Mining web content usage patterns of electronic commerce transactions for enhanced customer services

Sylvanus A. Ehikioya1 | Jinbo Zeng2

1Department of Computer Science, Baze University, Abuja, Nigeria
2Department of Computer Science, University of Manitoba, Winnipeg, Manitoba, Canada

Correspondence
Sylvanus A. Ehikioya, Department of Computer Science, Baze University, Abuja, Nigeria.
Email: ehikioya@gmail.com

Abstract
A successful business intelligence solution can help organizations improve the quality and speed of their decision-making processes by analyzing the consolidated information collected from their websites. Using the current Web server log standard, which indicates only the locations of served Web pages, may lead to inaccurate business analysis for data driven and frequently updated static content Web pages. A properly defined Web content usage data warehouse that captures both dynamic and static contents of web pages provides rich data source for discovering interesting business rules among users’ activities. This paper demonstrates the simplicity of data extract, transform and load procedures to import raw Web content usage log to various data models for data analysis, reporting and data mining tools. In this paper, we use two data mining techniques (expected maximization and the prefixspan algorithms) for visitor grouping and path analysis to find interesting patterns in Web content usage log data, an important component of e-commerce Web sites traffic analysis. Visitor grouping uses data clustering to group visitors/sessions with similar selected attributes’ value. Path analysis mines common visiting path sequence. The ultimate goal is to enable the online merchant provide enhanced and differentiated marketing services generally to its existing and potential customers and tell what items customers are interested in instead of ambiguous URLs.

KEYWORDS
data Mining, data warehouse, electronic commerce, web content usage data

1 | INTRODUCTION

One of the benefits of business intelligent systems is the effectiveness of business operations to continuously drive improvements. This is done through studies/analysis of trends or patterns from an organization’s operational data. Some examples of business intelligence applications include discovering valuable customers and maintenance of their unwavering loyalty by clustering and association rules, and discover how to conduct business more efficiently. These are covered within the research domain of data mining,1,2 well documented in the literature.

The initial research for this paper was carried out while Dr. Ehikioya was with the University of Manitoba, Winnipeg, Canada.
Web data mining is the application of data mining techniques to the large Web usage log database to discover patterns and analyze useful information from Web data, which includes Web content data and Web usage data.\textsuperscript{3-9} The Web data mining process analyzes large volumes of prepared Web log data in order to help organizations understand customers’ behaviors, and the effectiveness of their current Web site structure and promotional campaigns. Web data mining is characterized by the following\textsuperscript{10}:

- The data are mostly unstructured.
- The data on the web are dynamic and the volume keeps growing.
- The web contains any data type, such as structured and unstructured data.
- The web contains a variety of information; often with redundant pages.
- The web contains vast amounts of linked information.
- Noisy data

In other words, the huge dynamic volume of source dataset and large variety of data formats differentiate web data mining from data mining.

The traditional Web data mining approaches include two domain dependent phases; the raw Web access log pre-processing phase, and the offline business data integration phase. These two phases are often tedious. With the growing popularity of e-commerce and the characteristic feature of e-commerce transaction that could involve the simultaneous execution of many processes on different computers (possibly at different locations), the volume of consumer usage data is also growing phenomenally and the huge volume of data needs to be intelligently integrated to be able to generate business actionable meaning from it. Web data integration for e-commerce is well documented in the literature\textsuperscript{11-16} and will not be repeated here.

The raw Web access log file pre-processing phase converts raw Web access records into some formats that can be used by data mining algorithms to discover meaningful patterns. This phase includes, at least, two tasks: data cleaning and session identification. Data cleaning removes bad requests and unwanted requests, such as image file requests, and so forth. Session identification is the process of associating session ids and page references, generating users’ session navigation paths in each session from the cleaned log. The user session database is then integrated with online transaction processing (OLTP) data as source data for the knowledge discovery phase. The success of a business intelligent solution for e-commerce systems largely depends on how successful associating universal resource locators (URLs) in Web access log with OLTP data. Successful association requires associating request parameters embedded in URL with the correct data attributes and correct values at the moment of the user request.

