Web Mining Framework for E-Commerce Security

Fourcan Karim Mazumder

Abstract— Internet sales are increasing rapidly as consumers take advantage of lower prices offer by wholesalers retailing their products. But until now e-commerce is not fully effective for people because of security problems are harming e-commerce business. Web mining algorithms and security algorithms provide security on e-commerce websites. Web mining algorithms are used to develop web mining framework in e-commerce website. This paper describes the usage of web mining framework to provide security for e-commerce websites.

Index Terms— Web mining, security, web structure mining, web content mining, pagerank algorithm, trustrank algorithm, trust calculation model.

1 Introduction

Electronic commerce or e-commerce refers to a wide range of online business activities for products and services. E-commerce is usually associated with buying and selling over the internet or con-ducting any transaction connecting the transfer of ownership or rights to use goods or services through a computer-mediated network (Andam, 2003). E-commerce has an important impact on business costs and productivity. This increases competition and modernization, which are likely to boost overall economic efficiency (Hoq et al., 2005). Any businesses adopt an e-commerce business plan because it offers the owner greater flexibility in terms of operating location and hours. That is, e-commerce may present an individual with the chance to be a lifestyle entrepreneur and locate the business where the entrepreneur wants to live (Barkley et al., 2007). Web mining is the term of applying data mining techniques to automatically discover and extract valuable information from the web documents and services. The unstructured feature of web data generates more complexity of web mining (Wang, 2000). Here web mining framework is using to provide security for e-commerce website.

2 Web Mining Framework System

The E-commerce encounters challenges in terms of high security risks due to open nature of the internet and increasing technical knowledge, which enable the criminals to build up more sophisticated means to perform illegal attacks. Security is defined as “the protection of data against accidental or intentional disclosure to unauthorized persons or unauthorized alterations or destruction. High security’s risks belong to E-commerce are laws and regulations fault, systems and technology flaw and the internet (Jebur et al., 2012). Another way we can say that, E-commerce Security is a part of the Information Security framework and is specially applied to the components that affect e-commerce that include Computer Security, Data security and other wider areas of the Information Security framework (Niranjanamurthy et al., 2013). Web mining approach is classified into three areas: Web content mining, Web structure mining, and Web usage mining (Wang, 2000). Web mining framework is using to provide security for e-commerce which has different phases such as web structure mining, web content mining, decision analysis and security analysis (Karthik et al, 2013).

2.1 Web Structure Mining Analysis

It is the method by which we discover the model of link structure of the web pages. We record the links, generate the information such as the similarity and relations among them by taking the advantage of hyperlink topology. The goal of web structure mining is to produce structured summary about the website and web page. It tries to find out the link structure of hyper links at inter document level (Nithya, 2013). Larry page and Sergey Brin invented the concept of page rank algorithm. Necessity of page is calculated from the inbound link. Each page vote can move to other pages by a link. Page attached with high page rank increase rank of the page. Outgoing link increases its vote for n pages.

A. Pagerank Algorithm

Ranking became a critical factor because people are interested to look only few top list sites on the search engine. Google pursue pagerank. Computation of pagerank algorithm work as follows:

PR(x) = (1-d) + d(PR(I1)/c(I1)+ … + PR(In)/c(In))

- PR(ln) –First Page “PR(l1)” to last page “PR(l2)” has self necessity.
- C(ln) –Outgoing links increases its vote from “C(l1)” to “C(ln)”for n pages.
- PR(ln)/C(ln) – page A linked by “n” back link 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.

Obtain quality and relevant back links to our website can in-

- *Fourcan Karim Mazumder is currently working as a senior lecturer in Computer Science and Engineering Department, International University of Business Agriculture and Technology (IUBAT), Dhaka, Bangladesh, PH-008801738332159. E-mail: fk.mazumder@iubat.edu* crease or decrease our search engine rankings. Back link with related content increase rank with page links indexed by google and the page should also have a google page rank (Karthik et al, 2013).

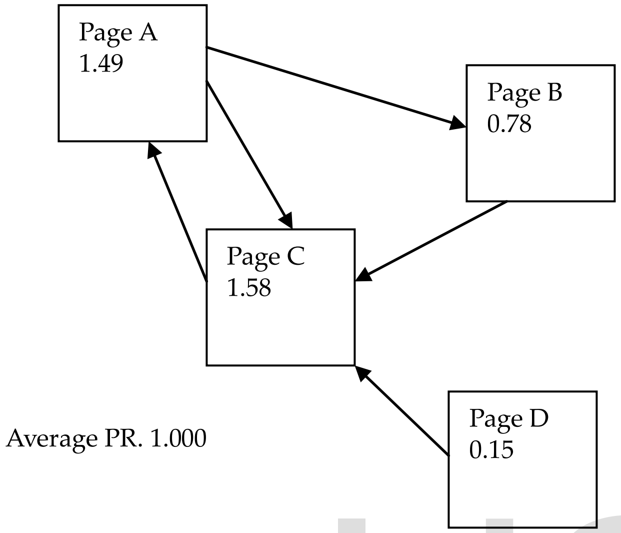

Figure: Average pagerank value is equal to 1.00.

B. Trustrank Algorithm
An effective trust assessment tool must be able to identify and decide the trustworthiness level of the web sites correctly (Soiraya et al, 2007). Trustrank algorithm is the formula to rate the quality of websites. Trustrank is analogous to that of pagerank. Taking the link structure evaluate the quality of a page.
Step1: Algorithm begins with the selection of trusted page
Step2: Trust can be moved to other pages by linking it to them.
Step3: Trust propagates analogous as pagerank.
Step4: Negative measure transmit backward which indicate measure of bad pages.
Step5: For ranking algorithm both determines can be taken in to account (Karthik et al, 2013).

2.2 Web Content Mining

Web content mining is the method of extracting useful information from the content of a website. It includes extraction of structured data or information from web pages, recognition, match and integration of semantically similar data, opinion extraction from online sources, and idea hierarchy, ontology, or knowledge integration (Herrouz et al, 2013). Content data is the set of facts in a webpage designed to convey to the user. Usually content may consist of text, image, audio, video or structured record such as list and tables

In these example job categories for the computer professional is taken to recognize associated skill needed for his job set. We perform clustering analysis in to two stages: Hierarchical agglomerative clustering first step to discover 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 recognize unique skill set cluster. Bottom up strategy of placing object with own clusters and then combine the cluster into larger and larger cluster, until all of the object in the single cluster. So each iteration it merges with nearby pair until all of the data is in one cluster.

Advantages
Attractive strategy to yield good result is obtained by hierarchical agglomerative which determines the number of clusters and find an early cluster and then use iterative relocation to improve the clustering (Karthik et al, 2013).

K-Means Cluster Analysis
This section describes the original kmeans clustering algorithm. One of the most popular clustering methods is k-means clustering algorithm. K-means clustering is a method of cluster analysis which aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean. Clustering is defined as grouping similar objects either physical or abstract. Each created group is called a cluster (Kaur et al, 2013). So that intracluster match is high but intercluster match is low.

Working Model for K-Means Algorithm
Step1: First is the random selection of k objects which firstly represents a cluster mean.
Step2: Remaining object is assigned to the cluster which it is mainly similar based down on the distance between object and cluster mean.
Step3: Calculate new mean for each cluster.
Step4: Process iterates until the principle function converges.

In this separation 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 holding n objects. Output set for k clusters method: choose random k objects from D as the initial cluster. Reiterate until k object from D cluster are matched. Reassign object to the cluster which are most analogous based on mean value of object in the cluster. Compute mean value for the cluster until no change exist with the cluster (Karthik et al, 2013).

B. User Identification
Recognition of the user falls in to different category such as new user who register in to the system. Existing member can logon to the system with their account. Regularly access URL is used to identify cluster users. Classifier is used to produce a profile for each cluster. Java and MySql is using to develop website.

C. Data Collection & Analysis

Grouping of data for job definition is obtained by gathering 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 estimate skill set frequency. Job description from different search engine is extracted and distilled to its required set using a web content data mining application. Only some cluster which are of similar skill. Set is needed 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 accurate job and learn accordingly (Karthik et al, 2013).

2.3 Decision Analysis

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

A. Trust Calculation of a Website
There are three trust level websites are High trust website, Moderate trust website and Un trust website.

