Secure web mining framework for e-commerce websites

Web Mining Framework

M.KARTHIK
Dept. of Computer science and engineering
Dr. M.G.R Education and Research Institute
Chennai.

S.SWATHI
Dept. of Computer science and Engineering
Bharathidasan University
Tiruchirappalli.

***Abstract*—This paper deals with e-commerce website and shows the usage of web mining technology to provide security for e-commerce websites. User behavior on the web is based on web mining, security and e-commerce. Customer behavior pattern is analyzed to improve e-commerce websites. Different web mining algorithm and security algorithm provide security on e-commerce websites. Web mining algorithms such as like pagerank and trust rank are used to develop web mining framework in e-commerce website. Generally web mining framework is only based on the web content mining or web usage mining. In this proposed system web mining consist of web structure mining, web content mining, decision analysis and security analysis.**

.

***Index Terms*— Web mining, E-commerce, Security.** *(key words)*

I. WEB MINING FRAMEWORK SYSTEM

Web mining is the application of data mining technique to extract and discover knowledge from the web documents. Web mining act as information service center for e-commerce, education, entertainment, advertisement, news, government etc.web mining task is generally classified as web content mining, web usage mining, web structure mining.web mining algorithm such as page rank and trust rank algorithm are used for search engine ranking. Web mining framework phases are web structure mining, web content mining, decision analysis and security analysis

II. WEB STRUCTURE MINING ANALYSIS

Larry page and sergey Brin invented page rank algorithm.Importance of page is calculated from the inbound link.Each page vote can transfer to other pages by a link.Page connected with high page rank increase rank of the page.Outgoing link spreads its vote for n pages.

A. pageRank Algorithm

Ranking became a crucial factor because people are interested to look only few top list sites on the search engine.GOOGLE follow pagerank. Calculation of pagerank algorithm work as follows .

$$PR(x) = (1-d) + d(PR(I1)/c(I1) + \ldots + PR(In)/c(In))$$

- PR(ln) –First Page “PR(l1)” to last page“PR(l2)” has self Importance.
- C(ln) –Outgoing links spreads its vote from “C(l1)”to’C(ln)”for npages.
- PR(ln)/C(ln) –pageA connected by“n” backlink pages hence share of the vote page A will be “PR(ln)/C(ln)”
  d is a damping factor in the range, 0<d<1 , Usually set to 0.85.

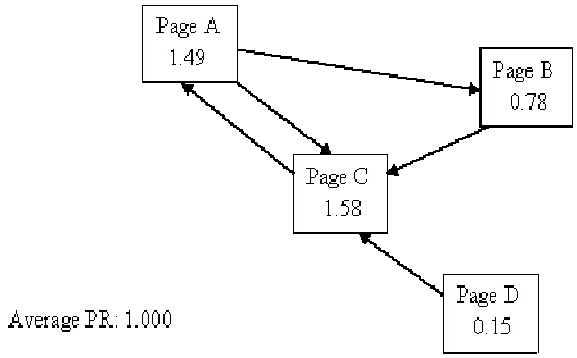

FIGURE 1: Average pagerank value value is equal to 1.00.

Acquire quality and relevant backlinks to our website can increase or decrease our search engine rankings. Backlink with relevant content increase rank with page links indexed

by Google and the page should also have a Google page rank.

B. TRUST RANK ALGORITHM

Trust rank algorithm is the procedure to rate the quality of websites. Trust rank is similar to that of pagerank. Taking the link structure measure the quality of a page.

Step1: Algorithm starts with the selection of trusted page

Step2: Trust can be transferred to other pages by linking it to them.

Step3: Trust propagates similar as pagerank.

Step4: Negative measure propagate backward which indicate measure of bad pages(spam).

Step5:For ranking algorithm both measures can be taken in to account.