The Web offers a critical channel of promoting a company’s products and services because e-commerce sites are important sales channels that enable online transactions real-time. E-commerce provides a critical path to future success of any commercial organization because it offers huge opportunities and market outlets worldwide. Therefore, companies must realize that winning in the e-commerce domain means more than mere simple sales transactions, but the careful integration of the appropriate strategies is the key to improving competitive power. One of these strategies is to analyze historical data from e-commerce activities. Therefore, analyzing the data from the activities carried out by visitors to these websites is vital. The goal is to discover actionable non-obvious yet useful information from massive sources of Web data, using unstructured, semi-structured and heterogeneous data, such as textual information and hyperlink structure, and generate information to improve the quality of services and products offering by e-commerce merchants. This can be achieved using web mining techniques, a breakthrough technology that can be used to gather information and build models for predicting customer [purchasing] behaviors/decisions with remarkable accuracy.

This study is motivated by the need for businesses to gain competitive advantage over their competition, gain customer loyalty, and enhance their profitability. Therefore, businesses need to have detailed information of the activities of online customers on their websites in order to have actionable and decisive knowledge. Adoption of adequate and appropriate marketing strategies of e-commerce enterprises has practical value on their competitiveness because the resolution of how enterprises can accurately match the characteristics of its products or services with consumers is easily achievable. In addition, it also helps to maintain a relatively stable customer group size and structure. Web data mining technique is applied to e-commerce data set to know the browsing behavior of customers, to determine the success of marketing efforts, to improve the design of e-commerce web site, to monitor and optimize website performance, and to provide personalized services. Thus, e-commerce enterprises need strong web analytics tools and skills to discover and extract actionable, useful and interesting information from the Web content usage data log. For example, information about the purchasing behavior of visitors can be derived from the e-commerce sites by analyzing sequence analysis of pages visited by a visitor during
a given session (web clicks) recording the translation in sales transactions. The main purpose is to show a connection between the actions of online visitors and purchases made on a website. Furthermore, we can perform shopping basket analysis which involves determining products that are purchased together by customers, and thus helps to improve a website’s performance by customizing it. The ultimate goal in analyzing an online merchant’s Web content usage data log is to increase revenue/profitability and customer satisfaction through careful analysis of visitor interaction at the website.

The method for analysis presently is to cluster Web transactions mainly on request URLs and this approach is inadequate to meet the ever growing needs of online merchants. For example, item-based patterns and rules can be missed when the transactions are not considered to be similar during data mining. By simply applying data mining technologies on transaction URLs is unable to reveal the relationships among items presented in the transactions in dynamic Web content scenario. Furthermore, the inadequacy in reporting the effectiveness of specific marketing and merchandising efforts is a limitation.

To examine in depth Web log data for business analysis, we use data mining to uncover patterns and relationships that cannot be seen in simple summary statistics by applying probabilistic and association discovery techniques through a three-phase processing step: preprocessing, patterns discovery, and patterns analysis. In the data preprocessing step, the raw data from Web server log is cleansed and transformed into the formats [and database schemas] that can be used by the data mining algorithms necessary for pattern analysis. The data mining results are subject to impactful presentation tools for easy understanding by business decision-makers.

This paper uses the integrated Web users’ personal profiles, business transactional data, and users’ visit log enterprise data for the discovery and analysis of the relationship between product sales and customers’ behavior by mining the integrated data using two selected algorithms, expected-maximization17 and the prefixspan18 algorithms. The results of our analysis provide valuable insights into customers’ behaviors and content effectiveness, thus helping online merchants to build a more effective and profitable e-commerce system. For example, online merchants can easily derive important clues for analyzing sales data and visitors’ behavior data of specific marketing campaigns for business decision.

This paper demonstrates the ability to discover business-oriented intelligence from granularized Web content usage data logged attributes using two existing data mining algorithms. The contributions of this paper are:

1. Examines the discovery and analysis of the relationship between product sales and customers’ behavior by mining integrated Web content usage data. The analyzed results provide valuable insights into customers’ behaviors at a merchant’s e-commerce Web site and content effectiveness, thus building a more effective and profitable e-commerce system.
2. Provides an experimental test-bed for reasoning about differentiated marketing strategies through intelligent data mining of Web content usage data. The results will help merchants to significantly improve customer relationships management and predictive, personalized, and targeted marketing.
3. Provides a practical study of using a probability-based clustering algorithm and a high efficient frequent sequence mining algorithm to mine Web content usage data.

