for.

The engine provides a list of results that best match what the user is trying to find. Today, there are many different search engines available on the Internet, each with their own abilities and features. The first search engine ever developed is considered **Archie**, which was used to search for **FTP** files and the first text-based search engine is considered **Veronica**.

Web Search Engines are a good example for massively sized Information Retrieval Systems.

Dictionary Definitions

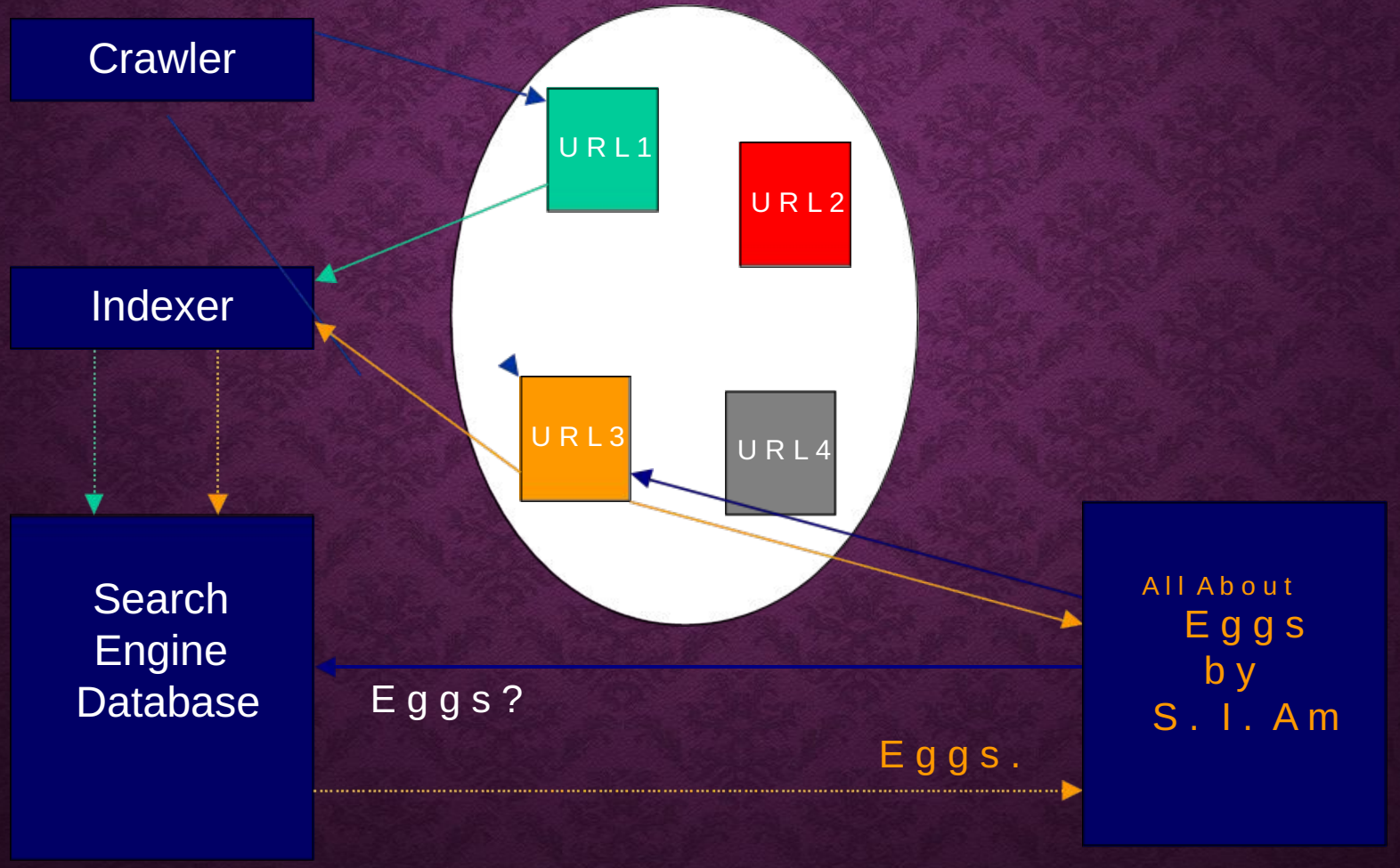
Search

COMPUTING (transitive verb) to examine a computer file, disk,database, or network for particular information

❖ HOW ACCESS A SEARCH ENGINE ?

