DISCOVERY OF PROCESS MODEL USING CLICK STREAM ANALYSIS IN WEB MINING

Dileep Kumar.Padidem*, 1Dr.C.Nalini 2

1Research Scholar, Bharath University, Chennai.
2Professor, Department of CSE, Bharath University, Chennai.

Email: padidemdileep@gmail.com

Received on 04-08-2016

Abstract:

Web mining uses the data mining techniques to automatically discover and extract information (Knowledge) from Web documents and services. The method of general web access pattern is extracted and analyzed using knowledge discovery techniques to understand the patterns. This usage mining can be done from the click stream data of the website by the users. Sequence of tasks are observed through click stream analysis and a business process model may be discovered in web structure mining. This paper suggests a plan and proposal about how to consider the user behavior in E commerce sites as process and to discover a process model which enhances business intelligence.

Key words: Process mining, Process Centric view, Click stream analysis, Process model, Web mining.

1. Introduction

Web mining is to extract knowledge from web data, i.e. web content, web structure, and web usage data using the data mining techniques. Web mining uses the data mining techniques to automatically discover and extract information (Knowledge) from Web documents and services. Web mining discovers three general classes of information:

- Web activity, tracking of server logs and Web browser activity.
- Web graph, links between pages, people and other data.
- Web content, for the data found on Web pages and inside of documents.

There are many other end uses for web mining results. A partial list consist of: Competitive intelligence, Business intelligence, Pricing analysis, Events, Product data, Popularity, Reputation

When extract Web content information using web mining, there are four typical steps.

1. Collect – To get the content from the Web
2. Parse – To extract usable data from formatted data (HTML, PDF, etc)

3. Analyze – To tokenize, classify, cluster, sort, rate, filter, etc.

4. Produce – To turn the results of analysis into something useful (search index, report, etc)

The dynamic nature of web creates large massive volumes of information in structured and semi structured nature. Discovering knowledge from these information are said to be web mining and this web mining process can be focused on three different categories of the information present. They are Web Structure Mining, Web Content Mining, & Web Usage Mining. Web mining is applied in different domains and researches are carried out in concentration in all the three categories with different methodologies.

**Web Structure Mining Analysis**

The model of link structure of the web pages is discovered through this method. The links are recorded to generate the information such as the similarity and relations among them by taking the advantage of hyperlink topology. The aim of web structure mining is to create structured summary about the website and web page. It tries to discover the link structure of hyper links at inter document level (Nithya, 2013) [3].

**Web Content Mining**

The method of extracting useful information from the content of a website is Web content mining. It includes extraction of structured data from web pages, detection, match and combination of semantically similar data, opinion extraction from several online sources, and ontology, idea hierarchy, or knowledge incorporation (Herrouz et al, 2013) [1] Content data is the set of facts in a webpage designed to express to the end user. Usually content may consist of text, image, audio, video or structured record such as list and tables.

**Web Usage Mining**

The method of general web access pattern is extracted and analyzed using knowledge discovery techniques to understand the patterns. This usage mining can be done from the click stream data of the website by the users.

2. **Web Mining Versus Data Mining**

There are three main differences to be considered when comparing web mining with traditional data mining system

1. **Scale** – In data mining, processing one million records from a database would be large job. In web mining, even though ten million pages wouldn’t be a large number.
2. **Access** – When doing data mining of corporate information, the data is private and often requires access rights to read. The data is public and rarely requires access rights in web mining. Also web mining has additional restriction, due to the implicit agreement with webmasters regarding automated (non-user) access to this data. This implicit agreement is that a webmaster allows crawlers access to useful data on the website, and in return the crawler.

(a) Promises not to overload the site, and

(b) Has the potential to drive more traffic to the website once the search index is published. With web mining, there often is no such index, which means the crawler has to be extra careful/polite during the crawling process, to avoid causing any problems for the webmaster.

3. **Structure** – A traditional data mining task gets information from a database, which provides some level of explicit structure. A typical web mining task is processing unstructured or semi-structured data from web pages.

3. **Web Mining To Process Mining**

Two different approaches were taken in initially defining web mining. First was a “process-centric view,” which defined web mining as a sequence of tasks (Etzioni 1996)[6]. Second was a “data-centric view,” which defined web mining in terms of the types of web data that was being used in the mining process (Cooley, Srivastava, and Mobasher 1997)[7]. There are many works on web mining in terms of data-centric view, and this paper considers its study on web mining in terms of process centric view which defined web mining as a sequence of tasks. This paper suggest its plan and view to consider the Click stream data in ecommerce websites as sequence of tasks and framing different optimum business models.

3.1 **Clickstream Data**

Click stream data are the electronic record of a user’s behavior on the web sites. This data trace the path a visitor takes while navigating the Web and this path reflects Choices, often very big in number, made by the user both within and across websites. For example, The data set of a click stream might include a record of every website and every page Click stream data are defined as the electronic record of a user’s activity on the Internet. Thus, the data trace the path a visitor takes while navigating the Web. This path reflects Choices, often very large in number, made by the user both within and across websites. For example, a click stream dataset might include a record of every website and every page visited, the time user spent on each site and the order the sites and pages were visited. An important unit of observation in clicks...
stream data is the page visit—the recording of a user’s visit to a given website page. Technically, the assembly of a “page view” from the user’s perspective can involve numerous “hits” to the Web server. These reflect the downloading of various page elements before they are assembled in the User’s Internet browser window. Click stream data is automatically aggregated from hits to page views but in some cases (e.g., raw server log files), the analyst may need to perform this step. Raw click stream data can be captured by server log files maintained by a website can record all the requests and information transferred between the server and the user’s computer system. The data are collected from a single website and they are known as “site-centric.” Site-centric click streams can provide very detailed records of users’ behavior that is about their navigation and interaction with a given site. Click-stream data provides the opportunity for a detailed look at the decision making process itself, and knowledge extracted from it can be used for optimizing, influencing the process, etc. (Ong and Keong 2003) Underhill (2000) has conclusively proven the value of process information in understanding users’ behavior in traditional sites. Research needs to be carried out in (1) extracting process models from usage data, (2) understanding how different parts of the process model impact various web metrics of interest, and (3) how the process models change in response to various changes that are made, i.e. changing stimuli to the user.