We draw on our experience19-21 in data mining and rely on the well found existing data mining algorithms17,18,20,22 in the literature for our purpose. Therefore, we abstract away many of the fundamentals of data mining in this paper. Interested readers should see the following additional resources.1,2,8,9 Furthermore, some studies23-30 have relied solely on the application of existing algorithms in the literature without the burden of developing any new algorithm. Besides, several studies examining the comparative performance analysis of various clustering and sequential patterns analysis algorithms exist.28,29,31-36 Also, Xu and Tian17 provide a comprehensive survey of clustering algorithms while Mooney and Roddick38 provide a detailed analysis of sequential pattern mining approaches and algorithms. These studies provide a knowledge reservoir on which one can rely. Therefore, we ignore any urge for a comparative performance analysis in this paper because doing so will shift the focus away from the more important task of the discovery and analysis of the relationship between product sales and customers’ behavior by mining the integrated Web content usage data. Our ultimate goal is to enable the online merchant have valuable insights into customers’ behaviors and content effectiveness, and thus provide enhanced marketing services generally and offer differentiated marketing services to its existing and potential customers thereby building a more effective and profitable e-commerce system. Similarly, we note that although

---

For example, one may discover that customers buy certain products, say egg and bread, together and can cause specific types of egg to be purchased with certain types of bread, such as white bread. This knowledge and many other previously undiscovered rules can be used to maximize profits by assisting successful optimization of marketing campaigns through strategic reasoning and adoption of practical mechanisms, such as the non-availability of discount on both products simultaneously.
we used existing algorithms (because of the matured research in this domain), the results of our experiments will provide business analysts better understanding of customers’ interests for potential opportunities to increase sales, lower costs, and efficiently manage supply chains.

While those studies above are significant, none has specifically examined the use of expectation–maximization algorithm and the sequential patterns mining to find relationships in customers’ transactions and activities on an online merchant’s website.

Also, although the expected maximization and the prefixspan algorithms are well established, their use in the e-commerce domain has not been extensively studied. To our knowledge, no documented study exists in the literature using these algorithms for visitor grouping and path analysis in order to find interesting patterns in Web content usage log data, an important component of e-commerce Web sites traffic analysis. In addition, we focus on web content usage log data rather than mere web site usage log data.

Our approach is novel because we focus on the mining of web content usage patterns in the e-commerce domain in order to improve and provide enhanced customer services and, thereby, enhance profitability of online merchants. Our intent is to espouse the discovering of how existing or potential customers use a website’s content with a view to improving the effectiveness of the website. We believe that online merchants will largely benefit from our approach in determining the marketing effectiveness of a site by quantifying user behavior while actually visiting the site, which allows the Web site administrator to provide personalized services and improve customer satisfaction. Thus, the e-commerce data analysis provides an indicator of the degree of user convenience in using the interface forms, shopping cart, payment, and so forth, in enhancing sales.

This paper is significant for the following reasons:

- Mining Web content usage log can tell exactly what items customers are interested in instead of ambiguous URLs, thereby enhancing the personalization of digital marketing and merchandising for customers, which offers enhanced shopping experience for consumers and improved sales/market share for the merchants.
- Experience gained in the implementation of our experimental test-bed offers us opportunity to gain insights into the intricacies of Web content usage data mining. Furthermore, it provides a stimulating learning experience to other system developers. The basic principles and thrust on which our design is anchored is applicable to solving real-world problems.

We modeled our prototype Web site as a collection of groups of related objects instead of a collection of pages. Using this approach offers a more effective way of modeling, organizing, and interconnecting Web data, and thus, provides a more efficient way to analyze the Web log data. Our prototype includes three major parts; a Web object content logging module for Web server, an object-oriented Web store running on the Web server with the enhanced Web object content logging module, and Web object log data mining. The Web content logging module captures and logs any information about the content that a client requests. This approach is a departure from the usual Web server logging method that does not log any information about the content that a client requests, so analyzing this kind of Web log data cannot give sufficient insight of how the Web site is serving the business. The Web server enhanced with a Web object content logging module is able to log interesting data attributes in the return pages as it supports logging for both static and dynamic Web pages. The Web content logging module is flexible in the configuration of the interesting data attributes names while supporting the ability to handle and capture information from both static and dynamic Web pages. The Web store is designed, modeled and implemented in an object-oriented approach, which makes the Web site easy to maintain and provides better understanding of the Web site content and analysis results for business analysts. The Web store provides a platform to demonstrate how to utilize the Web object content logging module to capture interesting business data. The log data of clients’ requests and responses from this Web store provides the source data for the next step, Web content usage data mining, to uncover hidden patterns and relationships to provide analysis report for business purpose. The Web content usage mining component provides an interactive user interface to give analysis results of the Web content usage data for business analysts. With better understanding of how the Web site is serving its customers, business decision-makers are able to make more profitable strategies for e-commerce.