FOR USERS, A SEARCH ENGINE IS ACCESSED THROUGH A BROWSER ONN THEIR COMPUTER, SMARTPHONE, TABLET, OR ANOTHER DEVICE. TODAY, MOST NEW BROWSERS USE AN OMNIBOX, WHICH IS A TEXT BOX AT THE TOP OF THE BROWSER. THE OMNIBOX ALLOWS USERS TO TYPE IN A URL OR A SEARCH QUERY. YOU CAN ALSO VISIT ONE OF THE MAJOR SEARCH ENGINES HOME PAGE TO PERFORM A SEARCH.

How does Search Engines Work?



how do search engines work? elaboration

- crawlers, spiders: go out to find content
  - in various ways go through the web looking for new & changed sites
  - periodic, not for each query
    - no search engine works in real time
  - some search engines do it for themselves, others not
    - buy content from companies such as Inktomi
  - for a number of reasons crawlers do not cover all of the web – just a fraction
  - what is not covered is “invisible web” ?

Elaboration …

- organizing content: labeling, arranging
  - **indexing** for searching – automatic
    - keywords and other fields
    - arranging by URL popularity - PageRank as Google
  - **classifying** as directory
    - mostly human handpicked & classified
- as a result of different organization we have basically two kinds of search engines:
  - search – input is a query that is searched & displayed
  - directory – classified content – a class is displayed
    - and fused: directories have search capabilities & vice versa

Elaboration (cont.)

- databases, caches: storing content
  - humongous files usually distributed over many computers
- query processor: searching, retrieval, display
  - takes your query as input
    - engines have differing rules on how they are handled
  - displays ranked output
    - some engines also cluster output and provide visualization
    - some engines provide categorically structured results
- at the other end is your browser

Similarities & Differences

- All search engines have these basic parts in common
- BUT the actual processes – methods how they do it – are based on various algorithms and they significantly differ
  - most are proprietary (patented) with details kept mostly secret (or protected) but based on well known principles from information retrieval or classification
  - to some extent Google is an exception – they published their method

Google Search

- In the beginning it ran on Stanford computers
- Basic approach has been described in their famous paper “The Anatomy of a Large-Scale Hypertextual Web Search Engine”
  - well written, simple language, has their pictures
  - in acknowledgement they cite the support by NSF’s Digital Library Initiative i.e. initially, Google came out of government sponsored research
  - describe their method PageRank - based on ranking hyperlinks as in citation indexing
  - “We chose our system name, Google, because it is a common spelling of googol, or ten on hundredth power”

LIST OF TOP 10 MOST POPULAR SEARCH ENGINES IN THE WORLD

- Google
- Bing.
- Yahoo
- Baidu
- yandex,.ru.
- Duckduck go.
- Ask.com.
- Aol. Com.
- Wolframalpha
- Internet archive

SEARCH ENGINES

- **Google :**

  is the best search engine with a worldwide market share between 81.5% and 92.96%

- Bing :

  market share between 2.34% and 5.29%. This place bing as the best alternative search engine to google.

- Yahoo :

  market share is between 1.64% and 2.04%.

- Baidu :

  has a global market share between 0.92% and 9.37% but is the most popular search engine in china.

- Yandex :

  Russian's most popular search engine has a global market share between 0.47 % and 0. 83%.

- Duck duck go **:**

  market share is between 0.28% and 0. 43 %

- Ask. Com **:**

  formerly known as ask jeeves, ask. Com receives approximately 0.42 % of the search share. Ask is based on a question answer format where most question are answered by other user or are answered by other users or are in the form of polls.

- Aol**. Com :**

  according to marketshare the old time famous AOL is still the top 10 search engines with market share that is close to 0.05%.

- Wolframalpha :

  wolframalpha is different than all the other search engine. They market it as a computational knowledge engine which can give you facts and data for a number of topics.

- Internet archive :