4. Process Mining

The process mining targets the automatic discovery of information from an event log. This discovered information can be used to create a new system that support the execution of business processes or as a feedback tool that helps in auditing, analyzing and improving already enacted business processes. Process Mining is a recent research discipline, which is used to discover knowledge from event logs are the processes that are extracted from the available information systems like transaction log file, Ms-Excel spread sheet or a normal data base table

The three powerful areas of process mining applications are (Fig.1):

![Fig.1. Three types of process mining: discovery, conformance and enhancement.](image-url)
Process model discovery constructing complete and compact process models able to reproduce the observed behavior.

Conformance checking reflecting on the observed reality, i.e. checking whether the modeled behavior matches the observed behavior and

Process model extension projection of information extracted from the logs onto the model, to make the knowledge explicit and facilitate better understanding

The idea of process mining is to discover the process model, monitor and enhancement of real processes by getting knowledge from event logs readily available in the systems. In the three powerful areas of process mining application, many organizations are interested in information about conformance of its processes to rules that should be observed. Among the three areas of process mining this paper makes an observation and a case study on conformance checking that is checking how far the observed behavior matches with the modeled behavior of the workflow model in an organization.

W. M. P. van der Aalst [4] stated that there is currently a missing link between business processes and the real processes with information systems. Process mining has arisen as a new scientific discipline to provide a link between process models and event data [4]. Simeonova [5] defined process mining as techniques that help to find, screen and enhance genuine procedures by concentrating learning from event logs. Data is gathered from assorted types of systems and examined to identify deviations from standard processes and see where the bottlenecks are. Process mining is based on fact-based data and starts with an analysis of data, followed by the creation of a process model.

The techniques of data mining and knowledge discovery could be applied efficiently on web sites or e-commerce sites. This specific application of data mining on e-commerce web pages called Web Mining and it has taken much attention of researches. A new research area was derived from Web Mining for guiding the solutions to its specific requirements. Some researchers has worked on mining the contents of a web site in web content mining, mean while others has decided to study the structure of a web site in web structure mining or analyze the usage of a web site (web usage mining).

Now a days, web usage mining has attracted much attention from researchers and e-business professionals, it offers many benefits to an e-commerce website such as:

- Targeting customers based on usage behaviour or profile (personalization).
- Adjusting web content and structure based on page access pattern of users (adaptive web site).
- Improving the service quality and delivery to the end user (cross-selling, up-selling).
– Enhancing web server system performance based on the web traffic analysis.

– Identifying significant area of the web site.

The data needed to accomplish such tasks is derived normally from a Web server log file – almost all e-commerce applications are Web based. Clickstream files are generated in order to represent information that is specific to each Web access attempt. Basically, a clickstream contains, among other things, the IP address of origin site, the access time, the referring site, the URL of the target site (i.e. the web page or object accessed), the browser method, and the protocol that was used. Nowadays, several commercial tools are available for clickstream analysis and many more are accessible free on the internet. The web logs goes through four stages. In the first stage, the data is filtered to remove irrelevant information and a relational database is created containing the meaningful remaining data. This database facilitates information extraction and data summarization based on individual attributes like user, resource, user’s locality, day, etc.

In the second stage, a data cube is constructed using the available dimensions. On-line analytical processing (OLAP) is used in the third stage to drill-down, roll-up, slice and dice in the web log data cube. Finally, in the fourth stage, data mining techniques are put to use with the data cube to predict, classify, and discoverer interesting correlations.

Therefore, process mining allows us to answer the following type of questions through click stream analysis

1. What is the buying process that customers follow?

   (a) Find distinctive groups of customers like the bargain shopper, the surgical shopper, the enthusiastic shopper and the power shopper.

   (b) Detect differences in behavior and buying pattern between the groups.

2. What kind of community relationship and personal attention the customers need?

3. Where do different type of customers deviate from an intention of purchasing?

So many research results has been produced using web analytics study which observes customers behavior through click stream behavior and market basket analysis which will not provide critical path of site visitors behavior and abstracted view of underlying customer processes. This paper suggests a plan of applying Business Process Methodologies (BPM) to event logs of ecommerce websites to study the challenges and potential benefits of such an approach.

1. Extracting process models from usage data. (Transforming web clicks into tasks, and corresponding URL as high level tasks for analytics using BPM tools.)
2. understanding how different parts of the process model impact various web metrics of interest,
3. how the process models change in response to various changes that are made, i.e. changing stimuli to the user
4. The difference of work flow model (originally planned) and the real time model can be studied.
5. And results can be obtained on different underlying process path through customer critical behavior, different process model can be extracted, deviated process model from planned work flow model can observed, and optimum work flow model can be reached to enhance the business process.

5. Conclusion
This paper sets a plan to study the web usage mining in terms of process centric view and to consider the usage pattern of ecommerce websites as sequence of process. Thorough this plan business models can be analyzed in terms of different process models and the ideas proposed in this paper are under study and implementation is future scope.

6. References

Corresponding Author:
Dileep Kumar.Padidem*,
Email: padidemdileep@gmail.com