Web content usage mining analyzes information about visited Web page content saved in Web site usage log file in order to discover interesting patterns previously unknown and potentially useful. Current Web data mining pattern discovery algorithms are based mostly on static pages. However, when these algorithms are applied on dynamic Web pages they present skewed results.

Our implementation includes the following two functionalities:

- Clustering of Web usage data and Web content objects. Clustering algorithms are used to group together a set of items having similar characteristics. These groups can be classified by using unsupervised inductive learning algorithms.
• Association rules, sequential patterns and dependency modeling of Web usage data and Web content objects. These
data mining techniques are used to discover possible relationships among items. The goal is to associate related infor-
mation and then attempt to discover the inter-session patterns in temporal ordered set of events. This function is used
to analyze customer groups’ and individuals’ behaviors in-depth, and as well as detect trends. The analysis results can
be used to build a more effective market campaign and convenient navigation.

We used XML and its related technologies, such as Stylesheet transformations (XSLT), to simplify some procedures,
such as data modeling, data parsing, and so forth. Introducing metadata for Web objects’ attributes by XML technologies
greatly assists data analysis for business purposes. Using XML also provides better compatibility and easier deployment.

We used our prototype experimental Web store over a three-month period, accessed randomly by 250 students and
faculty, to simulate online consumers in order to generate data for analysis to test the effectiveness of our selected algo-
rithms. A real online merchant’s website transaction log data are unavailable because the numerous online merchants we
approached refused to grant third party access, perhaps due to consumer data protection compliance requirements.40-43

This study of mining Web content usage patterns in e-commerce systems deals with issues that ultimately affect
the overall e-commerce system’s effectiveness and profitability, and hence the factors that affect customer satisfaction
of e-commerce website, and the enhancement of profitability of the online merchant. Thus, this paper has practical
implications for e-commerce website management, design and performance enhancement, and profitability.

Our approach provides affinity analysis for discovering co-occurrence relationships among activities performed by
online customers. Thus, it can help online merchants understand customers’ purchasing behavior. These insights can
drive online merchants’ revenue through smart marketing and sales strategies and can assist them in developing customer
loyalty programs, sales promotions, as well as discount plans.

The practical implications of our proposed analysis are many; its usefulness to the online merchant, difficulties in
consumer data protection conformance, and so forth. To the online merchant, our proposed analysis offers a wide variety
of insights, such as item affinity, pull items identification, revenue optimization, marketing, and operations optimization.

• Item affinity defines the likelihood of two (or more) items being purchased together. That is, it determines the likelihood
  that a set of items will be bought together.
• Pull items identification enables the identification of the items that pull people to the [web] store. These items always
  need to be in stock.
• Revenue optimization helps in determining the price points for the store, increasing the size and value of the market
  basket.
• Marketing insight helps in identifying more profitable advertising and promotions, targeting offers more precisely to
  improve return on investment, generating better customer loyalty promotions with longitudinal analysis, and helps in
  attracting more traffic to the store.
• Operations optimization helps in matching inventory to requirements by customizing the store and assortment to trade
  area demographics, optimizing store layout, and so forth.

The online merchant can effectively create a convenient and easy way to assess its customers with different charac-
teristics such as products’ units sold per customer, revenue per transaction, number of items per basket (shopping cart),
and so forth. These insights will help online merchants to make specific offers to the right customer segments/profiles,
as well as gain understanding of what is valid for which customer, predict the probability score of customers responding
to an offer, and understand the customer value gain from offer acceptance. Thus, product affinity information is critical
to enable the online merchant plan promotions appropriately because a price variation (decrease or increase) for some
items may cause corresponding demand variations on related high-affinity items; therefore, the related items would not
undergo any further promotion. Our analysis approach reveals regularities between products. In addition, the analysis
of an online merchant’s web data can provide further insights to improve the design of e-commerce web site, to monitor
and optimize website performance, and to provide personalized services.