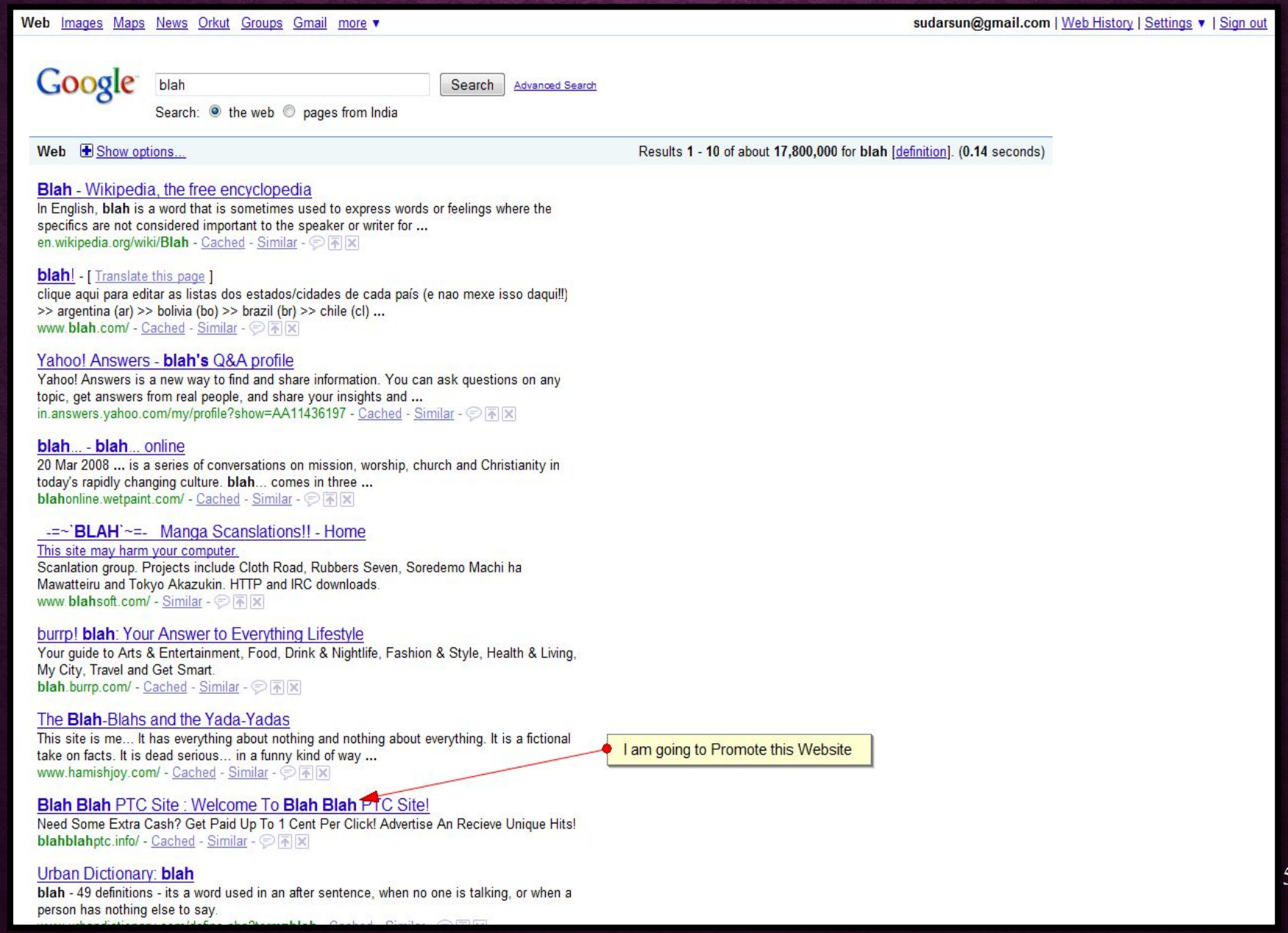
  it is a very useful tool if you want to trace the history of a domain and examine how it has changed over the years .