III. WEB CONTENT MINING

Web content mining is the process of searching useful information from the contents of web documents. Content data is the collection of facts in a webpage designed to convey to the user. Generally content may consist of text, image, audio, video or structured record such as list and tables. Usually text mining, information retrieval and NLP techniques are applied.

In these example job categories for the computer professional is taken to identify associated skill needed for his job set. We perform clustering analysis in to two phases : Hierarchical agglomerative clustering first step to identify unique skill set characters and perform k- means clustering algorithm for modules such as User identification, Job definition, Data collection and Data analysis.

A. Hierarchical clustering

In these we have used hierarchical agglomerative clustering to identify unique skill set cluster. Bottom up strategy of placing object with own clusters and then merge the cluster into larger and larger cluster, until all of the object in the single cluster. So each iteration it merges with closest pair until all of the data is in one cluster.

Advantages

Interesting strategy to yield good result is obtained by hierarchical agglomerative which determines the number of clusters and find an initial cluster and then use iterative relocation to improve the clustering.

K-means cluster Analysis

K-means algorithm takes input parameter k and partitions a set of n objects in to k clusters

So that intracluster similarity is high but intercluster similarity is low.

Working model for K-means algorithm

Step1: First is the random selection of k objects which initially represents a cluster mean.

Step2: Remaining object is assigned to the cluster which it is most similar based down on the distance between object and cluster mean.

Step3: Compute new mean for each cluster.

Step4: Process iterates until the criterion function converges.

In this partitioning each cluster is represented by the mean value of object in the cluster. input variable k is the number of cluster and D is the data set containing n objects. Output set for k clusters method: choose arbitrary k objects from D as the initial cluster. Repeat until k object from D cluster are matched. Reassign object to the cluster which are most similar based on mean value of object in the cluster. Calculate mean value for the cluster until no change exist with the cluster.

B. User identification

Identification of the user falls in to different category such as new user who register in to the system. Existing user can logon to the system with their account. Frequently access URL is used to identify cluster users. Classifier is used to generate a profile for each cluster. Website is developed using java as front end and MYSQL as backend.

C. Data Collection & Analysis

Grouping of data for job definition is obtained by collecting the values of job title, job description and skill set required from the candidate to satisfy the job set. Data collection values are analyzed in these module to calculate skill set frequency.

Job description from various search engine is extracted and distilled to its required set using a web content data mining application. Few cluster which are of similar skill. Set is required to map specific job makes faster and quicker result based on the user preference

D. Performance analysis result set:

Information about the system can be logged for future reference to identify gap between fresher students and industry helps graduate to get an exact job and learn accordingly.

IV. DECISION ANALYSIS

Trust calculation of web page is generated from web structure mining. Trust calculation of website and the application of suitable statistical technique to analyze the evaluation result

A. Trust calculation of a website

Three trust level websites are High trust website, Moderate tust website and Un trust website.

Trust calculation Model

Trust calculation model is classified based on the opinion type weight, source type experience weight and reputation weight.

```
<Trust calculation model>
<opinions>
<opinion Type = “1”Weight-“0.2”>
<source Type =  “Experience” Weight =”0.6”/>
< Source Type =” Reputation” Weight= “0.1”/>
</source>
</opinion>
<opinion Type = “2”Weight-“0.7”>
<source Type =  “Experience” Weight =”0.6”/>
< Source Type =” Reputation” Weight= “0.1”/>
</source>
</opinion>
</Trust Calculation Mode>
```

1) Un trust website

```
<owner Name=”trustvalue”>
 <Term Name=”Un trust web sites”>
  <points>
  <Point x=”0.0”y=”1.0”/>
   <point x=”0.4”y=”0.0”/>
   </points>
   </term>
```

2) Moderate Trust Website

```
<owner Name=”trustvalue”>
<Term Name=”Moderate Trust web sites”>
 <points>
 <Point x=”0.0”y=”O.0”/>
  <point x=”0.4”y=”1.0”/>
  <point x=”0.4”y=”1.0”/>
 </points>
 </term>
```

3) High Trust Website

```
<owner Name=”trustvalue”>
<Term Name=”High Trust web sites”>
 <points>
 <point x=”0.4”y=”1.0”/>
  <point x=”1.0”y=”1.0”/>
 </points>
</term>
```

Trust value converted in to degree member function.let us consider trust value o.11.