Although some item affinities might seem very obvious, there are some affinities that are not. For example, the product affinity between toothpaste
and tuna is non-trivial; however, it appears that people who eat tuna are more likely to brush their teeth immediately after a meal because of the smell
of tuna. On the other hand, it is very obvious that there is very high product affinity between frozen potato chips, ketchup, and vegetable oil as it is
very typical for people who buy frozen potato chips to also buy ketchup and vegetable oil. Similarly, there is high product affinity between vegetable
oil, egg, and bread. Therefore, it is important for online merchants to have a good understanding of product affinities.
Furthermore, strict consumer data protection requirements conformance is a limiting inhibitor to the wider applications of the technology we proposed. This is because all consumer-identifying information must generally be restricted, which could limit the amount and kinds of targeted marketing campaigns. Therefore, online merchants desirous of adopting our technology should be aware of the need to conform to the various consumer data protection requirements.

Web log data oftentimes contain incomplete data, have missing data points, or have unobserved (hidden) latent variables. Therefore, in our analysis, we used the expectation–maximization (EM) algorithm to provide an iterative way to approximate/find maximum-likelihood estimates for model parameters. However, it is necessary to preset a maximum number of iteration to prevent infinite loop when the likelihood does not converge. Our approach identifies item-based patterns and rules in the transactions during data mining; thus able to reveal the relationships among items presented in the transactions in dynamic Web content scenario.

It is important to note that association (that is, correlation) is not synonymous with causation; thus this clear distinction requires careful attention when interpreting rules in the analysis of results. We did not examine causation in this paper but only clustering and association.

The rest of this paper is organized as follows: Section 2 provides a general review of some of the research in the Web analysis of e-commerce data and some of the methodologies suitable for analyzing a customer’s/visitor’s behavior and customer’s purchasing habits on a website in order to increase revenue and customer satisfaction through careful analysis of visitor interaction on the website. In Section 3, we briefly discuss the fundamentals and development of data warehouse for Web content usage data to provide appropriate context. Section 4 discusses the implementation of our experimental test-bed for the mining of Web content usage data using two popular data mining algorithms (expected maximization and the prefixspan algorithms), while we present and discuss our experimental results in Section 5. Finally, we conclude the paper and outline some future directions for this research in Section 6.

2 | REVIEW OF RELATED LITERATURE

This section is organized around topical thematic areas to give meaning, structure, and insights into the materials covered in the subsequent sections. It provides a general review of some of the research in web analytics including web scraping and clickstream analysis and web analysis of e-commerce data, data mining algorithms, classification/clustering, association rules algorithms, precision marketing campaigns, and some of the methodologies suitable for analyzing a visitor’s behavior and customer’s purchasing habits on a website in order to increase revenue and customer satisfaction through careful analysis of visitor interaction on the website.

A user’s navigational behavior indicates their steps through the shopping process in the e-commerce environment/domain. Thus, analysis of logged users’ navigational data is critical to the understanding of customers’ behaviors and the success of any online business. Generally, traffic data analysis is based on session data. It can also be based on the individual users if the users can be identified. When users’ personal data become available from other sources, such as user inputs, user tracking tools, and so forth, advanced information can be obtained by combining these users’ data with path analysis.

Web analytics is a technique for understanding users’ online experience for improvement of the overall quality of experience of the users. In other words, web analytics is a technique used to collect, measure, report, and analyze website data in order to assess the performance of the website and optimize its usage with the ultimate goal of increasing the return on investment. Web analytics provides a tactical approach to track key metrics and analyze visitors’ activity and traffic flow and generate reports. Thus, it is an indispensable technique for e-commerce merchants.

Web analytics is a subject of high interest and attraction lately; and it is well documented in the literature. For example, Dykes presents a detailed analysis of the evolution of Web analytics, and many studies provide an in depth examination of the rationale for web analytics while some studies examine web analytics and web metrics tools and their characteristics, functionalities and types, and data acquisition approaches and the selection of web tools for particular business models. In addition, Clifton examines some available web analytics methodologies and their accuracy. These studies collectively elevate Web analytics as an indispensable tool for e-commerce merchants.

