Improve Web Information Quality Using Advance Searching Methodologies and Information Retrieval Techniques

Dissertation
Submitted in partial fulfillment of the requirements
for the degree of
Master of Technology

By

Vishal Bhargava

(Roll No. 18/IS/09)

Under the guidance of
Assistant Professor Rahul Katarya

Department of Information Technology

Delhi Technological University

New Delhi, 2011

CERTIFICATE

This is to certify that Mr. Vishal Bhargava (18/IS/09) has carried out the major project titled “Improve Web Information Quality Using Advance Searching Methodologies and Information Retrieval Techniques ” as a partial requirement for the award of Master of Technology degree in Information Systems by Delhi Technological University.

The major project is a bonafide piece of work carried out and completed under my supervision and guidance during the academic session 2009-2011. The matter contained in this report has not been submitted elsewhere for the award of any other degree.

(Project Guide)

Asst. Prof. Rahul Katarya
Department of Information Technology
Delhi Technological University
Bawana Road, Delhi-110042

Acknowledgements

I take this opportunity to express my sincere gratitude towards Asst. Prof. Rahul Katarya for his constant support and encouragement. His excellent guidance has been instrumental in making this project work a success.

I would like to thank Prof. O.P. Verma for his useful insights and guidance towards the project. His suggestions and advice proved very valuable throughout.

I would like to thank members of the Department of Information Technology at Delhi Technological University for their valuable suggestions and helpful discussions.

I would also like to thank my family and friends, who have been a source of encouragement and inspiration throughout the duration of the project.

I would like to thank the entire DTU family for making my stay at DTU a memorable one.

Vishal Bhargava
Roll No. 18/IS/09
M.Tech (Information Systems)
E-mail: vishalbharg@gmail.com

Abstract

This Thesis deals with the latest and advance technique for web searching i.e. image search using Natural language processing, security advice, statistical ranking with social tagging and website recommendation with security and purpose information retrieval. The result of a search will yield many thousands of results but what is important is how relevant those results are to what was intended. Natural language processing is partly used in text search but not in image search; this report proposes a Search Engine superior to any other, which uses natural language processing for images as well. This search engine is based on machine language, AIML with consideration of social tagging and also adds various features like security advice, spam checking and purpose of the intended site and also it also deal with advertising problem on web by proposing efficient and effective way of presentation of advertising.

Keywords: *NLP, search engine, recommendation engine, AIML, website advisor, social tagging, Web advertising.*

List of Figures

Contents

Chapter 1
Introduction

Information is base for all type of businesses. Internet is the fastest and inexpensive source of information today. Thousands of the people today are using internet for accessing E-Books, Notes, News, Stocks, Entertainment and Education. This information easily available by use of search engines. Search Engines, now a day is getting popularity because of its unlimited benefits. A person, sitting anywhere around the globe can find any E-Resources from his home. Getting proper information is a difficult task and it is also a big challenge to improve web information quality, because we totally rely on search engine’s result & their ranking algorithms.

The most exciting part about the Internet and its most visible component, the WWW, is that there are 100’s of millions of pages available, presenting information on a numerous topics. The bad news about the Internet is that there are 100’s of millions of pages available, most of them titled according to the whim of their author, almost all of them sitting on servers with cryptic names. When you need information about a particular subject, how do you know which pages to read? If you're like most people, you visit an Internet search engine**.**

1.1 Web Information Quality

In the fast growing science and superfast growing IT, many efforts have already been done to make the web accessible. All the work, from submission of bill to product purchase, is possible on internet (E-Commerce) but the major issue, assesses the Quality of Information at a Web.

Internet contains lots of information and search engines aims to provide efficient path to this useful information. Information providers by the search engine have different levels of knowledge, different views of the world and different intensions and mostly impersonalized [33]. Thus, provided information may be wrong, biased, inconsistent or outdated and interests of a user are specific, the most relevant link might not be among the top 10 shown by conventional

[34] search engines. Before information from the Web is used to accomplish a specific task, its quality should be assessed according to task-specific criteria.

1.2 Search Engine

The purpose of a web search engine is to search for the information the user needs. The search results are presented as a list of hits and can be web pages, images, information in various file formats like pdf, doc etc. The search engines of internet are nothing but collection of special sites on the Web those are designed to help people fetch information stored on other servers. Although there are differences in the way various search engines work, but basically they all perform three tasks:

- They search on the Internet based on important strings.
- They keep an index of the words/strings they find, and location.
- They allow users to look for words or various permutation & combinations of words found in that particular index.

Yesteryear search engines held an index of a few 100 thousand pages and documents, and received maybe 1-2 thousand inquiries each day. However, today a top search engine will index 100’s of millions of pages, and respond to 10’s of millions of queries each single day.

When people talk about Internet search engines, most of them really mean World Wide Web search engines. There were already search engines in place to help people find information on the Net, before the Web became the most visible part of the Internet. Programs like ‘gopher’ [35] and ‘Archie’[36] kept indexes of files stored on servers connected to the Internet, and considerably reduced the amount of time required to find programs and documents. In the late 1980s, getting serious value from the Internet meant knowing how to use gopher, Archie, Veronica [37]and the rest.

1.2.1 How Search Engine works [41]:

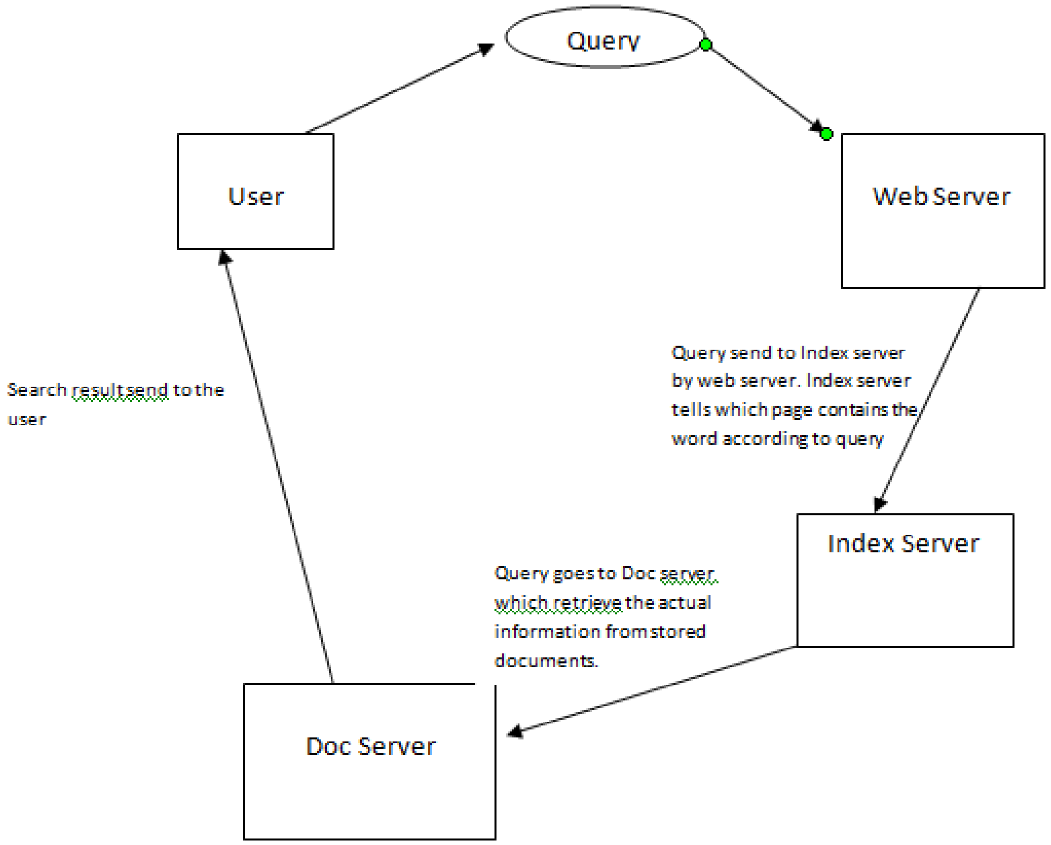

Query
User
Web Server
Query send to Index server by web server. Index server tells which page contains the word according to query
Search result send to the user
Index Server
Query goes to Doc server, which retrieve the actual information from stored documents.
Doc Server

Fig 1.1: Search Engine Working

Before a search engine can tell you where a file or document resides, it must be found before it is used. To find information on the 100's of millions of Web pages that exist, a search engine employs special software robots [43], called spiders, to build lists of the words found on Web sites. Web crawling [42] is the process of a spider building its lists. (There are few disadvantages to calling part of the Internet the World Wide Web -- a large set of arachnid-centric names for tools is one of them.) A search engine's spiders have to look at a lot of pages to gather information in order to construct and maintain a useful list of words.

How any spider does starts its travel over the Web? The usual starting points are lists of heavily hit servers and very frequently used pages. The spider will normally begin with a most known site, indexing the words on its pages and following each and every link found within the site. The spidering mechanism in this way swiftly begins to travel, spreading out across the most widely used portions of the Web.

In January 1996 Google started as a research project by Larry Page and Sergey Brin, at that time they were both PhD students at Stanford University in California. In their paper [38] they described how the system was built, Sergey Brin and Lawrence Page gave an example of how quickly their search engine's spiders can work. They built their initial system to use multiple spiders, usually three at one time. Each spider could keep about 300 connections to Web pages open at a particular instance of time. At its peak performance, using four spiders, their system could crawl over 100 PPS(pages per second), generating around 600 KB of data each second.

Maintaining everything running fast meant developing a system to feed necessary information to the spiders [44]. The initial phase of Google system had a server for providing URLs to the spiders. Rather than depending on an Internet service provider for the domain name server (DNS), that translates a server's name into an address, in order to keep delay minimum, Google had its own DNS.

When the Google spider looked at an HTML page, it looks for things:

- The words within the page
- Where the words were found

Strings/Words occurring in the title, subtitles, meta tags and other positions of relevant importance were noted for special consideration during a subsequent search by the user. The Google spider was built to index every significant word on a page, leaving out the grammatical articles "a," "an" and "the." Other spiders take other approaches.

These different approaches usually attempt to make the spider operate faster, allow users to search more efficiently, or both. For example, some spiders will keep track of the words in the title, sub-headings and links, along with the 100 most frequently used words on the page and each word in the first 20 lines of text. Lycos is said to use this approach to spidering the Web.

Other systems, such as AltaVista [46], go in the other direction, indexing every single word on a page, including "a," "an," "the" and other "insignificant" words.

Meta Tags

Meta tags allow the owner of a page to assign key words and concepts under which the page will be indexed. This can be helpful, especially in cases in which the words on the page might have double or triple meanings -- the meta tags can guide the search engine in selecting which of the several possible meanings for these words is correct. There is, however, a risk in over-reliance on meta tags, because a careless or unscrupulous page owner might add meta tags that fit very popular topics but have nothing to do with the actual contents of the page. To protect against this, spiders will correlate meta tags with page content, rejecting the meta tags that don't match the words on the page [45].

The above mentioned theory assumes that the owner of a particular page actually wants it to be included in the results of a search engine's activities. Many times, owner of the page doesn't want it showing up on a major search engine, or doesn't want the activity of a spider accessing the page. For example consider, a game that builds new, active pages each time sections of the page are displayed or new links are followed. If a Web spider accesses one of these new pages, and begins following all of the links for new pages, the game could mistake the activity for a high-speed human player and move out of control. To avoid such situations, the robot exclusion protocol was developed. This protocol, implemented in the meta-tag section at the beginning of a Web page, tells a spider to leave the page alone, neither index the words on the page nor try to follow its links.

1.2.2 Building the index

Once the spiders have accomplished the task of searching information on Web pages (and we should note that this is a task that is never actually accomplished -- the constantly changing nature of the Web means that the spiders are always crawling), the search engine must store the information in a way that makes it fruitful. There are two key components involved in making the gathered data accessible to users:

- The information stored with the data
- The method by which the information is indexed

In the simplest case, a search engine could just store the word and the URL where it was found. In reality, since there would be no way of telling whether the word was used in an important or a trivial way on the page, this would make for an engine of limited use, whether the word was used once or many times or whether the page contained links to other pages containing the word. If say in another word, there would be no way of building the ranking list that tries to present the most useful pages at the top of the list of search results.

To make for more useful results, most search engines store more than just the word and URL. An engine might store the number of times that the word appears on a page. The engine might assign a weight to each entry, with increasing values assigned to words as they appear near the top of the document, in sub-headings, in links, in the meta tags or in the title of the page. Each commercial search engine has a different formula for assigning weight to the words in its index. This is one of the reasons that a search for the same word on different search engines will produce different lists, with the pages presented in different orders.