Un trust websites:0.78.

Moderate tust websites:0.22

High trust web sites:0.00

Un trust website trust level value is none. If it is a moderate trust then the trust value is limited. Trust value of High trust website trust level is full.

B. Suitable application of statistical techniques:

Statistical analyzing of information is essential for websites.Population using random set of web data collected from the websites using descriptive statistics.It uses such as central tendency and dispersion measures.It provides summary about the sample information and observation.For ungrouped data measures are mean, median and mode.Pareto principal- states that,for many events,roughly 80% of the effects come from 20% of the causes.In this first 50% of untrust website is banned,next 25% is untrusted website followed by 12.5% untrust website.Various statistical techniques can be applied to evaluate better results.Various statistical techniques can be applied to evaluate better results.

V. SECURITY ANALYSIS

Most of the web development companies does not follow industrial standard of developing and hosting the websites.Customer using a website is unaware of whether it is a trusted website or untrusted website. In this paper security on e-commerce website is provided with trust path intermediate algorithm,false hit database algorithm and similarity search.Multistep processing is carried on nearest neighbor and similarity search. C-AMNC- used to reduce the size of false hit database.Query is authenticated and server maintains the database of trusted user details to reduce hang or lag in server. Provides accurate data with NN result-set. Security analysis module for providing security on e-commerce web sites.Module 1:Authentication; Module 2: Query processing; Module 3: Similarity Search; Module 4: False hit reduction.These techniques are used to provide security for e-commerce websites.

A. Module 1: Authentication

Member user access search facility in the job site. Admin updates the database from the false hit.

B. Module2: Query Processing

Server and the user interaction take place in the module .Client post the query and server respond it back from the criteria.

C. Module3: Similarity Search

Retrieve of relevant information from the database based with similar key word.

D. Module4:False hit reduction

Admin constantly checks the false hit record. He then finally post necessary response with search database for future verification

1) *Case 1:*

Search keyword is updated if it is not found in database

2) *Case2:*

If the search keyword is already present in database then admin post necessary response to the search database for future verification.

3) *Case3:*

User can access necessary search details from database. Admin checks false hit data and update database.

Administrator has a set of privileges to modify or update website based on user details.

FIGURE2:Welcome Administrator.

Recommendation posted by the user

User post his likes of interested job list based on his professional, salary, expectations etc. This helps to know the necessity of user for particular search.

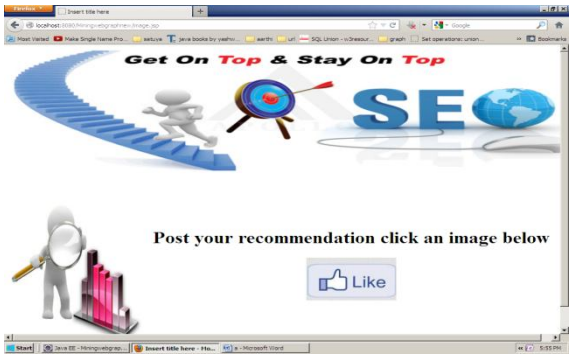

FIGURE3:Recommendation post on User like's.

VI. CONCLUSION

Web mining frame work consist of four different phases such as web structure mining analysis, Web cont mining, Decision Analysis, Security Analysis. Web structure mining uses pagerank and trust rank algorithm. Web content mining uses hierarchical clustering and K- means clustering algorithm.Decision analysis uses trust calculation of website and application of suitable statistical techniques.Finally Security module provides security to website.Security analysis perform using trust path intermediaries building algorithm,false hit database algorithm and nearest neighbor algorithm to provide security on e-commerce websites.