Bucklin and Sismeiro assert that two major categories of data are used for analysis; user-centric data and site-centric data. User-centric data, that is, data collected based on individual users, which include all browsing behavior of a user on all websites, is typically collected by Internet Service Provider (ISP). User-centric data permit the creation of a user’s profile of all Internet usage across multiple channels. Site-centric data, that is, data collected from a single website, represent the activities and behaviors of visitors on the website. Site-centric data permit focused data mining and understanding
of the context of the website. In this paper, we focus on site-centric data in performing website usage characterization by identifying patterns and regularities in the way users access and use web resources.

Rao\(^{60}\) provides a fundamental treatise of the path from raw data to stored knowledge, while Rao\(^{61}\) discusses the factors that lead to the creation of untapped data that organizations routinely store during normal operations (dark data), the applicable steps to curate and manage data more effectively, and the methods to extract and use dark data. To buttress the need for data analytics, Grover et al.\(^{62}\) examine the value proposition of big data analytics (BDA) by delineating its components and offer a framing of BDA value through extension of existing frameworks of information technology value based on effective use of data resources. Similarly, García et al.\(^{63}\) for example, provide an integrated approach that focuses on web analytics in e-commerce. Also, Correa et al.\(^{64}\) provide a concrete implementation of web mining techniques in the food delivery services as a more specialized e-commerce platform. Similarly, Shamout\(^{65}\) examines the effectiveness of supply chain analytics in enhancing firms’ supply chain innovation and robustness capability in the Arabian context by using variance-based structural equation modeling (PLS-SEM) to model the association between supply chain analytics, supply chain innovation and robustness capability. He concludes that supply chain analytics can help managers access timely and useful data for greater innovation. Makhabeletal.\(^{10}\) explore data mining techniques and show how to apply different mining concepts to various statistical and data applications in a wide range of fields, including description and implementation of a suite of social media mining techniques. Although these studies\(^{10,60-65}\) provide implementations of Web analytics, they differ from our approach in our choice of selected algorithms and goal.

Analyzing a visitor’s behavior and customer’s purchasing habits on a website using specific engagement metrics data provides critical insights into the performance of product pages, and the optimization and improvement of the effectiveness of the e-commerce solution. Ezzedin\(^{66}\) examines the top engagement metrics for each step of the purchasing cycle and show how to analyze the data collected for the different users’ segments using Google Analytics\(^{49,51,67}\) measurement platform.

An overview of methodologies suitable for analyzing websites in order to increase revenue and customer satisfaction through careful analysis of visitor interaction with a website is available in Booth and Jasen.\(^{58}\) The study presents how basic visitor information, such as number of visitors and visit duration, can be collected using log files and page tagging by including a “tracking code on every page of your website, and then access reports to view the data that is collected.”\(^{67}\) Usually, each user of a website creates a visitor path.\(^{64}\) A visitor path is the route a visitor uses to navigate through a website.\(^{58}\) Each visitor creates a path of page views and actions while on a website. By studying these paths, one can identify usage characterization of the website and any challenges a user has in using the website.

Nguyen et al.\(^{47}\) use web usage mining process to uncover interesting patterns in web server access log file gathered from Ho Chi Minh City University of Technology (HCMUT) in Vietnam. By incorporating attribute construction (or feature construction), one of strategies of data transformation of data pre-processing technique, they had wide knowledge about users access patterns for every country, province and ISP. Such knowledge is useful for optimizing system performance (such as deciding reasonable caching policies for web proxies) as well as enhancing personalization.

Clickstream data provide information about the sequence of pages or the path viewed by users as they navigate a website. Clickstream data enables consumer’s online behavior analysis, and explains the effectiveness of marketing actions implemented online. This is possible because clickstream data provide information concerning the sequence of pages viewed and actions taken by consumers as they navigate a website.\(^{68}\) The sequence of viewed pages and actions taken are commonly referred to as “paths”, and the clickstream data collected provide valuable insight into how the website is used by its users. However, as Clark et al.\(^{69}\) note, clickstream data do not reveal the true intentions of the user on a website, or other possible activities that the user engaged in during the use of the website.