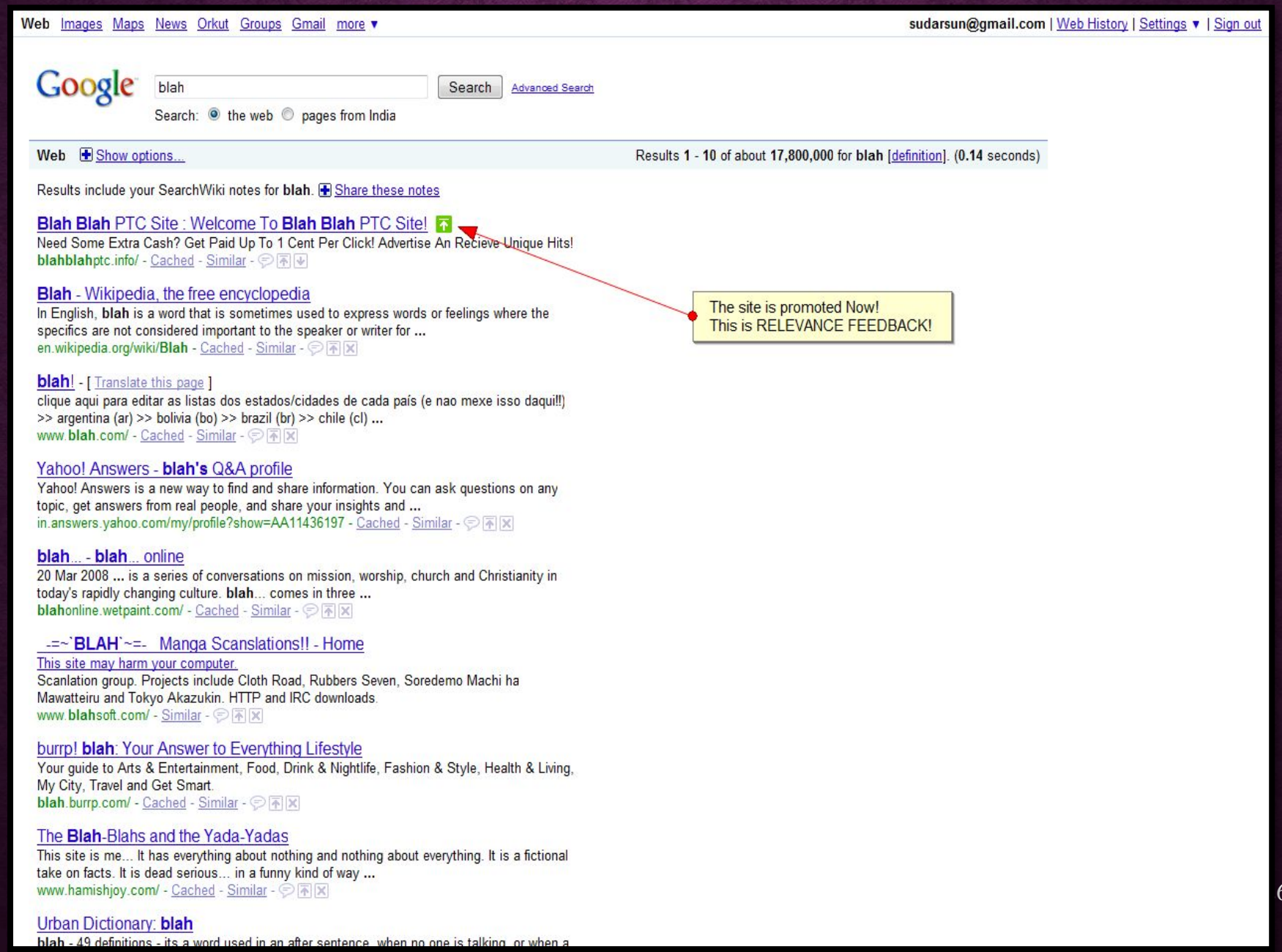
Coverage differences

- no engine covers more than a fraction of WWW
  - estimates: none more than 16%
  - hard (even impossible) to discern & compare coverage, but they differ substantially in what they cover
- in addition:
  - many national search engines
    - own coverage, orientation, governance
  - many specialized or domain search engines
    - own coverage geared to subject of interest
  - many comprehensive sources independent of search engines
    - some have compilations of evaluated web sources

Searching differences

- substantial differences among search engines on searching, retrieval display
  - need to know how they work & differ in respect to
    - defaults in searching a query
    - searching of phrases, case sensitivity, categories
    - searching of different fields, formats, types of resources
    - advance search capabilities and features
    - possibilities for refinement, using relevance feedback
    - display options
    - personalization options