Trust Calculation Model
Trust calculation model is categorized based on the opinion type weight, source type experience weight and reputation weight (Karthik et al, 2013).

```
<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 transformed in to degree member function. Let us consider trust value 0.11.
Un trust websites: 0.78.
Moderate trust websites: 0.22
High trust websites: 0.00
The trust level value of un trust website is none. The trust value is limited if it is a moderate trust. High trust website trust level value is full (Karthik et al, 2013).

B. Suitable Application of Statistical Techniques
Statistical analyzing of information is important for websites. Population using random set of web data gathered from the websites using descriptive statistics. It uses such as dispersion measures and central tendency. It gives summary about the sample information and observation. For ungrouped data determines are mean, median and mode. Pareto principal- says that, for many events, roughly 80% of the effects come from 20% of the causes. In this first 50% of untrust website is prohibited, next 25% is untrusted website followed by 12.5% untrust website. A variety of statistical techniques can be applied to evaluate better results (Karthik et al, 2013).

2.4 Security Analysis

Maximum web development companies does not follow industrial standard of developing and hosting the websites. Customer using a website is unconscious of whether it is a trusted website or untrusted website. In this paper security on e-commerce website is presented 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 decrease the size of false hit database. Query is authenticated and server maintains the database of trusted user details to decrease hang or lag in server provides exact data with NN result-set. Security analysis module for giving security on ecommerce web sites: Module 1: Authentication; Module 2: Query processing; Module 3: Similarity Search; Module 4: False hit reduction. These methods are used to provide security for e-commerce websites (Karthik et al, 2013).

1. Authentication Module
This module gives wide area for accessing capability to the user perspective job where the user puts a request for the pag-

es or websites. It restricts accessing where it does not confirm using authentication. This module plays a significant role in security analysis.
2. Query Processing Module
In Query processing module the communication take place between the user and server for the requested query. It is appropriate for large processing of query related request.
3. Similarity Search Module
This module searches equivalent alike factors among information in data mining. It probably searches basic similar keywords in the form of exceptional attributes. The most featuring application data retrieval is completed.
4. False Hit Reduction Module
In these module the administrator verifies the different records when there is a error is triggered or hit occurs. The admin focuses and does checking of the results of response and gives measure for future work for reducing the false hit which helps in establishment of accurate results of fact for search result in database (David et al, 2014).

i) Case 1:
Search keyword is updated if it is not available in database
ii) Case2:
If the search keyword is already available in database then admin post necessary response to the search database for future verification.
iii) Case3:
User can access essential search details from database. Admin verifies false hit data and update database. Administrator has a set of rights to modify or update website based on user details.

Recommendation Posted by the User
User post his likes of preferred job list based on his professional, salary, expectations etc. This helps to know the need of user for particular search (Karthik et al, 2013).

3 Conclusion

Web mining frame work contains of four different phases such as web structure mining analysis, web content mining, decision analysis, security analysis. Pagerank and trust rank algorithms are using in web structure mining. Hierarchical clustering and K- means clustering algorithms are using in web content mining. Trust calculation of website and application of suitable statistical techniques are using in decision analysis. Finally Security module serves security to website. Security analysis perform using trust path intermediaries building algorithm, false hit database algorithm and nearest neighbor algorithm to present security on e-commerce websites.

References

[1] Andam, Z. R. (2003), '*e-Commerce and e-Business', e-Asean Task Force UNDP-APDIP*, Available from: < http:// www. kau. edu. sa /Files /830 /Files / 61164_Ecommerce% 20and% 20E%20 Business.pdf >

[2] Hoq, Z., Kamal, M. S., Chowdhury, E. H. (2005), 'The Economic Impact Of E-Commerce', *BRAC University Journal*, Vol. 2, No. 2, pp. 49-56.

[3] Barkley, D. L., Markley, D. M., Lamie, R. D. (2007), *'E-Commerce as a Business Strategy: Lessons Learned From Case Studies of Rural and Small Town Businesses', UCED Working Paper 10-2007-01, EDA University Center for Economic Development.* Available from: < http:// www.joe.org/ joe/2011december / pdf/JOE_v49_6rb4.pdf >

[4] Wang, Y. (2000), 'Web Mining and Knowledge Discovery of Usage Patterns', Available from: < https: //cs.uwaterloo.ca/ ~tozsu/ courses/ cs748t/ surveys/ wang.pdf >

[5] Jebur, H., Gheysari H., Roghanian, P. (2012), 'E-Commerce Reality and Controversial Issue', *International Journal of Fundamental Psychology and Social Sciences*, Vol. 2, No. 4, pp. 74-79.

[6] Niranjanamurthy, M., Chahar, D. D. (2013), 'The study of E-Commerce Security Issues and Solutions', *International Journal of Advanced Research in Computer and Communication Engineering,* Vol. 2, No. 7.

[7] Karthik, M., Swathi, S. (2013), 'Secure web mining framework for e-commerce websites', *International Journal of Computer Trends and Technology*, Vol. 4, No. 5, pp. 1042-1046.

[8] Nithya, T. (2013), 'Link Analysis Algorithm for Web Structure Mining', *International Journal of Advanced Research in Computer and Communication Engineering*, Vol. 2, No. 8, pp. 2950-2954.

[9] Soiraya, B., Mingkhwan, A., Haruechaiyasak, C. (2007), 'eCommerce Web-Site Trust Assessment Framework Based on Web Mining Approach', paper presented in the *Proceedings of the 24th South East Asia Regional Computer Conference, Bangkok, Thailand.*

[10] Herrouz, A., Khentout, C., Djoudi, M. (2013), 'Overview of Web Content Mining Tools', *The International Journal of Engineering And Science*, Vol. 2, No. 6.

[11] Kaur, M., Kaur, N. (2013), 'Web Document Clustering Approaches Using K-Means Algorithm', *International Journal of Advanced Research in Computer Science and Software Engineering*, Vol. 3, No. 5, pp. 861-864.

[12] David, R. P., Karnewar, J. S. (2014), 'Security in Web Data Mining Framework', *International Journal of Engineering Science and Technology*, Vol. 6, No. 5S, pp. 10-15.