Montgomery et al.\(^{70}\) show how path information can be categorized and modeled using a dynamic multinomial model of clickstream data using data from a major online bookseller. Their results suggest that paths may reflect a user’s goals, which could be helpful in predicting future movements at a website. A potential application of their model is to predict purchase conversion. This technique is useful in the personalization of Web designs and product offerings based on a user’s path. Noreika and Drašutis\(^{71}\) propose website activity data analysis model based on a composition of website traffic and structure analysis models with intelligent methods. This approach enables theoretical predictions on how and what factor changes in website structure affect a visitor’s click paths and overall website activity. Their model relies on the main principle of dividing the website analysis into two parts; namely website structure analysis model and website traffic analysis model. They construct and formalize these models separately and then establish a relation function between them based on intelligent methods. One of the limitations of their work is that they only describe the models construction leaving out the key intelligent methods as a black box, which leaves too many unknowns.

Ehikioya and Lu\(^{44}\) propose a path analysis model as an effective way to understand visitors’ navigation of a website, which can provide a lot of useful information about users’ navigation and a website’s usage. Also, Ehikioya and Zheng\(^{39}\)
present a Web content usage logging system for Web administrators and business analysts to capture Web site visitors' interests at a fine granular level, and show how a Web site, designed using the object-oriented paradigm, can benefit from this logging system to capture interested objects' attributes and relationships among these attributes. The Web content usage log provides valuable data for Web usage data mining with minimal effort in data extraction, transformation, and loading (ETL). Furthermore, Ehikioya and Lu 72 discuss three different approaches (i.e., improved single-pixel image, JavaScript tracking and HTTP proxy server), that work together to track a user's activities. These approaches have fewer limitations than the existing approaches. Similarly, Fernandes et al. 73 propose an algorithm that uses paths based on tile segmentation to build complex clusters. The algorithm offers two advantages; it does not create overlapping clusters, which simplifies the interpretation of results and it does not demand any configuration parameters from users, making it easier to use. Also, Lavanya and Princy 74 discuss concept maps and data mining techniques, and graph reading algorithms used for concept map generation and tabulate popular data mining techniques used in BDA.

Many studies 55,75-79 examine consumers behavioral patterns online. For example, Ellonen et al. 75 analyze consumer behavioral patterns on a magazine website using a unique dataset of real-life clickstream data from 295 magazine website visitors. They found interesting behavioral patterns that 86% of all sessions only visit the blogs hosted by the magazine. Similarly, Ribeiro 76 examines the navigational patterns of users on the website of Shifter, an online media company, for a 3 months period using Microsoft Excel tool to obtain a context for each piece of content produced and published. Analysis of Shifter’s data resulted in recommendations for rethink, and the redesign, of the editorial content of the business to respond to different community’s needs. Also, Linden 77 examines behavioral patterns of web users on an online magazine website with a view to first find and visualize user paths within the data generated during collection, and then identify some generic behavioral typologies of user behavior using cluster analysis and sequential path analysis. He used a dataset of clickstream data generated from the real-life clicks of 250 randomly selected website visitors over a period of 6 weeks using Microsoft Excel to visualize user paths and analyze descriptive studies based on the clickstream data. The analytical process focuses on a combined methodology of cluster analysis and swim-lane diagrams. Similarly, Jain et al. 78 and Pani et al. 79 provide an analysis of Internet browsing and site usage behavior using sequential access pattern mining, while Siddiqui and Aljahdali 55 discuss Web mining tree structure.

Although many studies on sequential access patterns mining exist, 18-20,77-83 most of them focus on improving the efficiency of mining sequential access patterns. Agrawal and Srikan 80 introduce three algorithms (AprioriAll, AprioriSome, and DynamicSome) and the AprioriHybrid 81 based on an association rules mining algorithm, the Apriori algorithm, 80,81 for mining sequential patterns. Association rules mining finds frequent sets of items and, therefore, generates desired rules. Association rules mining usually ignores item sequence in transactions. Items within an itemset are kept in lexicographic order. Generally, mining association rules focuses on finding intra-transaction patterns while mining sequential patterns focuses on inter-transaction patterns. 80 These apriori-based algorithms differ majorly on how to generate candidates and prune the candidates that will not lead to large itemsets efficiently. Improving efficiency in these two areas can greatly improve the performance of the algorithms since the transaction database may be huge in size. The PrefixSpan algorithm 18 uses database projection for frequent sequences to make the candidates for next pass much less than the candidates generated by Apriori-based algorithms. This innovative approach makes the PrefixSpan algorithm much faster and highly efficient in mining large itemsets, hence it became our candidate choice algorithm for implementation.