Web Images Maps News Orkut Groups Gmail more ▾
sudarsun@gmail.com | Web History | Settings ▾ | Sign out
Google
blah
Search
Advanced Search
Search: the web pages from India
Web Show options...
Results 1 - 10 of about 17,800,000 for blah [definition]. (0.14 seconds)
Blah - Wikipedia, the free encyclopedia
In English, blah is a word that is sometimes used to express words or feelings where the specifics are not considered important to the speaker or writer for ...
en.wikipedia.org/wiki/Blah - Cached - Similar -
blah! - [Translate this page]
clique aqui para editar as listas dos estados/cidades de cada país (e nao mexe isso daqui!!) >> argentina (ar) >> bolivia (bo) >> brazil (br) >> chile (cl) ...
www.blah.com/ - Cached - Similar -
Yahoo! Answers - blah's Q&A profile
Yahoo! Answers is a new way to find and share information. You can ask questions on any topic, get answers from real people, and share your insights and ...
in.answers.yahoo.com/my/profile?show=AA11436197 - Cached - Similar -
blah... - blah... online
20 Mar 2008 ... is a series of conversations on mission, worship, church and Christianity in today's rapidly changing culture. blah... comes in three ...
blahonline.wetpaint.com/ - Cached - Similar -
-=~`BLAH`~=- Manga Scanslations!! - Home
This site may harm your computer.
Scanlation group. Projects include Cloth Road, Rubbers Seven, Soredemo Machi ha Mawatteiru and Tokyo Akazukin. HTTP and IRC downloads.
www.blahsoft.com/ - Similar -
burrp! blah: Your Answer to Everything Lifestyle
Your guide to Arts & Entertainment, Food, Drink & Nightlife, Fashion & Style, Health & Living, My City, Travel and Get Smart.
blah.burrp.com/ - Cached - Similar -
The Blah-Blahs and the Yada-Yadas
This site is me... It has everything about nothing and nothing about everything. It is a fictional take on facts. It is dead serious... in a funny kind of way ...
www.hamishjoy.com/ - Cached - Similar -
Blah Blah PTC Site : Welcome To Blah Blah PTC Site!
Need Some Extra Cash? Get Paid Up To 1 Cent Per Click! Advertise An Recieve Unique Hits!
blahblahptc.info/ - Cached - Similar -
Urban Dictionary: blah
blah - 49 definitions - its a word used in an after sentence, when no one is talking, or when a person has nothing else to say.
I am going to Promote this Website


Web Images Maps News Orkut Groups Gmail more ▾
sudarsun@gmail.com | Web History | Settings ▾ | Sign out
Google
blah
Search
Advanced Search
Search: the web pages from India
Web Show options...
Results 1 - 10 of about 17,800,000 for blah [definition]. (0.14 seconds)
Results include your SearchWiki notes for blah. Share these notes
Blah Blah PTC Site : Welcome To Blah Blah PTC Site!
Need Some Extra Cash? Get Paid Up To 1 Cent Per Click! Advertise An Recieve Unique Hits!
blahblahptc.info/ - Cached - Similar -
Blah - Wikipedia, the free encyclopedia
In English, blah is a word that is sometimes used to express words or feelings where the specifics are not considered important to the speaker or writer for ...
en.wikipedia.org/wiki/Blah - Cached - Similar -
blah! - [Translate this page]
clique aqui para editar as listas dos estados/cidades de cada país (e nao mexe isso daqui!!) >> argentina (ar) >> bolivia (bo) >> brazil (br) >> chile (cl) ...
www.blah.com/ - Cached - Similar -
Yahoo! Answers - blah's Q&A profile
Yahoo! Answers is a new way to find and share information. You can ask questions on any topic, get answers from real people, and share your insights and ...
in.answers.yahoo.com/my/profile?show=AA11436197 - Cached - Similar -
blah... - blah... online
20 Mar 2008 ... is a series of conversations on mission, worship, church and Christianity in today's rapidly changing culture. blah... comes in three ...
blahonline.wetpaint.com/ - Cached - Similar -
-=~`BLAH`~=- Manga Scanslations!! - Home
This site may harm your computer.
Scanlation group. Projects include Cloth Road, Rubbers Seven, Soredemo Machi ha Mawatteiru and Tokyo Akazukin. HTTP and IRC downloads.
www.blahsoft.com/ - Similar -
burrp! blah: Your Answer to Everything Lifestyle
Your guide to Arts & Entertainment, Food, Drink & Nightlife, Fashion & Style, Health & Living, My City, Travel and Get Smart.
blah.burrp.com/ - Cached - Similar -
The Blah-Blahs and the Yada-Yadas
This site is me... It has everything about nothing and nothing about everything. It is a fictional take on facts. It is dead serious... in a funny kind of way ...
www.hamishjoy.com/ - Cached - Similar -
Urban Dictionary: blah
blah - 49 definitions - its a word used in an after sentence, when no one is talking, or when a
The site is promoted Now!
This is RELEVANCE FEEDBACK!

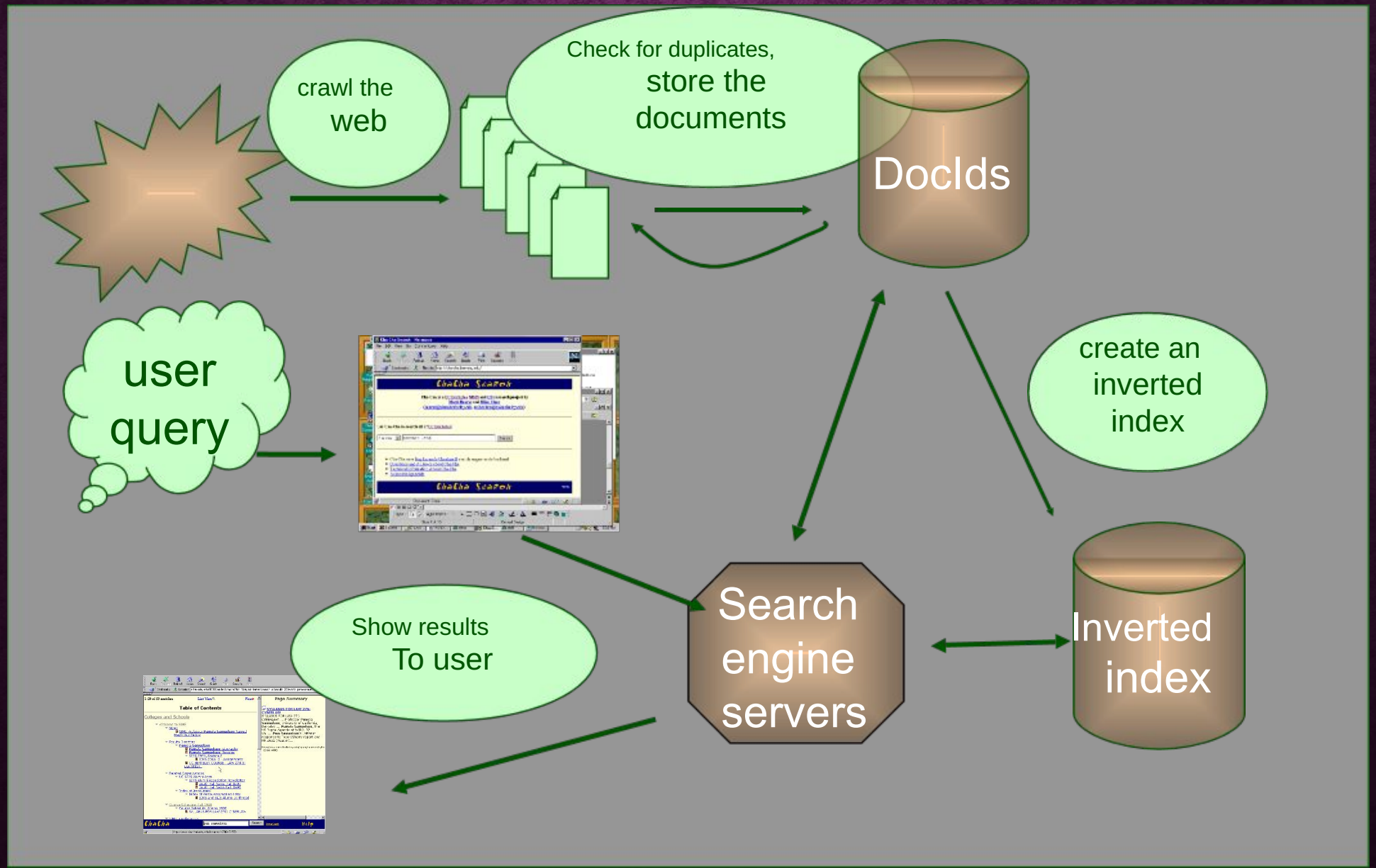
Limitations

- every search engine has limitation as to
  - Coverage: meta engines just follow coverage limitations & have more of their own search capabilities
  - finding quality information
- some have compromised search with economics
  - becoming little more than advertisers
- but search engines are also many times victims of spamdexing
  - affecting what is included and how ranked

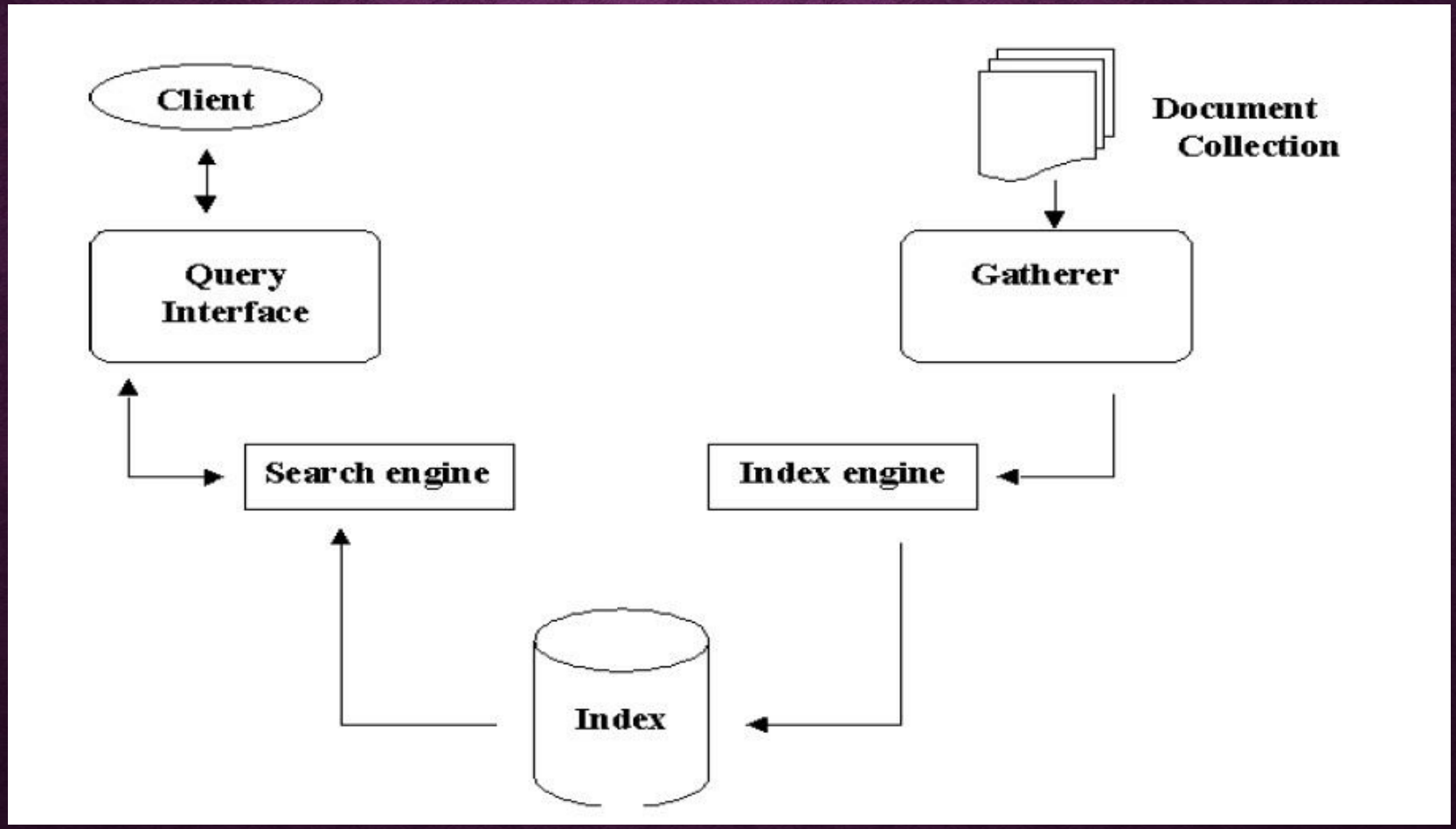
Spamming a search engine

- use of techniques that push rankings higher than they belong is also called **spamdexing**
  - methods typically include textual as well as link-based techniques
  - like e-mail spam, search engine spam is a form of adversarial information retrieval
  - the conflicting goals of accurate results of search providers & high positioning by content page rank