REFERENCE

[1]R.Manjusha and R.Ramachandran "Web Mining Framework for Security in E-commerce" IEEE-International Conference on Recent Trends in Information Technology, ICRTIT 2011,PP.1043-1048, June 3-5, 2011.

[2]Abdullah H. Wahbeh, Qasem A. Al-Radaideh, Mohammed N. Al-Kabi, and Emad M. Al-Shawakfa" A Comparison Study between Data Mining Tools over some Classification Methods" (IJACSA) International Journal of Advanced Computer Science and Applications,Special Issue on Artificial Intelligence.,PP.18-26,www.ijacsa.thesai.org..

[3]Sonal Tiwari "A Web Usage Mining Framework for Business Intelligence" International Journal of Electronics Communication and Computer Technology (IJECCT) Volume 1 Issue 1 | September 2011.

[4]Ayman Farahat ,Thomas Lofaro, Joel C. Miller Gregoryrae , and Lesley A. Ward " Authority rankings from HITS,PAGERANK,AND SALSA: Existence, Uniqueness, and effect of initialization. SIAM J. SCI. COMPUT. 2006 Society for Industrial and Applied Mathematics, Vol. 27, No. 4, pp. 1181–1201.

[5]Egger,F.N.(2003),From Interactions to Transactions: Designing theTrust Experience for Business – to – Consumer Electroniv Commerce. PhD Thesis, Eindhoven Univesity of Techonology(The Netherlands), ISBN 90-386-1778-X.

[6]Banatus Soiraya ,Anirach Mingkhwan & Choochart Haruechaiyasak " E-commerce Web Site trust assessment Based on Text Analysis" International Journal of Business and Information, Volume 3, Number 1, PP.86-114,June 2008.

[7]Banatus Soiraya ,Anirach Mingkhwan & Choochart Haruechaiyasak "An Analysis of Visual and Presentation Factors Influencing the Design of E-Commerce Web Sites" International conference on web intelligence and intelligent agent technology, PP.525-528 ,2008, IEEE/WIC/ACM.

[8]Chuck Litecky, Andrew Aken, Altaf Ahmad, and H. James Nelson,Southern Illinois University, Carbondale,Mining computing jobs,January/February 2010 IEEE SOFTWARE

[9]Yacine Ati fUnited Arab Emirates University, Building Trust in E-Commerce, IEEE INTERNET COMPUTING, January. February 2002

[10]Bing Liu, Robert Grossman, and Yanhong Zhai, University of Illinois at Chicago, Mining Web Pages for Data Records. Published by the IEEE Computer Society November/December 2004.

[11]J.Han and M.Kamber Data Mining: Concepts and Techniques, Morgan Kaurmann Publishers,2000.

[12]Tao, Y., Yi, K., Sheng, C., Kalnis, P. Quality and Efficiency in High Dimensional Nearest Neighbor Search. SIGMOD, 2009.

[13]Sankar K. Pal,Fellow,IEEE,Varun Talwar,Student Member,IEEE,and Pabitra Mitra, Student Member, IEEE, Web Mining in Soft Computing Framework:Relevance, State of the Art and Future Directions, IEEE TRANSACTIONS ON NEURAL NETWORKS, VOL. 13, NO. 5, SEPTEMBER 2002

[14]S. Chakrabarti, B. Dom, D. Gibson, J. Kleinberg, P. Raghavan, and S. Rajagopalan. Automatic resource compilation by analyzing hyperlink structure and associated text In Proceedings of the Seventh International World Wide Web Conference, 1998.

[15]S. Brin, R. Motwani, L. Page, and T. Winograd. What can you do with a web in your pocket. In Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, 1998.

[16]S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In Proceedings of the Seventh International World Wide Web Conference, 1998.

[17]Lawrence Page, Sergey Brin, Rajeev Motwani and TerryWinograd The PageRank Citation Ranking: Bringing Order to the Web(1998).