Regardless of the precise combination of additional pieces of information stored by a search engine, the data will be encoded to save storage space. For example, the original Google paper describes using 2 bytes, of 8 bits each, to store information on weighting -- whether the word was capitalized, its font size, position, and other information to help in ranking the hit [40]. Each factor might take up 2 or 3 bits within the 2-byte grouping (8 bits = 1 byte). As a result, a great deal of information can be stored in a very compact form. After the information is compacted, it's ready for indexing.

An index has a sole purpose: It allows information to be found as quickly as possible in relevant way for which user is searching for. There are quite a few ways for an index to be built, but one of the most effective ways is to build a hash table. In hashing, a formula is applied to attach a numerical value to each word. The formula is designed to evenly distribute the entries across a predetermined number of divisions. This numerical distribution is different from the distribution of words across the alphabet, and that is the key to a hash table's effectiveness.

In English, there are some letters that begin many words, while others begin fewer. You'll find, for example, that the "M" section of the dictionary is much thicker than the "X" section. This inequity means that finding a word beginning with a very "popular" letter could take much longer than finding a word that begins with a less popular one. Hashing evens out the difference, and reduces the average time it takes to find an entry. It also separates the index from the actual entry. The hash table contains the hashed number along with a pointer to the actual data, which can be sorted in whichever way allows it to be stored most efficiently. The combination of efficient indexing and effective storage makes it possible to get results quickly, even when the user creates a complicated search.

1.2.3 Building search

Searching through an index involves a user building a query [39] and submitting it through the search engine. The query can be quite simple, a single word at minimum. Building a more complex query requires the use of Boolean operators that allow you to refine and extend the terms of the search.

1.3 Spam & Spam Filter

What is spam?

A spam is junk electronic mail or junk newsgroup postings. Some people give generic definition of spam as any unsolicited Email. However, if a remotely related person gets your e-mail address and sends you a message, this could hardly be called spam, even though it's unsolicited. True spam is generally e-mail advertising for some product sent to a mailing list or newsgroup or lots of advertise posted on website.

In surplus of wasting people's time with unwanted e-mail, spam also eats up a lot of network bandwidth. Consequently, there are numerous organizations, as well as professionals, who have taken it upon themselves to fight spam with a variety of techniques. But because the Internet is public, there is really little that can be done to prevent spam, just as it is impossible to prevent junk mail. However, some online services have formed policies to prevent spammers from spamming their subscribers.

Spam filter

A spam filter is a set of instruction (program or code) that is used to detect unwanted and unsolicited email and prevent those messages from sneaking to a user's inbox. Like other types of filtering programs, a spam filter also looks for certain criteria on which it bases judgments. For example, the simplest and earliest versions (like the one available with Microsoft's Hotmail) can be set to watch for particular words in the subject line of messages and to restrict these from the user's inbox. This method is not especially effective, too often omitting perfectly legitimate messages (these are called false positives) and letting actual spam crawl through. More sophisticated algorithms or programs, such as Bayesian filters or other heuristic filters [10], attempt to identify spam through suspicious word patterns or word frequency.

1.4 Natural Language Processing

Natural language processing (NLP) is a field of computer science and linguistics concerned with the interactions between computers and human (natural) languages [47]. In theory, natural language processing is a very attractive method of human–computer interaction. Natural language understanding is sometimes referred to as an AI-complete problem because it seems to require extensive knowledge about the outside world and the ability to manipulate it. NLP has significant overlap with the field of computational linguistics, and is often considered a sub-field of artificial intelligence.

1.5 Social Networking Service

A social networking service is an online service, platform, or site that focuses on building and reflecting of social networks [49] or social relations among people, e.g., who share interests and/or activities. A social network service essentially consists of a representation of each user (often a profile), his/her social links, and a variety of additional services. Most social network services are web based and provide means for users to interact over the internet, such as e-mail and instant messaging. Online community services are sometimes considered as a social network service, though in a broader sense, social network service usually means an individual-centered service whereas online community services are group-centered. Social networking sites allow users to share ideas, activities, events, and interests within their individual networks.

1.5.1 Social Tagging

Social tagging or Social bookmarking is a method for Internet users to organize, store, manage and search for bookmarks of resources available online to its users. Unlike file sharing, the resources themselves aren't shared, merely bookmarks [50] that reference them about their location.

Descriptions may be added to these bookmarks in the form of metadata, so users may understand the content of the resource without first needing to download it for themselves. Such descriptions may be free text comments, votes in favor of or against its quality, or tags that collectively or collaboratively become a folksonomy. Folksonomy is also called social tagging, "the process by which many users add metadata in the form of keywords to shared content".

With the advancement in Internet and web technologies the world we live in has become a small place but on the other hand with increasing users the web database is becoming huge. It is hence an extremely difficult task to extract information that is useful and secure. To facilitate the search of useful data the web search engines are based on complex algorithmically defined programs and of course human input.

It is therefore very important that these search algorithms are so defined that not only do they produce timely results but most important the results produced should be accurate. The need of the hour is a search engine [1, 2] which not only which just does not find specific words on the internet but actually understands the need of the user. That is the Requirement today is not spotting certain words and matching them with pattern matching algorithms but to understand the connection between those words- Natural Language Search. Moreover several algorithms come, artificial intelligence doing well but yet it is not equivalent to the human perspective so page ranking should be according to the users, not only depend on machine. Facebook, Twitter, Orkut like social networking websites provide best way to give rank to the page depends on impact of page on humans.

Natural language processing is now being used for text search (in websites like ask.com) but the a major problem lies with image search, our search engine provides a method that utilizes AIML i.e. To say a Q&A service for image search. AIML (Artificial Intelligence Markup Language) is an XML complaint markup language. It is an easy to learn language and has been known to be used for Intelligent tutoring Systems (ITS) [3].

One more problem persist, spam crawled by search engine, than the site which contain spam data may come into top 10 result, so need of spam free crawling there. Our system also deals with that problem so spam full site does not indexed by our search engine.

Not only does our search engine provide the functionality of a natural language search for images, spam free crawling but it also adds a security feature to the web results displayed. It provides the users with information regarding how secure a website is. And it also deals with problem of advertising on web by providing a better way to present advertise of user interest without information degradation.

The organization of this report is as follows: Chapter II presents Related Work and Research Objectives. Evolutionary Algorithms and Techniques comes in Chapter III Description of the system and its implementation are dealt with in Chapter IV, Results in Chapter V and finally Conclusions and Future work are presented in Chapter VI.

Chapter 2

Related Work and Research Objectives

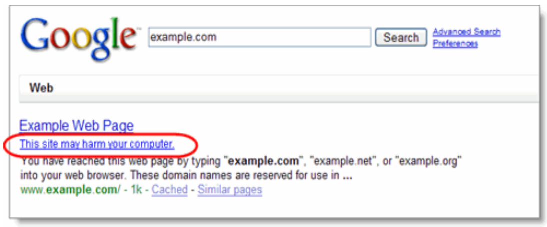
A search engine generally refers to a system to search the web for any required information. Natural language based search engine is a new type of search engine wherein the objective is to retrieve the information in the form a short precise answer to the question asked[6]; a natural language based search engine like ask jeeves[7] is typically capable of answering any type of factual queries such as “What is the second highest mountain in the world?” This type of system is generally used to answer the queries in the form of text from all kinds of textual documents that exist and not videos, images, audio etc.

Biggest web application of this century is social networking [48], Facebook, Orkut like websites just change the life of people, these not impact only their lives[8] also effect on other web service like online gaming, e-commerce, tagging. By use of social tagging people can map the authenticity and popularity of any service, product or most imported information. Delicious[9] type of web service have great impact on web service but audience poll also not effected like statistical method of searching, so need of combination of all these type of service in proper ratio.

Spam filter, we seen in email & email related application[10], but it never seen in search engine. Naivian bayss spam filter used in email client and it work according to spam word which is mostly used in email, like British lottery system, bank account number. But it rarely seen on time of crawling of web-page that search engine check spam in page content.

All the above search tools add complexity to the information retrieval activities but none of them talk about security of the search results.

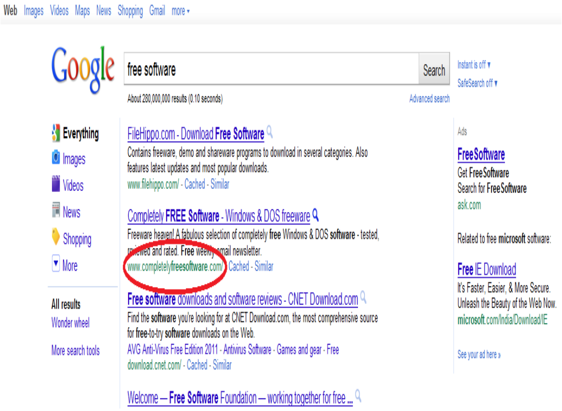
Some search engine like Google also show warning [11] but only for highly suspicious results and websites showing strange behavior, but not for every site.



Example Web Page
This site may harm your computer.
You have reached this web page by typing "**example.com**", "example.net", or "example.org" into your web browser. These domain names are reserved for use in ...
www.**example.com**/ - 1k - Cached - Similar pages

Fig 2.1: Google advice

The results shown when we search for “free software” on google are as shown below (fig 2.2).



About 280,000,000 results (0.10 seconds)

FileHippo.com - Download **Free Software**
Contains freeware, demo and shareware programs to download in several categories. Also features latest updates and most popular downloads.
www.filehippo.com/ - Cached - Similar

Completely **FREE Software** - Windows & DOS freeware
Freeware heaven! A fabulous selection of completely **free** Windows & DOS **software** - tested, reviewed and rated. **Free** weekly email newsletter.
www.completely**freesoftware**.com/ - Cached - Similar

Free software downloads and software reviews - CNET Download.com
Find the **software** you're looking for at CNET Download.com, the most comprehensive source for **free**-to-try **software** downloads on the Web.
AVG Anti-Virus Free Edition 2011 - Antivirus Software - Games and gear - Free
download.cnet.com/ - Cached - Similar

Welcome — **Free Software** Foundation — working together for free ...

FreeSoftware
Get **FreeSoftware**
Search for **FreeSoftware**
ask.com

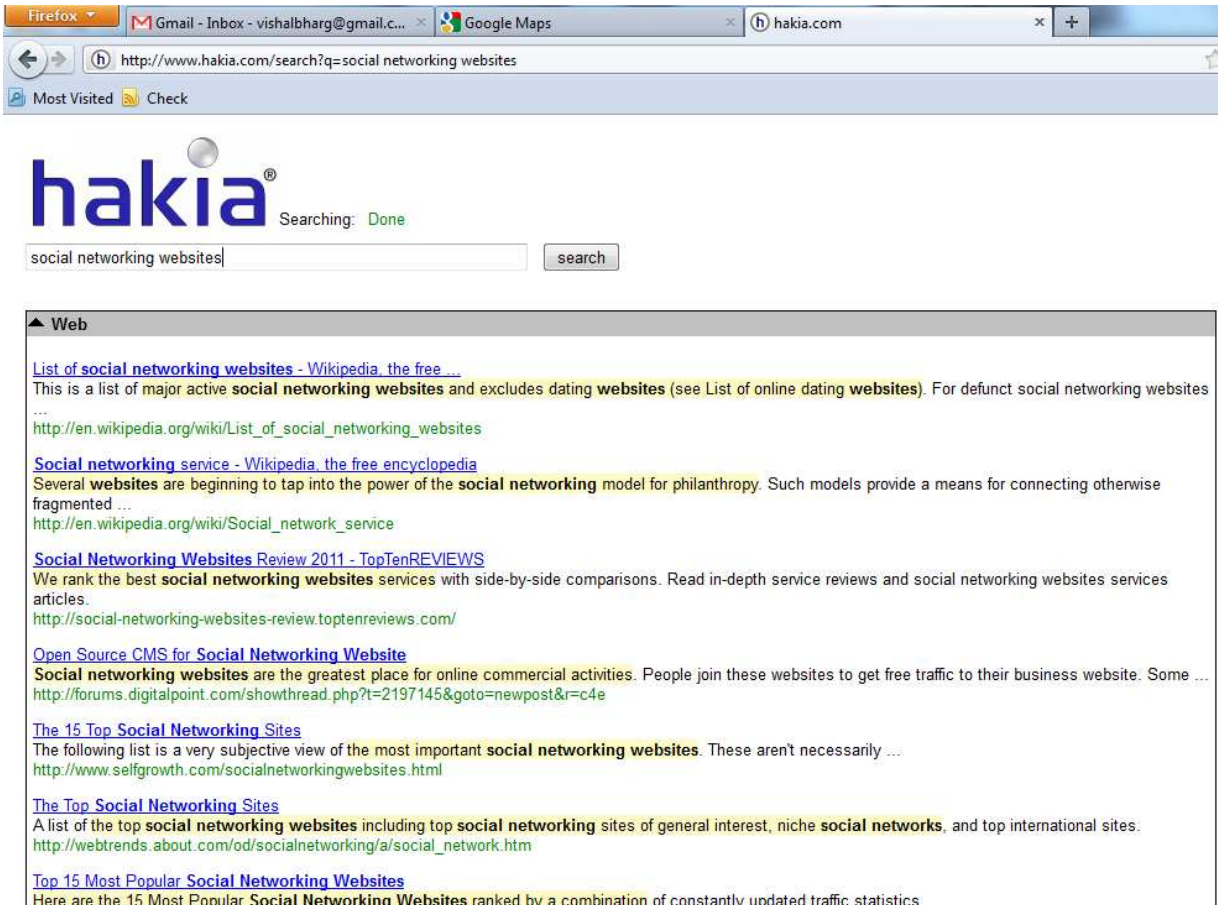
Related to free **microsoft** software:

Free IE Download
It's Faster, Easier, & More Secure.
Unleash the Beauty of the Web Now.
microsoft.com/India/Download/IE

Fig 2.2: Google search for “free software”

According to anti-virus companies many sites that are displayed as search results are unsecure. The difference lies in their risk factors. Some content can be classified as average risk, some high risk and still others can be very dangerous. Already existing search engines are not capable of understanding the difference between these terms and thus cannot provide the user with security information about each website which the user deserves. To use such features usually add-ons are provided which have to be separately installed, this is not only a tedious task but also makes the search engine platform dependent.

Semantic technology also comes into field of search engines[14], several engines are coming into pictures using web 3.0(fig 2.3) . Ontology based products provides service in several information system, knowledge management gadgets[15, 51], desktop tools.



Firefox
Gmail - Inbox - vishalbharg@gmail.c...
Google Maps
hakia.com
http://www.hakia.com/search?q=social networking websites
Most Visited
Check
hakia
Searching: Done
social networking websites
search
Web

List of **social networking websites** - Wikipedia, the free ...
This is a list of major active **social networking websites** and excludes dating **websites** (see List of online dating **websites**). For defunct social networking websites ...
http://en.wikipedia.org/wiki/List_of_social_networking_websites

Social networking service - Wikipedia, the free encyclopedia
Several **websites** are beginning to tap into the power of the **social networking** model for philanthropy. Such models provide a means for connecting otherwise fragmented ...
http://en.wikipedia.org/wiki/Social_network_service

Social Networking Websites Review 2011 - TopTenREVIEWS
We rank the best **social networking websites** services with side-by-side comparisons. Read in-depth service reviews and social networking websites services articles.
http://social-networking-websites-review.toptenreviews.com/

Open Source CMS for **Social Networking Website**
Social networking websites are the greatest place for online commercial activities. People join these websites to get free traffic to their business website. Some ...
http://forums.digitalpoint.com/showthread.php?t=2197145&goto=newpost&r=c4e

The 15 Top **Social Networking** Sites
The following list is a very subjective view of the most important **social networking websites**. These aren't necessarily ...
http://www.selfgrowth.com/socialnetworkingwebsites.html

The Top **Social Networking** Sites
A list of the top **social networking websites** including top **social networking** sites of general interest, niche **social networks**, and top international sites.
http://webtrends.about.com/od/socialnetworking/a/social_network.htm

Top 15 Most Popular **Social Networking Websites**
Here are the 15 Most Popular **Social Networking Websites** ranked by a combination of constantly updated traffic statistics

Fig 2.3 : Hakia a semantic search engine

Advertisement on web always tricky things, website wants to earn money and only ads is the medium and user want the information, so make balance between both is so difficult. Advertise should be according to users so lots of work done for it, it can be interactive [12], or it can be come from local proxy[13].

Moreover the existing search engines do not provide the user any advice regarding what type of website is the user attempting to open.

Inspired from all the above mentioned difficulties we present an easy to use and a fast accessible search engine which can provide a richer faster and secure way to search from the current web scenario.

Chapter 3
Evolutionary Algorithms and Techniques

3.1 Naïve Bayss Theorem

Naive Bayesian Classifier: A statistical classifier is called Naive Bayesian classifier[16, 17]. This classifier is based on the Bayes' Theorem and the maximum posteriori hypothesis.

Bayesian is used to predict class membership probabilities, such as the probability that a given sample belongs to a particular class. Bayesian classifier is based on Bayes' theorem. Naive Bayesian classifiers assume that the effect of an attribute value on a given class is independent of the values of the other attributes. This assumption is called class conditional independence.It is made to simplify the computation involved and, in this sense, is considered ”naive”.

Bayes' Theorem [18]

Let X = {x1, x2, . . . , xn} be a sample, whose components represent values made on a set of n attributes. In Bayesian terms, X is considered ”evidence”. Let H be some hypothesis, such as that the data X belongs to a specific class C. For classification problems, our goal is to determine P(H|X), the probability that the hypothesis H holds given the ”evidence”, (i.e. the observed data sample X). In other words, we are looking for the probability that sample X belongs to class C, given that we know the attribute description of X.

P(H|X) is the a posteriori probability of H conditioned on X.

According to Bayes' theorem, the probability that we want to compute P(H|X) can be expressed in terms of probabilities P(H), P(X|H), and P(X) as

$$P(H|X) = P(X|H)\,P(H) / P(X)\ ,$$

The naive Bayesian classifier works as follows:

1. Let T be a training set of samples, each with their class labels. There are k classes, C1,C2, . . . ,Ck. Each sample is represented by an n-dimensional vector, X = {x1, x2, . . . , xn}, depicting n measured values of the n attributes, A1,A2, . . . ,An, respectively.

2. Given a sample X, the classifier will predict that X belongs to the class having the highest a posteriori probability, conditioned on X. That is X is predicted to belong to the class Ci if and only if

$$P(Ci|X) > P(Cj\,|X) \text{ for } 1 \_ j \_ m, j\ 6{=}\ i. \qquad \text{Eq. (1)}$$

Thus we find the class that maximizes P(Ci|X). The class Ci for which P(Ci|X) is maximized is called the maximum posteriori hypothesis. By Bayes' theorem

$$P(Ci|X) = P(X|Ci)\ P(Ci) \,/\, P(X)\ . \qquad \text{Eq. (2)}$$

3. As P(X) is the same for all classes, only P(X|Ci)P(Ci) need be maximized. If the class a priori probabilities, P(Ci), are not known, then it is commonly assumed that the classes are equally likely, that is, P(C1) = P(C2) = . . . = P(Ck), and we would therefore maximize P(X|Ci). Otherwise we maximize P(X|Ci)P(Ci). Note that the class a priori probabilities may be estimated by P(Ci) = freq(Ci, T)/|T|.

4. Given data sets with many attributes, it would be computationally expensive to compute P(X|Ci). In order to reduce computation in evaluating P(X|Ci) P(Ci), the naive assumption of class conditional independence is made. This presumes that the values of the attributes are conditionally independent of one another, given the class label of the sample. Mathematically this means that

$$P\ (X \mid i\ ) \approx \prod\nolimits^{n}{}_{k=1}\ P\ (xk \mid Ci\ ) \qquad \text{Eq. (3)}$$

be estimated from the training set. Recall that here xk refers to the value of attribute Ak for sample X.

(a) If Ak is categorical, then P(xk|Ci) is the number of samples of class Ci in T having the value xk for attribute Ak, divided by freq(Ci, T), the number of sample of class Ci in T.

(b) If Ak is continuous-valued, then we typically assume that the values have a Gaussian distribution with a mean μ and standard deviation Ω defined by

$$g(x, \mu, \sigma) = 1 /\sqrt{2\pi\sigma} \quad \exp - (x - \mu)2/2\sigma\ 2 \qquad \text{Eq. (4)}$$

so that

$$p(xk\ |Ci\ ) = g(xk\ , \mu Ci\ , \sigma Ci\ ) \qquad \text{Eq. (5)}$$

We need to compute μCi and _Ci , which are the mean and standard deviation of values of attribute Ak for training samples of class Ci.
5. In order to predict the class label of X, P(X|Ci)P(Ci) is evaluated for each class Ci. The classifier predicts that the class label of X is Ci if and only if it is the class that maximizes P(X|Ci)P(Ci).

The Naïve Bayes theorem has the following characteristics as advantages:

Advantages:

1. It can handle discrete and quantitative data
2. Robust
3. During probability estimate calculations
4. Efficient regarding space and fast working.
5. Insensitive to irrelevant features
6. Quadratic decision boundary
7. By ignoring the instance it can handles missing values

3.2 Document object module (DOM)

The Document Object Model (DOM) [19] is an application programming interface (API) for valid HTML and well-formed XML documents. Basically it is used to define the logical structure of the documents and is used to access a document and document manipulation. In the DOM specification, the term "document" is used in the broad sense - increasingly, XML is being used as a way of representing discrete information that may be stored in diverse systems, and much of this would traditionally be seen as data rather than as documents. So XML represents this data as documents, and the DOM may be used to manage this data.

With the Document Object Model, programmers can build documents, navigate their structure, and add, modify, or delete elements and content. Anything data entity in an HTML or XML document can be accessed, changed, deleted, or added using the Document Object Model, but it included some exceptions like the DOM interfaces for the XML internal and external subsets have not yet been specified.

As a W3C specification, one important objective for the Document Object Model is to provide a standard programming interface that can be used in a wide variety of environments and applications. The DOM is designed to be used with any programming language. In order to provide a precise, language-independent specification of the DOM interfaces, we have chosen to define the specifications in Object Management Group IDL.

What the Document Object Model is

The DOM is a programming API for documents. It is based on an object structure that closely resembles the structure of the documents it models. For instance, consider this table, taken from an HTML document:

```
<TABLE>
<TBODY>
<TR>
<TD>DTU</TD>
<TD>IT</TD>
```

```
</TR>
<TR>
<TD>MTECH</TD>
<TD>IS</TD>
</TR>
</TBODY>
</TABLE>
```

A graphical representation of the DOM of the example table is:

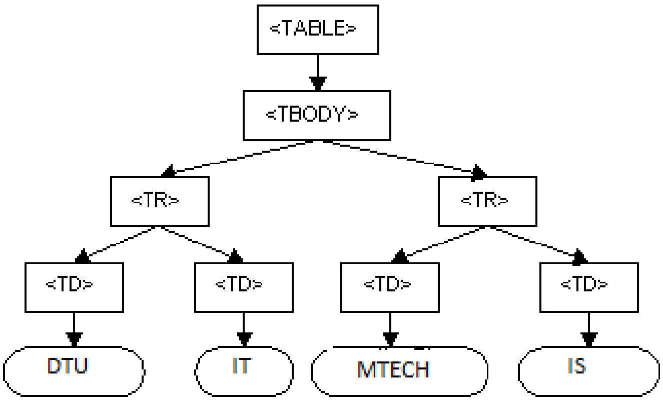


Fig 3.1: Graphical representation of the DOM of the example table

In the DOM, documents have a logical structure which contains information in the tree forms. Each document contains zero or one doctype nodes, one root element node, and zero or more comments or processing instructions; the root element serves as the root of the element tree for the document. However, the DOM does not specify that documents must be implemented as a tree or a grove, nor does it specify how the relationships among objects be implemented. The DOM is a logical model that may be implemented in any convenient manner. In this specification, we use the term structure model to describe the tree-like representation of a document. We also use the term "tree" when referring to the arrangement of those information items which can be reached by using "tree-walking" methods; (this does not include attributes). One important property of DOM structure models is structural isomorphism: if any two Document Object Model implementations are used to create a representation of the same document, they will create the same structure model, in accordance with the XML Information Set.

3.3 AIML

The XML dialect called AIML [20] was developed by Richard Wallace and a worldwide free software community between the years of 1995 and 2002. It formed the basis for what was initially a highly extended Eliza called "A.L.I.C.E." ("Artificial Linguistic Internet Computer Entity"), which won the annual Loebner Prize Contest for Most Human Computer three times, and was also the Chatterbox Challenge Champion in 2004.

Because the A.L.I.C.E. AIML set was released under the GNU GPL, and because most AIML interpreters are offered under a free or open source license, many "Alicebot clones" have been created based upon the original implementation of the program and its AIML knowledge base. Free AIML sets in several languages have been developed and made available by the user community. There are AIML interpreters available in Java, Ruby, Python, C++, C#, Pascal, and other languages. A semi-formal specification and a W3C XML Schema for AIML are available.

Elements of AIML

AIML contains several elements. The most important of these are described in further detail below.

Categories

Categories in AIML are the fundamental unit of knowledge. A category consists of at least two further elements: the pattern and template elements. Here is a simple category:

```
<category>    <pattern>WHAT IS YOUR NAME</pattern>
  <template>My name is John.</template>
 </category>
```

When this category is loaded, an AIML bot will respond to the input "What is your name" with the response "My name is John."

Patterns

A pattern is a string of characters intended to match one or more user inputs. A literal pattern like

WHAT IS YOUR NAME

will match only one input, ignoring case: "what is your name". But patterns may also contain wildcards, which match one or more words. A pattern like

WHAT IS YOUR *

will match an infinite number of inputs, including "what is your name", "what is your shoe size", "what is your purpose in life", etc.

The AIML pattern syntax is a very simple pattern language, substantially less complex than regular expressions and as such not even of level 3 in the Chomsky hierarchy. To compensate for the simple pattern matching capabilities, AIML interpreters can provide preprocessing functions to expand abbreviations, remove misspellings, etc.

Template

A template specifies the response to a matched pattern. A template may be as simple as some literal text, like

My name is John.

A template may use variables, such as the example

My name is <bot name="name"/>.

which will substitute the bot's name into the sentence, or

You told me you are <get name="user-age"/> years old.

which will substitute the user's age (if known) into the sentence.

Template elements include basic text formatting, conditional response (if-then/else), and random responses.

Templates may also redirect to other patterns, using an element called srai. This can be used to implement synonymy, as in this example (where CDATA is used to avoid the need for XML escaping):

```
<category>
  <pattern>WHAT IS YOUR NAME</pattern>
  <template><![CDATA[My name is <bot name="name"/>.]]></template>
</category>
<category>
  <pattern>WHAT ARE YOU CALLED</pattern>
  <template>
    <srai>what is your name</srai>
  </template>
</category>
```

The first category simply answers an input "what is your name" with a statement of the bot's name. The second category, however, says that the input "what are you called" should be redirected to the category that matches the input "what is your name"--in other words, it is saying that the two phrases are equivalent.

Templates can contain other types of content, which may be processed by whatever user interface the bot is talking through. So, for example, a template may use HTML tags for formatting, which can be ignored by clients that don't support HTML.

3.4 HTTP Methods Extraction

Common methods for HTTP/1.1 [21] is defined below

GET

The GET method means retrieves whatever information (in the form of an entity) is identified by the Request-URI. If the Request-URI refers to a data-producing process, it is the produced data which shall be returned as the entity in the response and not the source text of the process, unless that text happens to be the output of the process.

The semantics of the GET method change to a "conditional GET" if the request message includes an If-Modified-Since, If-Unmodified-Since, If-Match, If-None-Match, or If-Range header field. A conditional GET method requests that the entity be transferred only under the circumstances described by the conditional header field(s). The conditional GET method is intended to reduce unnecessary network usage by allowing cached entities to be refreshed without requiring multiple requests or transferring data already held by the client.

POST

The POST method is used to request that the origin server accept the entity enclosed in the request as a new subordinate of the resource identified by the Request-URI in the Request-Line. POST is designed to allow a uniform method to cover the following functions:

- Annotation of existing resources;
- Posting a message to a bulletin board, newsgroup, mailing list, or similar group of articles;
- Providing a block of data, such as the result of submitting a form, to a data-handling process;
- Extending a database through an append operation.

The actual function performed by the POST method is determined by the server and is usually dependent on the Request-URI. The posted entity is subordinate to that URI in the same way that a file is subordinate to a directory containing it, a news article is subordinate to a newsgroup to which it is posted, or a record is subordinate to a database.

The action performed by the POST method might not result in a resource that can be identified by a URI. In this case, either 200 (OK) or 204 (No Content) is the appropriate response status, depending on whether or not the response includes an entity that describes the result.

Data transfer to a webpage using these methods and extraction of web data using HTTPWebRequest and HTTPWebResponse and parsing the data with the help of return tags.

3.5 Google Gears Geo Location

The Geolocation API provides service (web application) to obtain a user's geographical position.

The Geolocation API enables a web application to:

- Obtain the user's current position, using the getCurrentPosition method
- Watch the user's position as it changes over time, using the watchPosition method
- Quickly and cheaply obtain the user's last known position, using the lastPosition property

The Geolocation API provides the best estimate of the user's position using a number of sources (called location providers). These providers may be onboard (GPS for example) or server-based (a network location provider). The getCurrentPosition and watchPosition methods support an optional parameter of type PositionOptions which lets you specify which location providers to use. Geolocation API network protocol helps the Gears uses to communicate with network location providers.

On the basis of above information, the procedure which Google uses to give information about the location can be described below.Now, Google stores the Mac address and the SSID together with the absolute location in a geo database that is kept updated when possible.

When you connect to Google Latitude using your laptop, Google can determine your accurate location using the Wi-Fi towers only because it has the information already in its database, which came from other queries of you or from other persons, about the precise location of the Wi-Fi towers. The above procedure can be shown in the following figure where the blue person sends GPS and the information about the GSM and Wi-Fi towers. Google compute the info about the

Wi-Fi towers. The location of the green person can be known by querying Google with Wi-Fi towers only.

Chapter 4

System Features

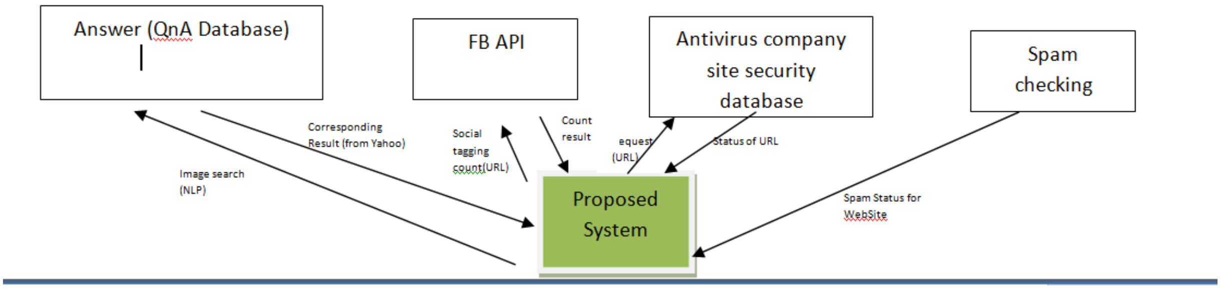


Fig 4.1: Proposed System

4.1 Spam checking on crawling time

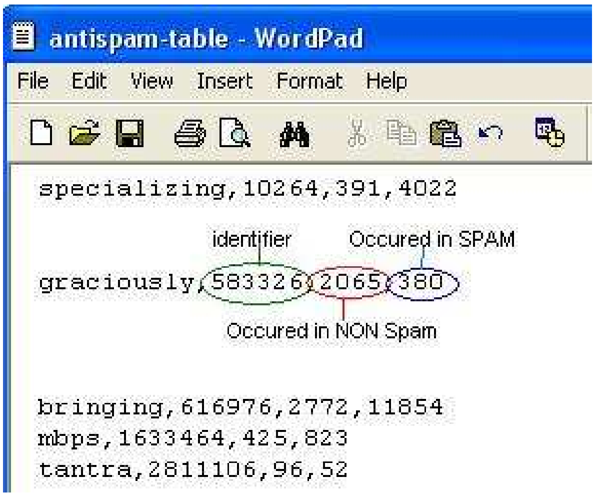
Means of spam content on webpage: Usually spam word refers for Electronic junk mail or junk newsgroup postings or an unsolicited e-mail. But in terms for a webpage, a webpage full with scraped content, lots of ads, and several links to other pages and very less useful information is called spam webpage.

Our Method:

To categorize a spam and non – spam data, first data retrieve [23,24] from webpages and following algorithm and approximations are used :

1. First, sorting will be done on the basis of language (spam or non spam), then on words and lastly on count.
2. If a word is not there, take in account to bring it as close as possible to P(word|class) using Laplacian.
3. Following table will be used for analysis

To filter a message, the antispam-table.txt [22] file is used which consist of words which differentiate that whether the message is spam or not. In addition to words, there are three numbers. The first one act as an identifier which is assigned by the anti-spam engine. The second number gives the information about the number of times a word has appeared in a non-spam message. Third number determines the number of times that a word has appeared in a spam message.



```
specializing,10264,391,4022

graciously,583326,2065,380

bringing,616976,2772,11854
mbps,1633464,425,823
tantra,2811106,96,52
```

Fig 4.2: Anti-spam table for Statistical Filtering and a new phrase list for Phrase Filtering.

Statistical Filtering is done using the anti-spam table and Phrase Filtering through a new phrase list.

CSV file has been used to generate the database. Excel computation is also included which can be used as a reference. The method as described below:

1. Eliminate any type of special characters so that only words can be used for the fair and probable computation. It is good to go with lower case letters or to convert each word in lower case**.**

   String content = TextBox1.Text;

   content = inputContent.Replace("\t\n,.;:?!&", ""); //

2. Detach each word and find the equivalent from the database. The value determines the no of times the word appeared in the spam or no-spam message. If the word has not occurred even a single time then use or just make it equal to 1.
3. Final probability is calculated by multiplying each word occurrences to the total probability:
4. Show the results in a table: Comparison of the results with each other can be used to determine the highest probability which could be the basis of deciding spam or non-spam messages.
5. After determining a message to be a spam, update the values in the database for each word. Same procedure can be applied for non-spam also. If the value is not there in the database, just stick it in for future reference. Classification could be better with more references.

4.2 Image search using NLP

The system here is a natural language search engine. It is based on Natural Language Programming and AIML. Moreover it not only uses Natural Language Processing for text[7,25] but should also returns images and video results related to the query, i.e. It applies Natural Language Processing to images.

The use of natural language processing is done using AIML.

There are approximately 20 additional tags often found in AIML files, and it's possible to create your own "custom predicates".

Example of pattern is as follows:

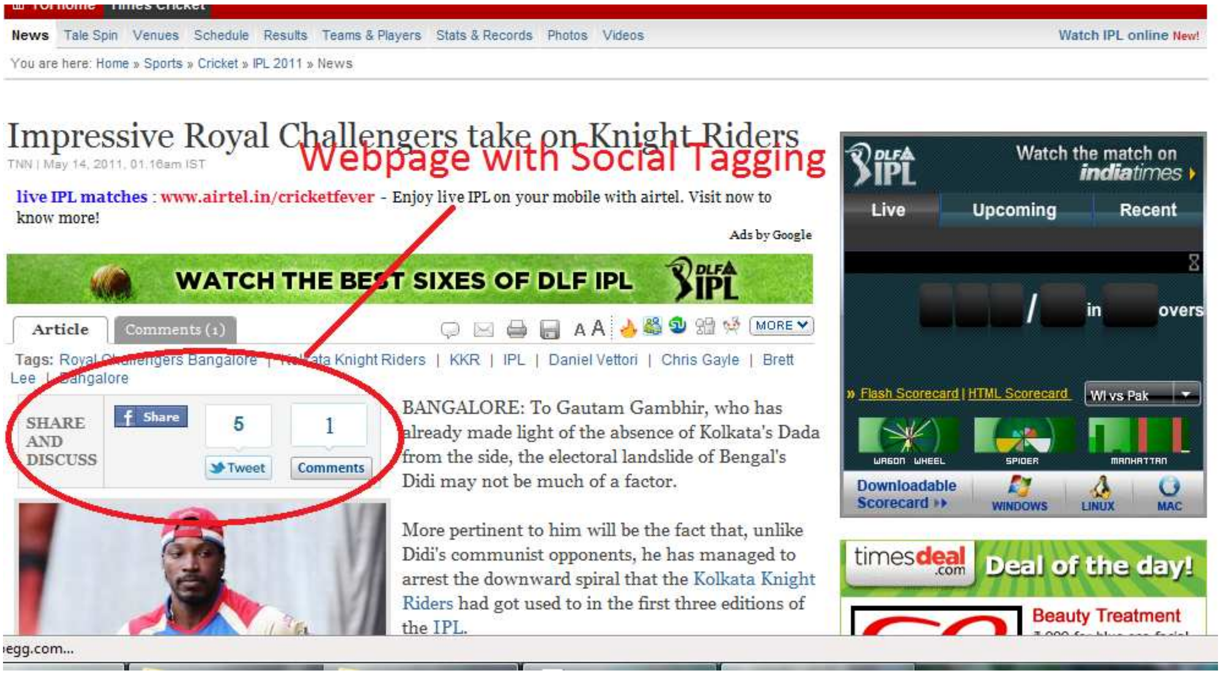
```
<category>
   <pattern> who is jai in sholay movie? <pattern>
   <template>
      <think><set name="topic">Me</set></think>
Amitabh Bachan
   </template>
</category>
```

But this is syntax dependent, difference come between queries like “Who is president of India?” And “who is the president of India?” So our intelligent system removes helping verbs and articles from the query so that the question will always remain “who president of India? “So in both the cases the answer will remain the same.

4.3 Social tagging ranking with statistical ranking

Social bookmarking is a method for Internet users to organize, store, manage and search for bookmarks of resources online. Unlike file sharing, the resources themselves aren't shared, merely bookmarks that reference them.

With regard to creating a high-quality search engine, a social bookmarking system has several advantages over traditional automated resource location and classification software, such as search engine spiders. All tag-based classification of Internet resources (such as web sites) is done by human beings, who understand the content of the resource, as opposed to software, which algorithmically attempts to determine the meaning of a resource. Also, people can find and bookmark web pages that have not yet been noticed or indexed by web spiders [26]. Additionally, a social bookmarking system can rank a resource based on how many times it has been bookmarked by users, which may be a more useful metric for end-users than systems that rank resources based on the number of external links pointing to it (although both types of ranking are vulnerable to fraud, and both need technical countermeasures to try to deal with this).


News | Tale Spin | Venues | Schedule | Results | Teams & Players | Stats & Records | Photos | Videos — Watch IPL online New!

You are here: Home » Sports » Cricket » IPL 2011 » News

Impressive Royal Challengers take on Knight Riders

TNN | May 14, 2011, 01.16am IST

Webpage with Social Tagging

live IPL matches : www.airtel.in/cricketfever - Enjoy live IPL on your mobile with airtel. Visit now to know more!

Ads by Google

WATCH THE BEST SIXES OF DLF IPL

Article | Comments (1) | A A | MORE

Tags: Royal Challengers Bangalore | Kolkata Knight Riders | KKR | IPL | Daniel Vettori | Chris Gayle | Brett Lee | Bangalore

SHARE AND DISCUSS — Share 5 — Tweet — 1 Comments

BANGALORE: To Gautam Gambhir, who has already made light of the absence of Kolkata's Dada from the side, the electoral landslide of Bengal's Didi may not be much of a factor.

More pertinent to him will be the fact that, unlike Didi's communist opponents, he has managed to arrest the downward spiral that the Kolkata Knight Riders had got used to in the first three editions of the IPL.

Fig 4.3: Social tagging

We used facebook API to get number of social tag on particular webpage and merge it with traditional statistical ranking of Google with a threshold value. Actually impact of social tagging individually not works very well.

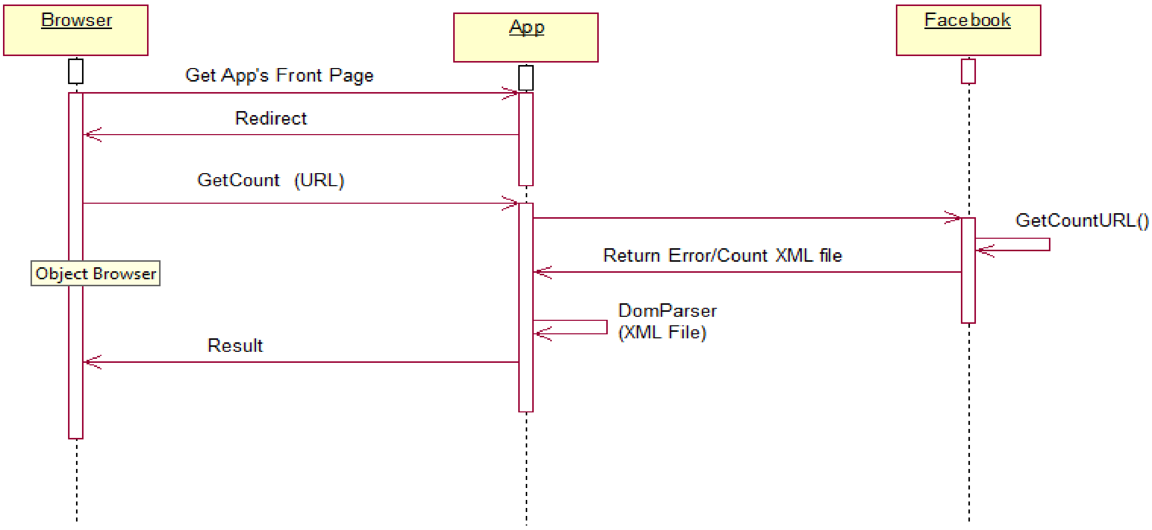


Fig 4.4: Sequential Diagram – use facebook API

For example facebook results for website www.dce.edu

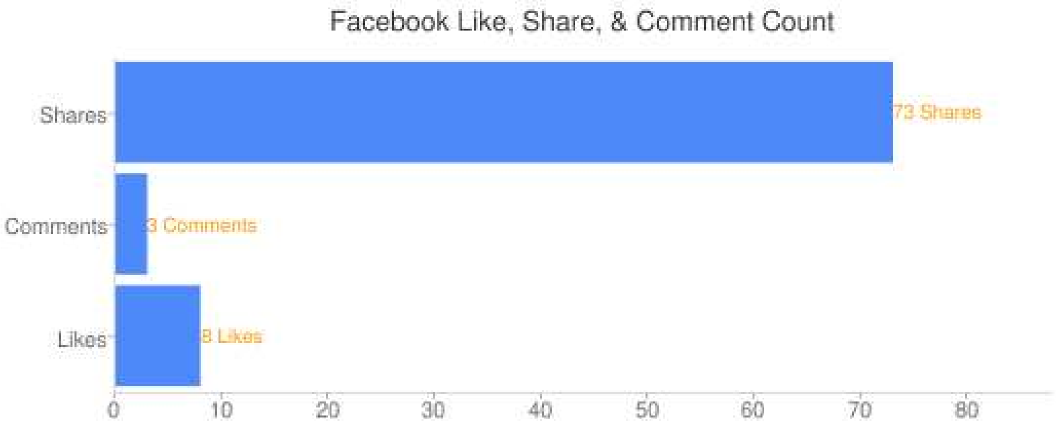


Fig 4.5: Facebook counts bar graph

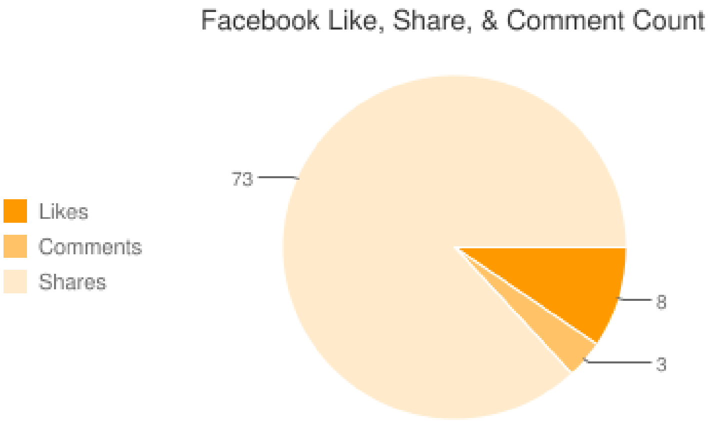


Fig 4.6 Facebook counts pie chart

4.4 Security Adviser

This search engine finds the security information for each site from antivirus sites such as Mcaffé and AVG and informs the user about the security of the site he wishes to visit.

We have used McAfee site advisor utility[27], in the current scenario to use this facility you have to install this utility to your system, so for a crawler website the URL has to be sent to the McAfee website, and in return the status of that URL is received (fig 4.7) and is saved. This is then displayed with the search results.

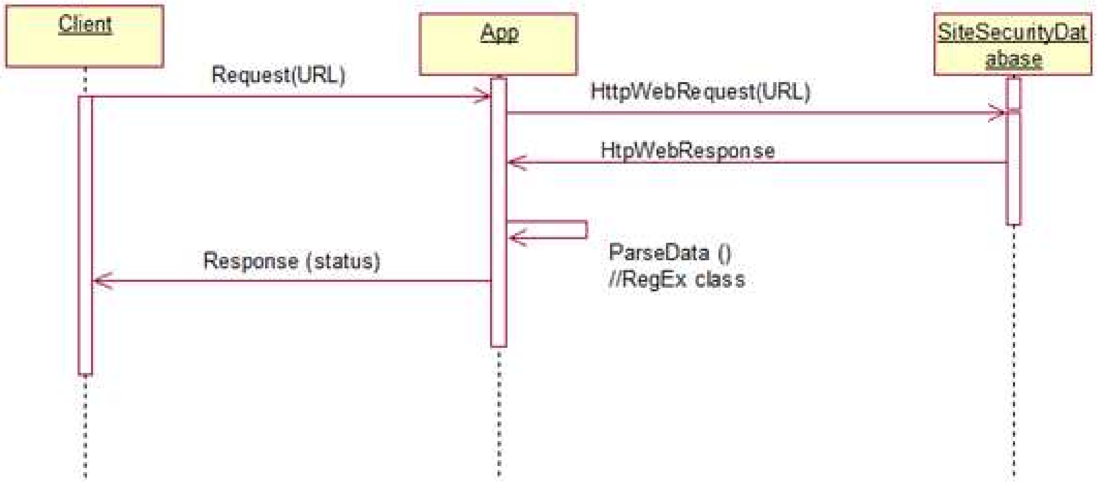


Fig 4.7: Sequence Dig fetch security status of an URL

To fetch the result from Mcafee web-extraction class libraries of .Net used. It done in two steps:

- First extract data, and performed GET & POST methods submission of data using HTTPWebRequest and HTTPWebResponse [28].
- Second Parsing data of regular expressions through RegEx class [29].
- Save status into form of annotations.

```
<Annotations>
  <Annotation about="www.dce.ac.in/*">
    <Label src=“green_label.png"/>
    <Comment>Secure Site</Comment>
  </Annotation>
    <Annotation about="www.songs.pk/*">
    <Label src=“red_label.png"/>
    <Comment>Unsecure Site</Comment>
  </Annotation>
 </Annotations>
```

The site ratings are based on tests conducted by mcafee site advisor. Symbols used to lead the user to web safety:

✓ - Secure site

⚠ - Minor risk

✖- High risk

●- Very high risk

4.5 Purpose Adviser

Another basic functionality this system proposes is Purpose Advice. By purpose advice we mean that this search engine will inform the user about the purpose of the site or say domain specific knowledge[30].

For e.g. If a user query is “Social Networking Site”

It will return the intended websites along with their purpose i.e.

Linkedin.com	For Business Purposes
Orkut.com	For Social Purposes
Facebook.com	For Social Purposes

This functionality provided by semantic web technology using ontology. Ontology represents class hierarchy for entities, their attributes and their relations at runtime without ambiguities. Ontologies provide dynamic and rich semantic information.

The logic is to extend the class hierarchy by ontologies:

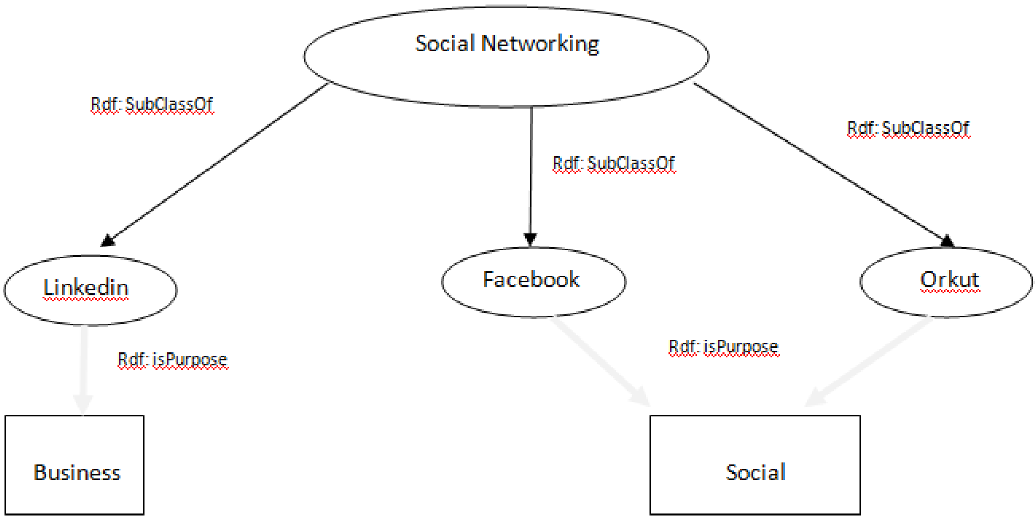


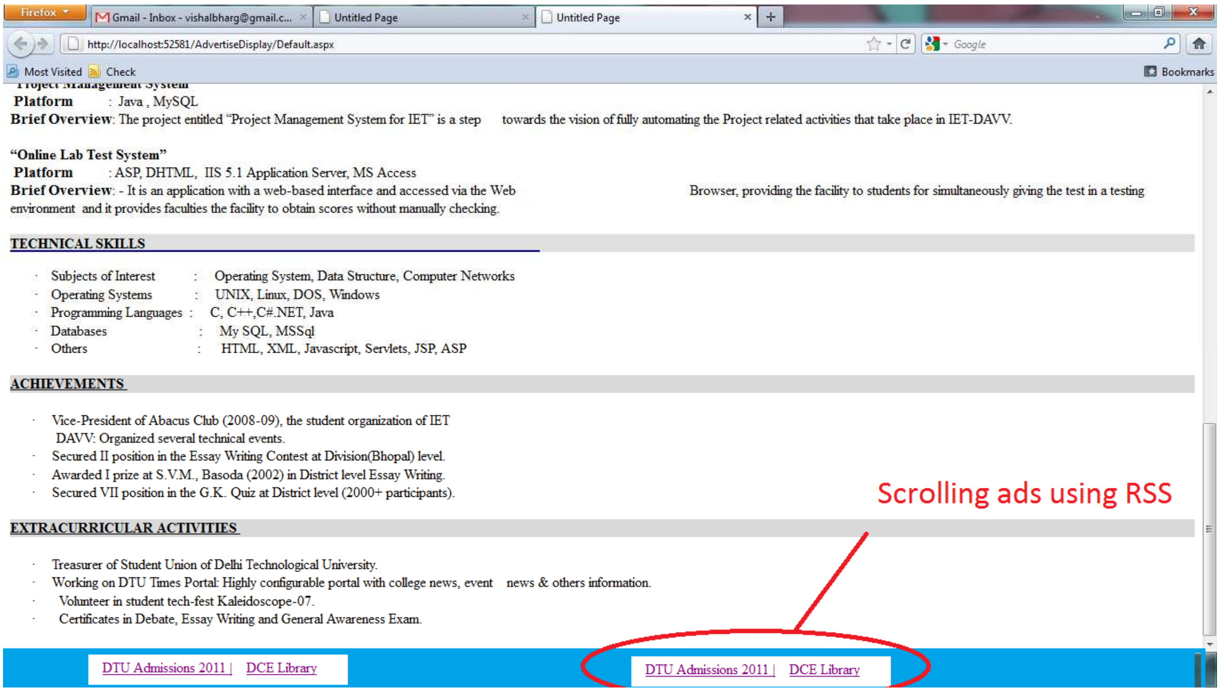
Fig 4.8: Ontology presentation

Depending on the used ontology, a SE node is either a Source or a SinglePort. The advantage is that the ambiguity of the father-child relationship does not lead to a conflict and a single (clean) class hierarchy is maintained.

4.6 Advertise with cable operator

The weirdest problem with web information quality is advertising. Website wants to earn money and that comes from advertise, so emphasize on advertise rather content. With growth of web services web advertising also grow rapidly, but it never compete with local cable advertise which directly touch with people.

So we make website advertise with local cable operator add, which scroll at bottom of webpage like bottom of television using RSS feed which same combined use by cable channel.


Firefox
Gmail - Inbox - vishalbharg@gmail.c...
Untitled Page
Untitled Page
http://localhost:52581/AdvertiseDisplay/Default.aspx
Google
Most Visited
Check
Bookmarks

Project Management System

Platform : Java , MySQL

Brief Overview: The project entitled "Project Management System for IET" is a step towards the vision of fully automating the Project related activities that take place in IET-DAVV.

"Online Lab Test System"

Platform : ASP, DHTML, IIS 5.1 Application Server, MS Access

Brief Overview: - It is an application with a web-based interface and accessed via the Web Browser, providing the facility to students for simultaneously giving the test in a testing environment and it provides faculties the facility to obtain scores without manually checking.

TECHNICAL SKILLS

Subjects of Interest	:	Operating System, Data Structure, Computer Networks
Operating Systems	:	UNIX, Linux, DOS, Windows
Programming Languages	:	C, C++,C#.NET, Java
Databases	:	My SQL, MSSql
Others	:	HTML, XML, Javascript, Servlets, JSP, ASP

ACHIEVEMENTS

- Vice-President of Abacus Club (2008-09), the student organization of IET DAVV: Organized several technical events.
- Secured II position in the Essay Writing Contest at Division(Bhopal) level.
- Awarded I prize at S.V.M., Basoda (2002) in District level Essay Writing.
- Secured VII position in the G.K. Quiz at District level (2000+ participants).

EXTRACURRICULAR ACTIVITIES

- Treasurer of Student Union of Delhi Technological University.
- Working on DTU Times Portal: Highly configurable portal with college news, event news & others information.
- Volunteer in student tech-fest Kaleidoscope-07.
- Certificates in Debate, Essay Writing and General Awareness Exam.

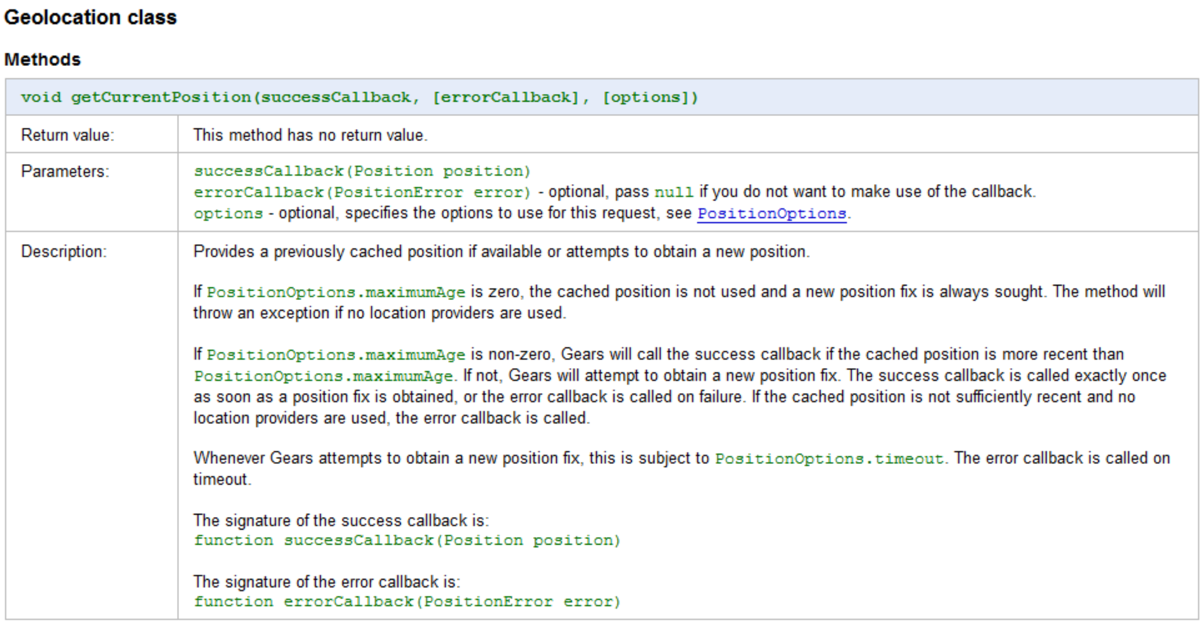
DTU Admissions 2011 | DCE Library

DTU Admissions 2011 | DCE Library

Scrolling ads using RSS

Fig 4.9: Scrolling advertise using rss feed

Using RSS (Really Simple Syndication) provides the facility of updation of dynamic content rather static content and this same content send to cable software which display on television. To know the location and latitude of user on website we use Google gears API and same use for cable software so nearest cable operators advertise displays on website. Google gears API totally free and it can easily embed [32] with your application. To access the location of user google provides geolocation APIs [31]. We override the getCurrentPosition method of this API to access current location of the user.



Geolocation class

Methods

void getCurrentPosition(successCallback, [errorCallback], [options])	
Return value:	This method has no return value.
Parameters:	`successCallback(Position position)` `errorCallback(PositionError error)` - optional, pass `null` if you do not want to make use of the callback. `options` - optional, specifies the options to use for this request, see `PositionOptions`.

Description:	Provides a previously cached position if available or attempts to obtain a new position.
	If `PositionOptions.maximumAge` is zero, the cached position is not used and a new position fix is always sought. The method will throw an exception if no location providers are used.
	If `PositionOptions.maximumAge` is non-zero, Gears will call the success callback if the cached position is more recent than `PositionOptions.maximumAge`. If not, Gears will attempt to obtain a new position fix. The success callback is called exactly once as soon as a position fix is obtained, or the error callback is called on failure. If the cached position is not sufficiently recent and no location providers are used, the error callback is called.
	Whenever Gears attempts to obtain a new position fix, this is subject to `PositionOptions.timeout`. The error callback is called on timeout.
	The signature of the success callback is:
	`function successCallback(Position position)`
	The signature of the error callback is:
	`function errorCallback(PositionError error)`

Fig 4.10: Geolocation class defined by Google geolocator API

Actually by using this API web-service got the name of all access point and their location and a combined database of access point and their location with nearest cable operator also manage by this service, so after got the all sufficient data nearest cable's advertise published on the web.

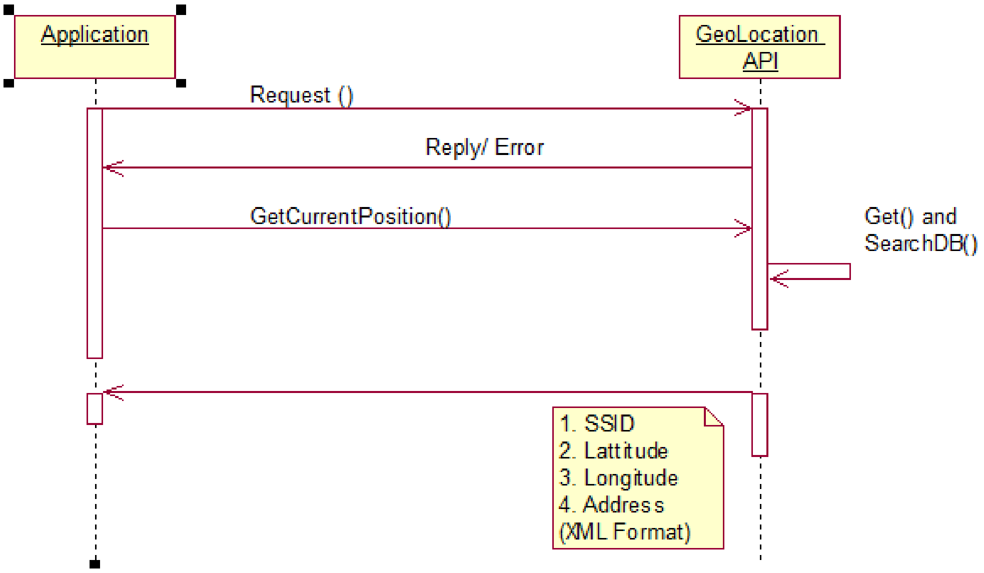
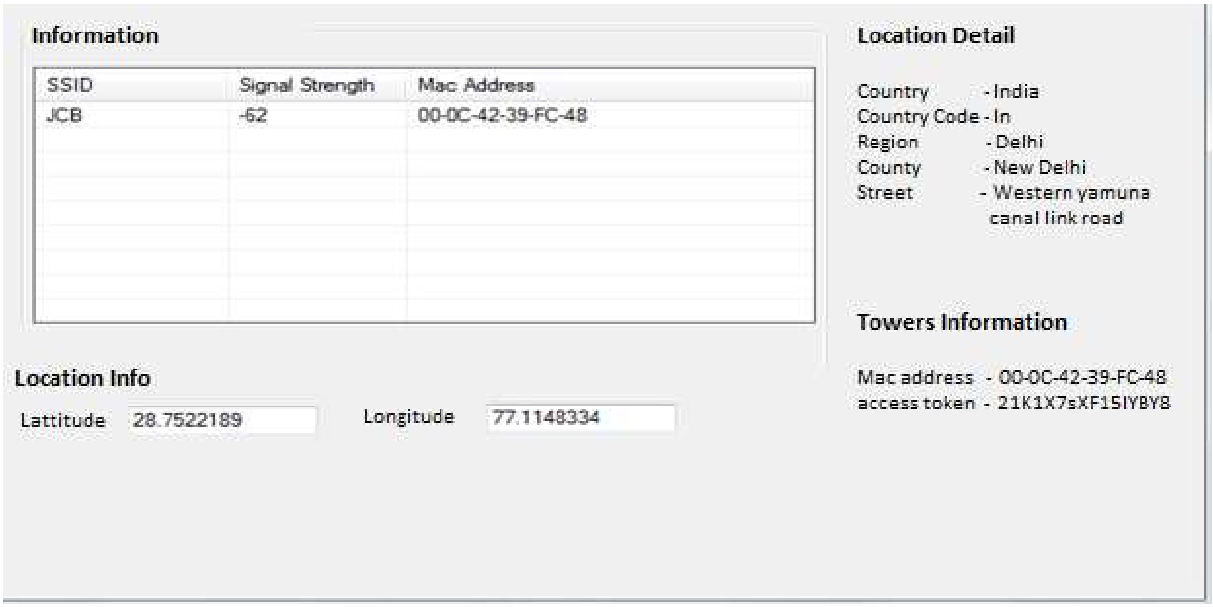


Fig 4.11:Sequence Dig to get user location by geolocator APIs



Information

SSID	Signal Strength	Mac Address
JCB	-62	00-0C-42-39-FC-48

Location Info

Lattitude 28.7522189

Longitude 77.1148334

Location Detail

Country - India
Country Code - In
Region - Delhi
County - New Delhi
Street - Western yamuna canal link road

Towers Information

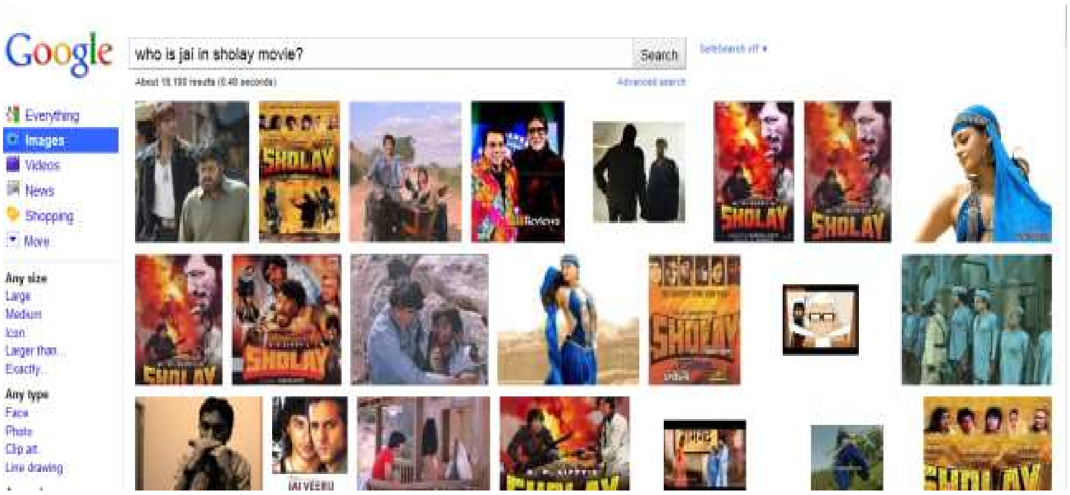
Mac address - 00-0C-42-39-FC-48
access token - 21K1X7sXF15IYBY8

Fig 4.12: Calculate user location by geolocator APIs

Chapter 5

Results

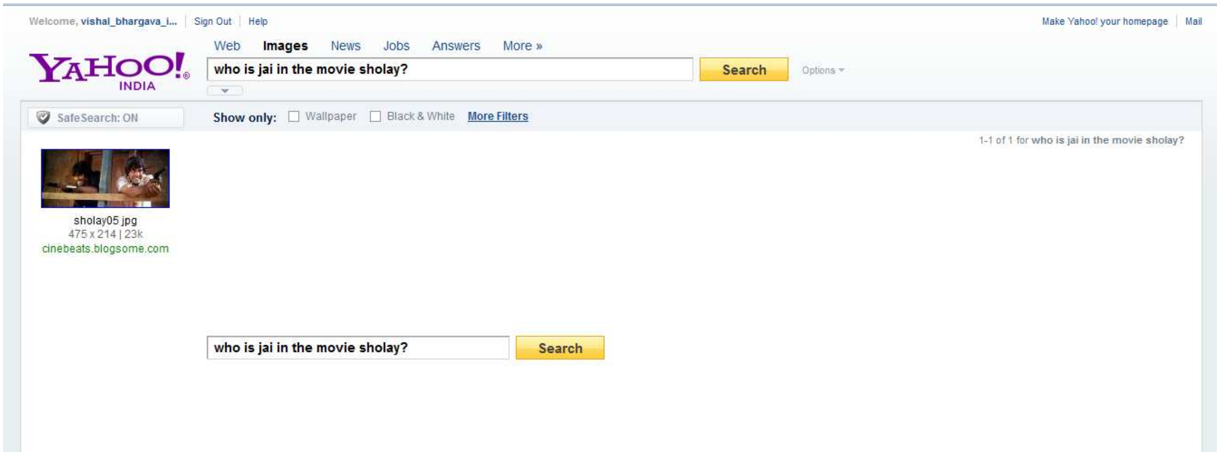
An image search for the query “who is jai in the movie sholay?” On Google & Yahoo we get the result shown below (fig 5.1 ,5.2).



who is jai in sholay movie?

About 18,100 results (0.40 seconds)

Fig 5.1: Google’s result



who is jai in the movie sholay?

SafeSearch: ON

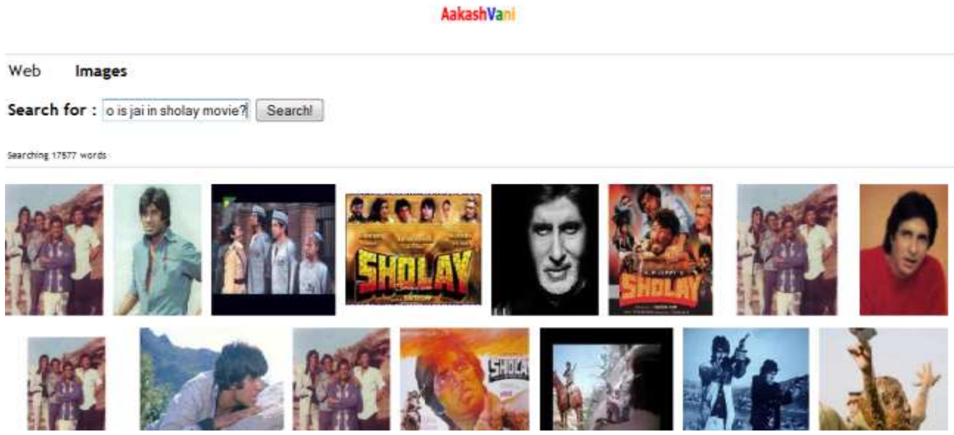
1-1 of 1 for who is jai in the movie sholay?

sholay05.jpg
475 x 214 | 23k
cinebeats.blogsome.com

who is jai in the movie sholay?

Fig 5.2: Yahoo’s result

The result by our search engine when presented with the same query is more accurate. The result is as shown below: (fig 5.3).



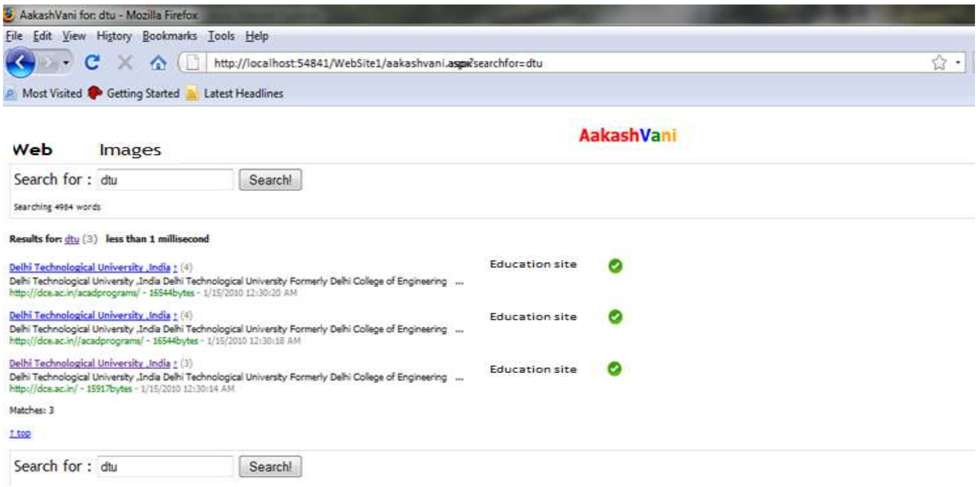
Search for : o is jai in sholay movie?

Searching 17577 words

Fig 5.3: Our search engine’s result

On the basis of Google, Yahoo and our results, it can conclude that our search engine return more accurate result in term of query.

Also our search engine provides advice for all the results in terms of security as well as recommendation as shown below (Fig 5.4).



http://localhost:54841/WebSite1/aakashvani.aspx?searchfor=dtu

Search for : dtu

Searching 4984 words

Results for: dtu (3) less than 1 millisecond

Delhi Technological University ,India ↑ (4)
Delhi Technological University ,India Delhi Technological University Formerly Delhi College of Engineering ...
http://dce.ac.in/acadprograms/ - 16544bytes - 1/15/2010 12:30:20 AM
Education site ✓

Delhi Technological University ,India ↑ (4)
Delhi Technological University ,India Delhi Technological University Formerly Delhi College of Engineering ...
http://dce.ac.in//acadprograms/ - 16544bytes - 1/15/2010 12:30:18 AM
Education site ✓

Delhi Technological University ,India ↑ (3)
Delhi Technological University ,India Delhi Technological University Formerly Delhi College of Engineering ...
http://dce.ac.in/ - 15917bytes - 1/15/2010 12:30:14 AM
Education site ✓

Matches: 3

Search for : dtu

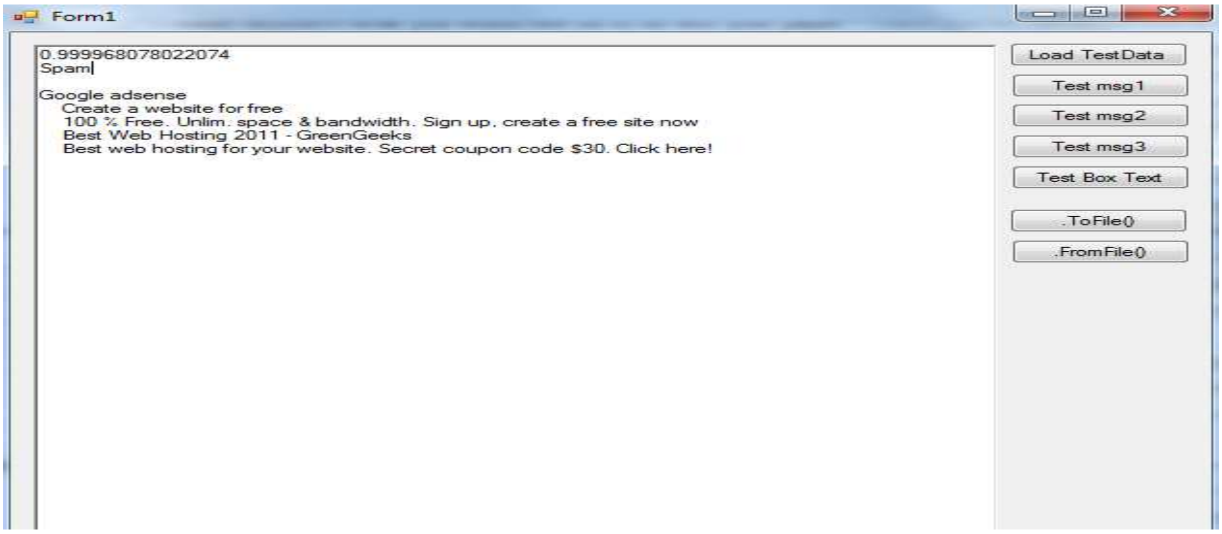
Fig 5.4: Security and Recommendation advice

We perform our research on 50 users and check their click behavior according to their preference on security advice, so with security advice following result found on web visit according to their security status in compare of previous visits.

Table 1: Websites visit status with and without security advice

Secure URLs visited	100%
Minor risk URLs visited	78.75%
High risk URLs visited	32.25%
Very high risk URLs visited	12%

Google and Yahoo also providing advertise service on the web. Even google's adsense service open to all, it can be use by anyone, just put some ads content provided by google on your website and earn according to pay per click rule. Google just provide little bit share to adsense user what it got from their client. This is best business for search engines so on crawling time of website they don't take their advertise in their account so a website with little information and lots of advertise also come into top ranking by these search engines, but our search engine also check advertise as content then generate the score for particular website.



```
0.999968078022074
Spam

Google adsense
    Create a website for free
    100 % Free. Unlim. space & bandwidth. Sign up, create a free site now
    Best Web Hosting 2011 - GreenGeeks
    Best web hosting for your website. Secret coupon code $30. Click here!
```

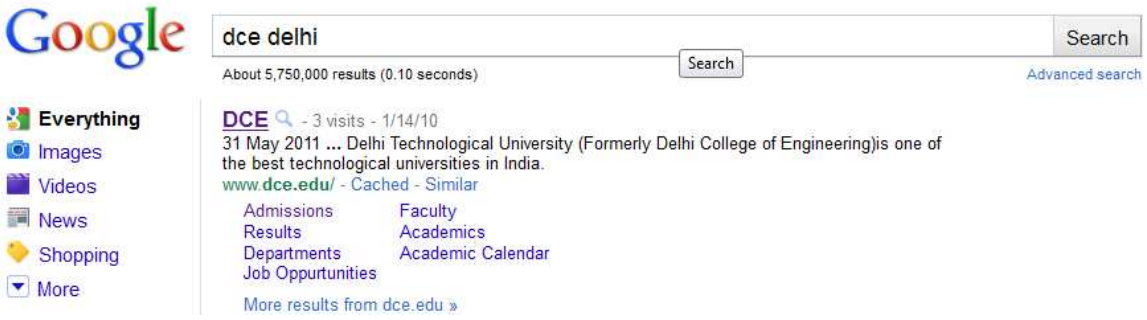
Fig 5.5: Checking Google adsense

When we merge social tagging with statically ranking, results for search query “DCE Delhi”

Table 2: Comparison of results for query “DCE Delhi”

Google’s Result	Yahoo’s Result	Our Search Engine’s Result
www.dce.edu/ http://maps.google.com/maps/place?dce www.admissions.dce.edu/ www.dce.edu/web/.../BTech_Admission_Policy_2010-11.php www.dce.edu/web/Sections/.../undergraduateprogramme.php dce.ac.in/ dce.ac.in/aboutdce/history.php dcedelhi.batchmates.com/ en.wikipedia.org/wiki/Delhi_Technological_University **www.successcds.net/.../Delhi-College-of-Engineering-DCE.html**	www.dce.edu www.dce.ac.in www.dce.edu/web/Sections/Admissions/undergraduateprogramme.php dce.ac.in/departments www.dceonline.net delhiassembly.nic.in/index.asp globalshiksha.com/college/Delhi-College-of-Engineering-DCE-/ dce-edu.com www.admissionnews.com/collegedetails.aspx?colid=11276 delhi.justdial.com/book-dce_Delhi.html	www.dce.edu www.dce.ac.in dcedelhi.batchmates.com/ rutsum.com/dce-becomes-dtu en.wikipedia.org/wiki/Delhi_Technological_University news.education4india.com/4573/dce-cee-delhi-college-of-engineering/ www.indiaedu.com/top-educational.../top.../dce-delhi.html www.collegekhabar.com/news511.htm ttp://www.jobsamiksha.com/index.php?option=c….

On the basis of results for the query we can see google provides mostly results from same domain (dce.edu) while it provides the feature to group of all results (fig 5.6), yahoo provides statistical result but our search facilitate results from several domain mostly suggest by users (like collegekhabar).



About 5,750,000 results (0.10 seconds)

DCE - 3 visits - 1/14/10
31 May 2011 ... Delhi Technological University (Formerly Delhi College of Engineering)is one of the best technological universities in India.
www.dce.edu/ - Cached - Similar

Admissions Faculty
Results Academics
Departments Academic Calendar
Job Oppurtunities

More results from dce.edu »

Fig 5.6: grouping of results from same domain

Chapter 6

Conclusion and Future Work

Search engines with increasingly complex functionalities and behavior are being developed. This report presents an easy to use, flexible and secure search engine.

Our search results satisfy the security criteria as well as our requisite NLP and Q-A features:

(1)It displays the credibility of Web sites as recommended by security companies and antivirus firms.
(2) It represents the most voted information available in terms of social tagging and
(3) Image search result remains absolutely relevant to the query (NLP).
(4) Site recommendation put great impact on information.
(5) Impacted way to represent advertise on web has proposed.

Thus our search engine provides not only the basic features of matching the query will results that any standard search engine provides but also provides additional features like security, site advisor and query search in images, video and audio etc.

However our system is not able to differentiate between queries like “Who is James Bond?” And “Who was James Bond?” Merging of social tagging with statistical searching algorithm also depend on user behavior so need of personalization search behavior is there

Thus it is more than a search engine but a recommendation engine.

Future work:

1. Merging of social ranking with statistical ranking also depends on user specification so the need of personalized search engine felt.
2. To develop proper ontology for purpose advice.

3. Sites which reports are not given by Mcafee, their ranking not determined. It can be secure or unsecure so can’t predict.
4. With image search NLP it can be apply on other things video, books search.
5. Advertise websites are always detected as spam site, so separate classification for this type of site needed.
6. Advertise can also be interactive that means on demand advertise.

References

1. Atsaros, G.; Spinellis, D.; Louridas, P.,“ Site-Specific versus General Purpose Web Search Engines: A Comparative Evaluation” , Informatics, 2008. PCI '08. Panhellenic Conference,pp 44-48
2. Smith, L.S.; Hurson, A.R.,"A search engine selection methodology”, Information Technology: Coding and Computing [Computers and Communications], 2003. Proceedings. ITCC 2003. International Conference,pp 122-129
3. Fernando A. Mikic Fonte,Juan Carlos Burguillo Rial,Martín Llamas Nistal,”TQ-Bot: An AIML-based Tutor and Evaluator Bot”,Journal of Universal Computer Science, vol. 15, no. 7 (2009),pp 1486-1495
4. Walter, F., Battiston, S., & Schweitzer, F. (2008). A model of a trust-based recommendation system on a social network. Autonomous Agents and Multi-Agent Systems, 16(1), 57–74.
5. Gori,M.,&Witten, I. (2005). The bubble of web visibility. Communications of the ACM, 48(3), 115–117.
6. Cai Dongfeng ; Cui Huan ; Miao Xuelei ; Zhao Chenguang ; Ren Xiangshi ; Shenyang Inst. of Aeronau “ A Web-based Chinese automatic question answering system” Computer and Information Technology, 2004. CIT '04.
7. http://www.ask.com
8. A report “THE EFFECT OF SOCIAL NETWORKING SITESON PERSONAL LIVES OF THE PEOPLE 2009” http://www.scribd.com/doc/13653301/The-Effect-of-Social-Networking-Sites
9. http://www.delicious.com/
10. Haiyan Wang , Runsheng Zhou , Yi Wang “An anti-spam filtering system based on the Naive Bayesian Classifier and Distributed Checksum Clearinghouse” 2009 Third International Symposium on Intelligent Information Technology Application 978-0-7695-3859-4/09
11. http://www.google.com/support/websearch/bin/answer.py?answer=45449

12. Rebeca P. D´ıaz Redondo, Ana Fern´andez Vilas, Jos´e Juan Pazos Arias, Manuel Ramos Cabrer, Albert Gil Solla, and Jorge Garc´ıa Duque "Bringing Content Awareness to Web-Based 1IDTV Advertising" 1094-6977/$26.00 © 2011 IEEE
13. Jing Deng, Chi-Hung Chi "Local Web Advertisement Through Dynamic Active Proxy" Multimedia and Expo, 2000. ICME 2000. 2000 IEEE International Conference ISBN No. 0-7803-6536-4
14. Junaidah Mohamed Kassim and Mahathir Rahmany "Introduction to Semantic Search Engine" 2009 International Conference on Electrical Engineering and Informatics 5-7 August 2009, Selangor, Malaysia
15. Konstantinos Christidis, Niki Papailiou and Gregoris Mentzas "Exploring Gadget-based Interfaces for the Social Semantic Desktop", Proc. of the 2009 13th Panhellenic Conference on Informatics.
16. M. Kantardzic, Data Mining - Concepts, Models, Methods, and Algorithms, IEEE Press, Wiley-Interscience, 2003, ISBN 0-471- 22852-4.
17. Jiawei Han and Micheline Kamber, Data Mining: Concepts and Techniques, Elsevier 2006, ISBN 1558609016.
18. Naive Bayesian Classifier by K. Ming Leung
19. W3C Web DOM Specifications www.w3.org/DOM/
20. www.alicebot.org/aiml.html
21. www.w3.org/Protocols/HTTP/Methods.html
22. http://support.ipswitch.com/kb/IM-20030513-DM01.htm
23. M. Kantardzic, Data Mining - Concepts, Models, Methods, and Algorithms, IEEE Press, Wiley-Interscience, 2003, ISBN 0-471-22852-4.
24. Jiawei Han and Micheline Kamber, Data Mining: Concepts and Techniques, Elsevier 2006, ISBN 1558609016.
25. Nikravesh, M.,"Neuroscience and precisiated natural language-based search and question answering system: neusearch", Granular Computing, 2005 IEEE International Conference on 25-27 July 2005,pp 45-50, Vol1

26. Heymann, Paul; Paul; Koutrika, Georgia; Garcia-Molina, Hector (February 12, 2008). "Can Social Bookmarking Improve Web Search?”First ACM International Conference on Web Search and Data Mining. http://dbpubs.stanford.edu:8090/pub/2008-2. Retrieved 2008-03-12.
27. http://www.siteadvisor.com
28. http://msdn.microsoft.com/en-us/library/system.text.regularexpressions.regex.aspx
29. http://msdn.microsoft.com/en-us/library/8y7x3zz2%28vs.71%29.aspx
30. Lei Zhang; Jun-Liang Chen; Shang-Meng Li; Yong Peng,"Study on domain-specific search engine and its automated generation",Machine Learning and Cybernetics, 2008 International Conference Volume: 3,pp 1637-1642
31. http://code.google.com/apis/gears/api_geolocation.html
32. With the New Google Latitude API, Build Latitude and Location Into Your App, by Ana Ulin, Google Code Blog, May 19, 2010. Retrieved November 25, 2010
33. Walter, F., Battiston, S., & Schweitzer, F. (2008). A model of a trust-based recommendation system on a social network. Autonomous Agents and Multi-Agent Systems, 16(1), 57–74.
34. Gori,M.,&Witten, I. (2005). The bubble of web visibility. Communications of the ACM, 48(3), 115–117.
35. http://en.wikipedia.org/wiki/Gopher_protocol
36. http://en.wikipedia.org/wiki/Archie_search_engine
37. http://en.wikipedia.org/wiki/Veronica_(computer)
38. Brin, Sergey and Page Lawrence. The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems, April 1998
39. Baldi, Pierre. Modeling the Internet and the Web: Probabilistic Methods and Algorithms, 2003.
40. Chakrabarti, Soumen. Mining the Web: Analysis of Hypertext and Semi Structured Data, 2003
41. Grossan, B. “Search Engines: What they are, how they work, and practical suggestions for getting the most out of them,” .http://www.webreference.com
42. Pant, Gautam, Padmini Srinivasan and Filippo Menczer: Crawling the Web, 2003. http://dollar.biz.uiowa.edu/~pant/Papers/crawling.pdf

43. Search Indexing Robots and Robots.txt, 2002 http://www.searchtools.com/ robots/robots-txt.html
44. Garcia-Molina, Hector. Searching the Web, August 2001 http://oak.cs.ucla.edu/~cho/papers/cho-toit01.pdf
45. Liu, Jian. Guide to Meta-Search Engines. Reference Department, Indiana University Libraries. June 1999 http://www.indiana.edu/~librcsd/search/meta.html
46. www.altavista.com
47. Bates, M. (1995).”Models of natural language understanding”. Proceedings of the National Academy of Sciences of the United States of America, Vol. 92, No. 22 (Oct. 24, 1995), pp. 9977–9982.
48. Bausch, S., Han, L.(2006). Social networking sites grow 47 percent, year over year, reaching 45 percent of web users, according to nielsen//netratings. Retrieved December 16, 2007, from http://www.nielsen-netratings.com/pr/pr_060511.pdf
49. Boyd, D. M., & Ellison, N. (2007). Social Network Sites: Definition, History, and Scholarship. Journal of Computer-Mediated Communication, 13(1), article 11. Retrieved December 16, 2007, from jcmc.indiana.edu/vol13/issue1/boyd.ellison.html
50. O'Hanlon, C. (2007). If You Can't Beat 'Em, Join 'Em. T.H.E. Journal. 34(8), 39-40, 42, 44. Retrieved December 16, 2007, from ERIC database.
51. Om Prakash Verma, Rahul Katarya, Vishal Bhargava & Nikhil Maheshwari “Use of Semantic Web in Enabling Desktop based Knowledge Management” Apr 2011 3rd International Conference on Electronics Computer Technology, Kanyakumari.