Besides the above studies, Jokar et al. 84 examine Web mining and Web usage techniques while presenting an efficient framework for Web personalization based on sequential and non-sequential patterns, and analyze the structure of the web pages using compression of tree structure method.

Generally, in Web log mining domain, sequential patterns are interesting because they not only report popular itemsets, but they also reveal underlying relationships among them. With the understanding of frequent sequential patterns, businesses can organize Web contents, control inventory, manage logistic more efficiently, and make effective target marketing strategies and predict future trend.

Lewis and White 85 present a method for web usage mining based on a linear ordering of the age transition matrix created from web server access logs. This ordering facilitates the categorization of web pages into different classes (such as origins, hubs, or destinations) based on position in the linear order; thus providing a measure of the orderliness of website traffic. They applied this technique to website traffic of a university over time by comparing the website traffic immediately after a major change to the website design and the traffic 2 years later since changes in website organization could also dramatically change visitors flow. The results show the traffic is more ordered. Similarly, Asha and Rajkumar 86 discuss web usage mining techniques for enhanced quality of experience of customers shopping online on websites and also discuss web mining techniques to find dishonest recommenders in open social networks. They propose a recommendation system that uses semantic web mining process integrated with domain ontology which can be used to extract interesting patterns from complex and heterogeneous data.
Ohta and Higuchi\textsuperscript{87} analyze store layout that underlie supermarket store design and product display styles and then examine the interaction between shop floor layout and customer behavior from the perspective of the supermarket owner to discover the main sections within the shop likely to attract customers into the store. The authors made a general classification between the standard layout, which accounted for approximately 90\% of the survey sample, and the minority layout, used by less than 10\% of the survey sample. Using the survey results, they analyzed the customer circulation rates and section drop-by rates as influenced by the store layout. They concluded that the standard layout is superior. This study is fundamental and analogous to the behavior of online visitors to e-commerce websites.

Zheng et al.\textsuperscript{88} propose a way of detecting fraud in users transactions by extracting the behavior profiles (BPs) of users based on their historical transaction records, and then verify if an incoming transaction is a fraud or not in view of their BPs. The Markov chain models\textsuperscript{89} are popular in representing BPs of users, which is effective for those users whose transaction behaviors are relatively stable. However, with the development and popularity of e-commerce, it is more convenient for users to shop online, which diversifies the transaction behaviors of users. Therefore, Markov chain models are unsuitable for the representation of these behaviors. However, they\textsuperscript{88} propose use of logical graph of BP (LGBP), a total order-based model, to represent the logical relation of attributes of transaction records. Based on the LGBP and users’ transaction records, one can compute a path-based transition probability from one attribute to another, and diversity coefficient to characterize users’ transaction behaviors and diversity. In addition, to capture temporal features of a user’s transactions they also defined a state transition probability matrix. Their experiments over a real data set illustrate that the LGBP method can characterize the users transaction behaviors precisely, and abstracts and covers all different transaction records.

Traffic data tracking and analysis is pivotal for Web site management and marketing in e-commerce. While several analyses use sequential pattern discovery (i.e., path analysis) techniques\textsuperscript{90-93} to discover frequent path patterns, some authors use advanced path analysis to achieve more complex tasks such as serving as a basis of personalization\textsuperscript{93} or recommendation systems.\textsuperscript{94-98} Recommendation systems and personalization are two related and popular areas in Web data mining. They both apply statistical and knowledge discovery techniques to achieve serving/selling of more products and services, thereby enhancing the profitability of e-commerce sites.\textsuperscript{98} In a recommendation system, a new user is matched against a pre-built database, which stores consumers’ preferences for products. If some neighbors, who are customers already in the database and have the same taste as the new user, are found, products favored by those neighbors are recommended to the new user. An example of using path analysis for recommendation systems is to predict HTTP requests,\textsuperscript{99} which is based on path profiles and recommends an URL with a high probability to the user before the user makes such a request.

Kahya-Özyirmidokuz\textsuperscript{100} analyzes large amounts of Facebook social network data which are generated and collected for valuable decision making information about online shopping firms in Turkey in order to have a competitive advantage by translating social media texts into something more quantitative to extract information. The author used web text mining techniques to determine Facebook patterns in 200 popular Turkish online shopping companies’ web URLs via similarity analysis and clustering. