Semantic Based Multiple Search Engine with Anti Phishing Technique

Dr. S. Latha Shanmugavadivu [1], Dr. M. Rajaram [2]

[1.] Department of ECE, Tamilnadu College of Engineering, Coimbatore - 641656.
[2.] Department of Electrical Engineering, Government College of Technology, Coimbatore - 641013.
latha_tce@yahoo.co.in

Abstract: Search engines are computer programs that travel the Web, gather the text of Web pages and make it possible to search for them. Search engines are rugged individualists and none of them index the same Web information, and none of them search the content in the same way. Search engines are the wonder of the Internet. The main purpose of this approach is to develop a search engine based on ontology matching within the Semantic Web. In order to make the Semantic Web work, well-structured data and rules are necessary for agents to roam the Web. Extensible Markup Language (XML) and Resource Description Framework (RDF) are two important approaches used. Technically, in order to be called a Search Engine, a search tool must be made up of three major components the Interface an Index and Crawlers or Spiders. Multi- search enables the user to gather results from its own search index as well as from one or more search engines, metasearch engines, databases or any such kind of information retrieval programs. Multisearch is an emerging feature of automated search and information retrieval systems which combines the capabilities of computer search programs with results classification. The basic idea is reducing the amount of time required to search for resources by improvement of the accuracy and relevance of individual searches as well as the ability to manage the results. The next proposed approach focuses on the Anti-Phishing technique to avoid phishing attacks. In this approach, visual features are evaluated from the computer screen by image processing technicques.

[S. Latha Shanmugavadivu, M. Rajaram. **Semantic Based Multiple Search Engine with Anti Phishing Technique**. *Life Sci J* 2013;10(9s):284-288] (ISSN:1097-8135). http://www.lifesciencesite.com. 40

Keywords: Multisearch engines; XML; resource description framework; ontology; semantic web

1. Multi-Search Engine

There are various search engines available which covers both general and specific subjects, and which search specific elements of the Internet such as Web pages or Usenet.

While some of the search engines are particularly effective and sophisticated, none of them are entirely comprehensive. These search engines may only use a small database to create the set of results (Yahoo for example only indexes a very small proportion of the 3 billion pages indexed by Google), or they may not be updated quickly (All the other web update every fortnight or so, while Google updates monthly). Their spider programs are not very fast, which means that their currency might not be a real reflection of the state of play on the Internet.

Therefore, even if the user have a favourite search engine, or even several of them, to guarantee anything like a comprehensive search, the user may need to use few search engines before the user is satisfied that every required information is obtained on a particular topic. A Multi-search engine avoids the trouble of going to a variety of different sites in order to run the search, or it may suggest a search engine which the user had not considered, or perhaps the user had never know.

1.1. Characteristics of Multi-Search Engine

In any words, all words, phrase searching, there is little that the Multi-search engine can do directly about this since they are unable to affect the internal workings of individual search engines. However, it is an option that should be offered to the end user; if one search engine can search on a phrase out of the list available it seems short sighted not to offer this. Other engines on the list will simply ignore the phrase aspect and search on the words using an implied OR. If this is not given as an option though, it reduces the effectiveness of those search engines which can undertake phrase searching.

Phishing is a form of amalgamation of Web technology and social engineering. The most popular phishing tricks are executed using phishing web pages. Phishing web pages are forged to imitate certain legal companies' web pages. Phishing websites generally trigger users into leaking their sensitive information and private data by faking trustworthy web identities. Trusted users may easily be deceived by such tricks. Victims of phishing web pages may be forced to expose their bank accounts, passwords, credit card numbers, or other important information to malicious people. Nowadays, phishing activities have continued to flourish in spite of the technological measures put in place by organizations.

The anti-phishing techniques can be structured into three levels (Anthony Yingjie Fu (2006)): visual assessment (graphic level), semantic assessment (text level), and human computer interaction (HCI) enforcement. Both graphic level phishing attacks and text level attacks occur through the interface between human and computer, such that the improvement of HCI approaches directly impacts the security and usability of Web applications. Another very important and essential problem is protecting web pages by providing strong evidence to prove their originality, such that we can define phisher and victim.

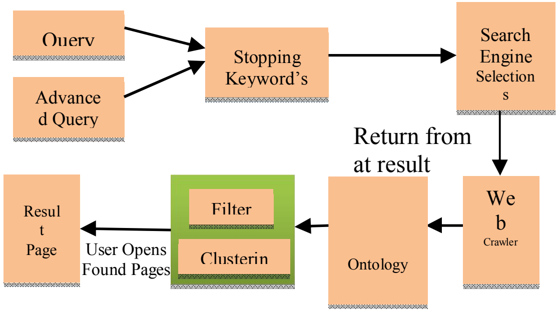


Figure 1. Semantic Web infrastructures (prototype)

2. Research Motivation

The main motivation behind the very basic relevance criterions underlying their information retrieval strategies (Anh and Moffat 2002), rely on the presence of query keywords within the returned pages. Moreover, the time taken by the search engines for the query result is also meaningless. The hit ratios of the web pages are very important factor for assessing the importance and the efficiency of the websites.

The other problem in the websites is the phishing attacks. Phishing is an emerging type of social engineering crime on the Web. Most phishers initiate attacks by sending emails to potential victims. These emails lure users to access fake websites, and induce them to expose sensitive and/or private information. Phishing detection techniques are available but the results are not eloquent.

2.1 Analysis of the Phishing Techniques

There are several techniques available in the literature for avoiding the phishing attacks. But most of the phishing techniques suffer from several drawbacks. Some of the drawbacks of the phishing techniques are

- Time consuming and algorithm complexity
- Failure rate in detection is more
- Less burden on the user and high burden on the attacker

3. Objectives of the Research

The main aim of this research to develop a multiple search engine with efficient search query result and less time complexity, to develop a typical phishing detection technique using image processing technique which shows indicative performance, to diminish network traffic and Latency time.

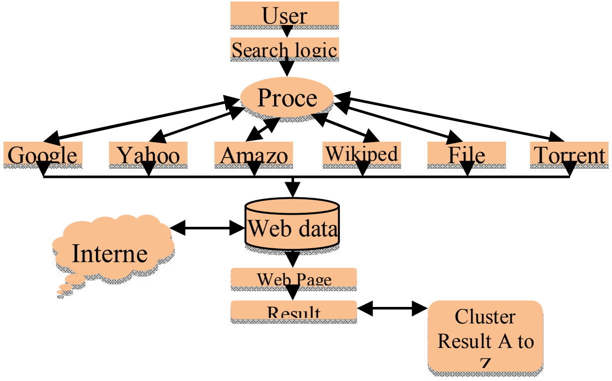


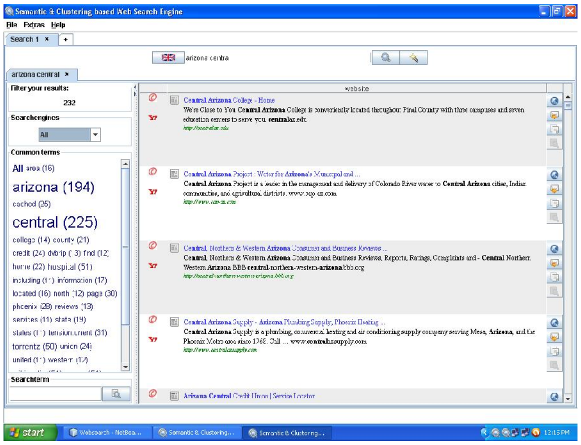
Figure 2. Diagrammatic Representation of the Proposed Approach

3.1 Search Services

Search services broadcast user queries to various search engines and various other information sources concurrently. Then they combine the results submitted by these sources, check for duplicates, and present them to the user as an HTML page with clickable URLs. For example, *IBM* InfoMarket searches Yahoo, OpenText, Magellan, various business resources, and Usenet newsgroups simultaneously and generates a rank-ordered query output. MetaCrawler is another search service that sends queries to eight different search engines: OpenText, Lycos, WebCrawler, InfoSeek, Excite, AltaVista, Yahoo, and Galaxy. MetaCrawler supports both Boolean and phrase search.

4. Proposed Work

The tremendous and explosive growth of information available to end users through the Web makes the search engines very important in the current scenario. Nevertheless, because of their general-purpose approach, it is always less uncommon that obtained results provide a burden of useless pages. It is not uncommon that even the most renowned search engines return results including many pages that are definitely useless for the user. This is mainly due to the fact that the very basic relevance criterions underlying their information retrieval strategies rely on the presence of query keywords within the returned pages. Moreover, the time taken by the search engines for the query result is also unimportant. The hit ratios of the web pages are very important factor for assessing the importance and the efficiency of the websites. Hence new approaches are needed to deal with the above important issues. This research focus on the development of search engines which is very effective and efficient.



arizona centra

232

All area (16) arizona (194) cached (25) central (225) college (14) county (21) credit (24) dwbip (13) find (12) home (22) hospital (51) including (11) information (17) located (16) north (12) page (30) phoenix (28) reviews (13) services (11) state (19) states (11) tension.rent (31) torrentz (50) union (24) united (11) western (12)

Central Arizona College - Home
We're Close to You Central Arizona College is conveniently located throughout Pinal County with three campuses and seven education centers to serve you. centralaz.edu
http://centralaz.edu

Central Arizona Project : Water for Arizona's Municipal and ...
Central Arizona Project is a leader in the management and delivery of Colorado River water to Central Arizona cities, Indian communities, and agricultural districts. www.cap-az.com
http://www.cap-az.com

Central, Northern & Western Arizona Consumer and Business Reviews ...
Central, Northern & Western Arizona Consumer and Business Reviews, Reports, Ratings, Complaints and - Central Northern Western Arizona BBB central-northern-western-arizona.bbb.org
http://central-northern-western-arizona.bbb.org

Central Arizona Supply - Arizona Plumbing Supply, Phoenix Heating ...
Central Arizona Supply is a plumbing, commercial heating and air conditioning supply company serving Mesa, Arizona, and the Phoenix Metro area since 1968. Call ... www.centralazsupply.com
http://www.centralazsupply.com

Arizona Central Credit Union | Service Locator

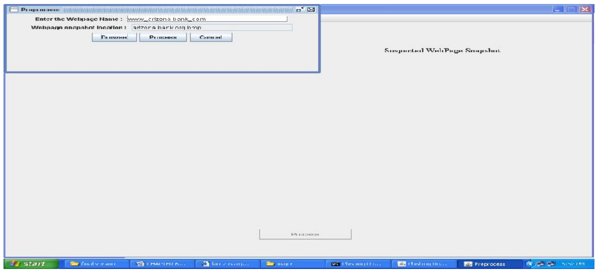
Figure 3. Result window for Semantic Based Multiple Search Engine

New approaches are necessary in the areas of web search engines which can provide higher efficiency and accuracy. This research mainly deals with "Semantic Based Multiple Web Search Engine with antiphising technique".

The ontology file and metadata extractor package is used to extract the metadata from multiple web pages. The web crawler travels the web and collects the content on the Web. Query is based on metasearch. The query builder is the bridge between the user and machine learning agent. It accepts the input of the user and constructs a kind of "query language". Then it transfers the query to machine learning agent. It has a friendly interface that will also ease the user input. The query value is matched into multiple websites with maximum occurrence of the term, the result based on the term frequency.

The main aim of this thesis is to develop a search engine based on ontology matching within the Semantic Web. This work is to match those elements, and then bring more accuracy in the search result which can provide higher efficiency. The methodologies used in the implementation of antiphising techniques are:

i. Java – For Business Logic
ii. Java Swings – For User Interface
iii. JAI – Advanced Image handling package for java
iv. Webpage Capture – To capture webpage image
v. ImageMagick – To crop the captured webpage images



Enter the Webpage Name : www_arizona bank_com
Webpage snapshot location : arizona bank org.bmp

Figure 4 Pre-processing of the Original and Suspected Web Page

5. Taxonomy for Search Tools and Services

Automated techniques for retrieving information on the Web can be broadly classified as search tools or search services. Search tools utilize robots for indexing Web documents. They feature a user interface for specifying queries and browsing the results. The main part of a search tool is the search engine, which is the main cause for searching the index to retrieve documents relevant to a user query. Search services offer users a layer of abstraction over various search tools and databases and aim at simplifying the Web search. Search tools are described along the following dimensions:

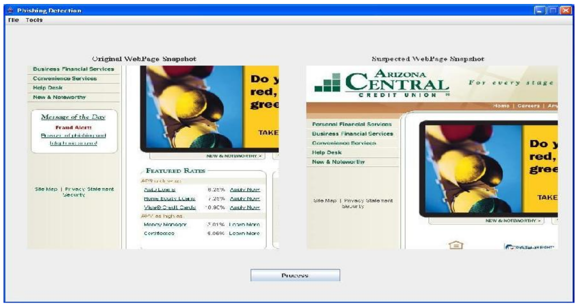
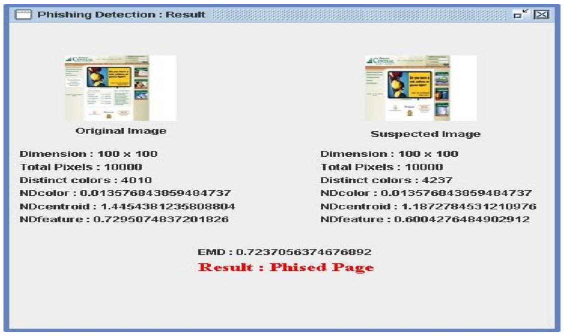

Phishing Detection
File Tools
Original WebPage Snapshot
Suspected WebPage Snapshot
Business Financial Services
Convenience Services
Help Desk
New & Noteworthy
Message of the Day
Fraud Alert!
Site Map | Privacy Statement Security
Do y red, gree
TAKE
NEW & NOTEWORTHY »
FEATURED RATES
Auto Loans 8.25% Apply Now
Home Equity Loans 7.25% Apply Now
Visa® Credit Cards 10.90% Apply Now
Money Manager
Certificates
Learn More
ARIZONA CENTRAL CREDIT UNION
For every stage
Home | Careers
Personal Financial Services
Process

Figure 5 Processing of the Original and Suspected Web Page

- Methods for Web navigation,
- Indexing techniques,
- Query language or specification approach for expressing user queries,
- Approaches for query-document matching, and
- Methods for presenting the query output

In Figure 1, the crawler application gathers annotated Web pages from the Semantic Web including RDF metadata and originating OWL ontology. RDF metadata are interpreted by the OWL parser and stored in the knowledge database. A graphics user interface facilitates for the definition of a query, which is passed on to the relation-based search logic. The ordered result set constructed by this latter module is finally presented to the user. Figure 2 shows the diagrammatic representation of the proposed approach.



```
Dimension : 100 x 100                       Dimension : 100 x 100
Total Pixels : 10000                        Total Pixels : 10000
Distinct colors : 4010                      Distinct colors : 4237
NDcolor : 0.013576843859484737              NDcolor : 0.013576843859484737
NDcentroid : 1.4454381235808804             NDcentroid : 1.1872784531210976
NDfeature : 0.7295074837201826              NDfeature : 0.6004276484902912

                    EMD : 0.7237056374676892
                    Result : Phised Page
```

Figure 6 Phishing Detection Result

6. Advantages

- The Semantic Web will offer the way for solving this problem at the architecture level. In fact, in the Semantic Web, each page possesses semantic metadata that record additional details concerning the Web page itself.
- This method is applied on each property individually and requires exploring all the Semantic Web instances.

7. Implementation

The results of an extensive set of simulation tests are shown in figure 3 which the proposed approaches are compared under a wide variety of different scenarios.

For a single user query Arizona Central the result is got from multiple search engines, such as Google, yahoo, AltaVista, Clusty, Excite, All the web, File tube, You tube, Amazon, Cite seer, Wikipedia, Isohunt etc., here the answer is got from 25 search engines, searches take place simultaneously from different search engines, the main advantage is that searches do not have to wait for each search engine, the results are computed simultaneously from multiple search engines and they are displayed on the screen. The result is therefore much faster.

Java in relation to other programming languages lets one write special programs called applets that can be downloaded from the Internet and played safely within a web browser. Even more dangerous software would be promulgated if any web page visited could run programs on a system. No way of checking these programs for bugs or for out-and-out malicious behavior before downloading and running them.

Java solves this problem by severely restricting what an applet can do. A Java applet cannot write to hard disk without permission. It cannot write to arbitrary addresses in memory and thereby introduce a virus into a computer. It should not crash the system.

As a programming language, Java can create all kinds of applications that one could create using any conventional programming language. As a development environment, Java technology provides a large suite of tools like compiler (javac), an interpreter (java) and a documentation generator (javadoc).

Java is used in this research work. The platform used for these proposed approaches is Windows XP. The processor used is Pentium IV. The experimentation needs a system RAM of 2 GB.

The proposed approaches used for the experimental observations are

- Semantic Based Multiple Web Search Engine
- Detection of Phishing Web Pages Based On Image Processing Techniques

The performances of these approaches were compared based on certain parameters like Search result, Accuracy, Time Complexity, Web cache optimization and Phishing Result.

8. Detection of Phishing Web Pages on Image Processing Techniques

The implementation process contains the following modules Pre-processing of Original and Suspected Web pages, Signature Generation, Computation of Earth Movers Distance (EMD) and detecting phishing.

8.1 Pre-processing of Original and Suspected Web pages

The original and suspected web pages are pre-processed as shown in Figure 4. The snapshots of the web page images are taken and the interest regions are cropped. The image is resized to 100 x 100 using Lanczos algorithm.

8.2 Signature Generation

From the 100 x 100 pixel resized images, the dominant colours and its centroid were computed. The colour and centroid were computed for original as well as suspected web pages. Figure 5 shows the processed image.

8.3 Computation of EMD and Detecting Phishing

The EMD is calculated between the two web page images. If the EMD value is equal to 0, then the suspected page is original else if the EMD value is equal to 1, then the suspected page is phished page. Figure 6 depicts the computation of the EMD, which detects phishing.

9. Conclusion

The proposed approaches were evaluated on the java platform.It is seen that semantic based multiple web search engine shows very high accuracy and reduces time complexity. Thus the semantic approach is very significant in terms of search accuracy. The time complexity is also reduced greatly in this approach. The searches take place simultaneously from different search engines, and the searches do not have to wait for each search engine, the results are computed simultaneously from multiple search engines and they are displayed on the screen. The result is therefore much faster.

In order to further prove the efficiency of the proposed semantic search engine, the time taken to access a typical website was calculated for all types of search engines individually. It was found that the proposed semantic search engine was able to catch the website 75% faster than other types.

The cached web page from the designed semantic search engine is processed using ImageMagick software and by assessing using visual similarity based on Earth Movers Distance it is being reported as phished web page or original web page.

Corresponding Author:
Dr. S.Latha Shanmugavadivu,
Department of Electronics & Communication Engineering, Tamilnadu College of Engineering,
Coimbatore - 641656.
E-mail: latha_tce@yahoo.co.in

References

1. Agichtein E, Brill E, and Dumais S. "Improving Web Search Ranking by incorporating User Behavior Information". In Proceedings of ACM SIGIR. (2006); 19-26.
2. Ahuja L and Kumar E. "Development of expert search engine for web environment," 2nd IEEE International Conference on Information Management and Engineering (ICIME). 2010; 288–291.
3. Anthony Yingjie Fu. "Web Identity Security: Advanced Phishing Attacks and Counter Measures", Doctor of Philosophy, City Univerdity of Hong Kong; 2006.
4. Anti-Phishing Group of City University of Hong Kong; 2005.
5. Anti-Phishing Working Group, http://www.antiphishing.org; 2005.
6. ArtemChebotko and Shiyong Lu. "Querying the Semantic Web: An Efficient Approach Using Relational Databases", LAP Lambert Academic Publishing, ISBN 978-3-8383-0264-5; 2009.
7. Bollegala .F, Matsuo .Y and Ishizuka .M (2010), "A Web Search Engine-based Approach to Measure Semantic Similarity between Words," IEEE Transactions on Knowledge and Data Engineering. 99; 2010:1.
8. Ding L, Finin T, Joshi A, Peng Y, Pan R and Reddivari P. "Search on the Semantic Web" IEEE Journals on Computer. 2005; 38(10): 62-69.
9. Downs J S, Holbrook M and Cranor L F. "Behavioral Response to Phishing Risk", Proceedings of the 2nd Annual Crime Researchers. 2007; 37-44.
10. Egelman S, Cranor L F and Hong J. "You've Been Warned: An Empirical Study of the Effectiveness of Web Browser Phishing Warnings", Proceedings of the 26th annual SIGCHI Conference on Human Factors in Computing Systems. 2008; 1065-1074.
11. Elias Iosif and Alexandros Potamianos. "Unsupervised Semantic Similarity Computation using Web Search Engines". IEEE/WIC/ACM International Conference on Web Intelligence. 2007; 381–387.
12. Galina Vitkova. "The Semantic Web – great expectations". http://techenglish.wordpress.com/2011/10/31/the-semantic-web-%E2%80%93-great-expectations/ 2011.
13. Grauman K and Darrell T. "Fast Contour Matching Using Approximate Earth Mover's Distance". Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2004; 1:1220-227.
14. Mangold C. "A survey and classification of semantic search approaches". International Journal of Metadata, Semantics and Ontologies (IJMSO). 2007; 2(1):23-34.
15. Medved E, Kirda E and Kruegel C. "Visual-Similarity-Based Phishing Detection". Proceedings of the 4th International Conference on Security and Privacy in Communication Networks. 2008; 234-245.

8/24/2013