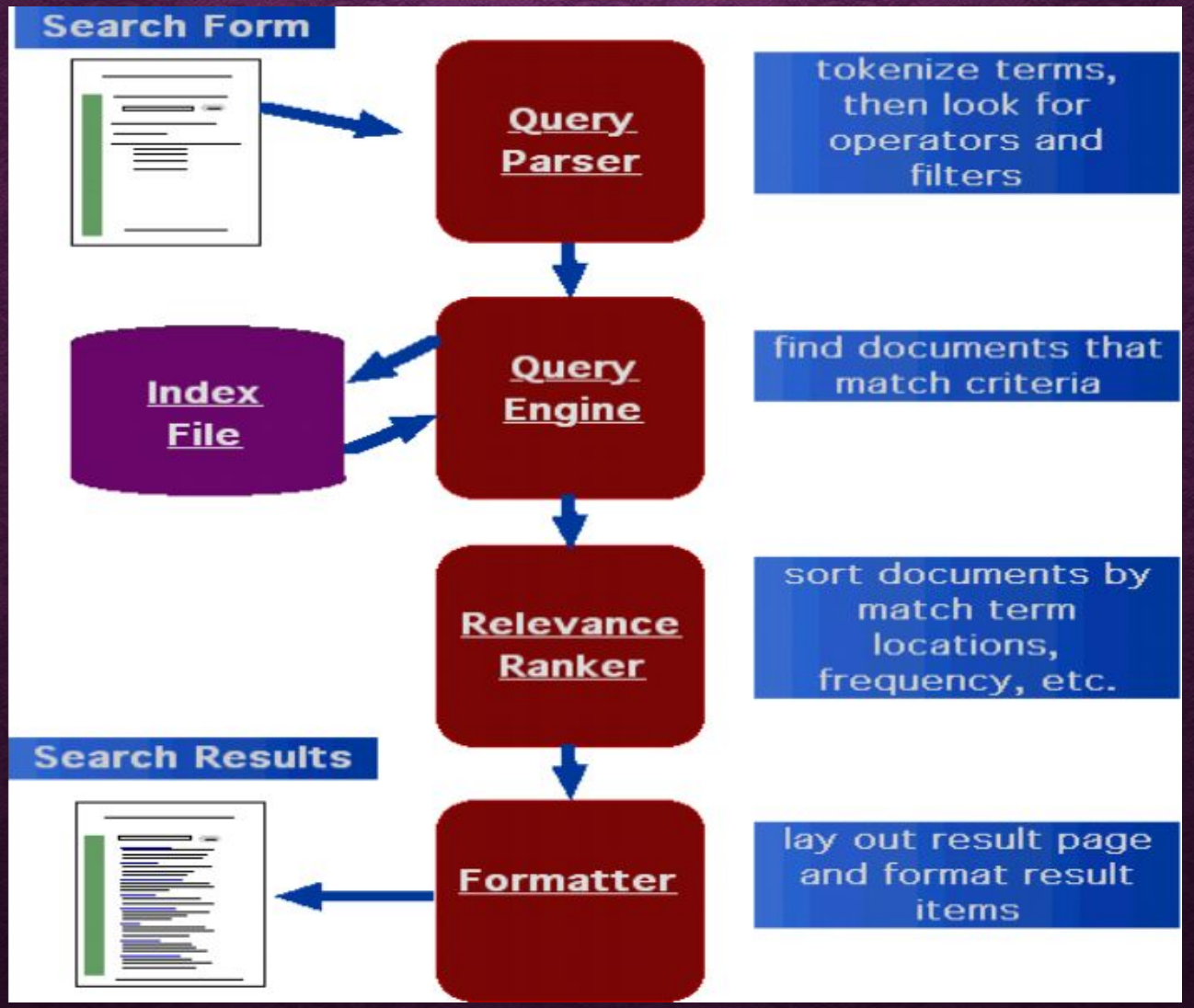
Standard Web Search Engine Architecture



Typical Search Engine



Searching – Process Flow



Advantage

- **Time Savings:**

A search engine saves you time in two ways: by eliminating the need to find information manually, and by performing searches at high speeds.

- **Relevance:**

When a search engine scans a website, it scores the content for relevance to particular search words.

- **Free Access:**

Some search engines, such as LexisNexis, specialize in legal or other specialized, scholarly information; these sites charge a fee to use their services. Google, Bing and Yahoo pay for their operations through advertising; searches are free to the user, without restrictions for the information you seek.

- **Comprehensive:**

Search engines scan the entire Web and keep comprehensive data on every page they have.

- **Advanced Search:**

In addition to keywords, search engines let you use advanced search options to refine your results.

- **Allow you to build brand credibility:**

If your page has some great rankings, there will be some credibility on behalf of the brand.

Disadvantage

- **Getting too much notice:**

One of the biggest disadvantages of search engine optimization is that you end up getting a lot more notice than you imagined you ever.

- **Too much success:**

Even though every single business wants to become a success, there is a big danger of way too much happening at a fast pace. If you can't handle too many leads, then there is a chance that customers will feel let down and opportunities will slide by.

- **Results Take Time:**

Unfortunately, SE strategies do not work their magic overnight. As it takes time for search engines to index the content of an SEO campaign, your website won't rank on certain terms for days or even weeks.

- **Difficulty of Competitive Keywords:**

Your competitors are likely trying to rank on the same keywords as you are, and if there are major corporations in your market you likely won't be able to beat out their domination over general keywords.

- **Results Are Not Guaranteed:**

With all of these variables, it can be hard for SE service-providers to absolutely guarantee that SE will position your website as the first result. But results are not accurate very time.

Thank You...!!!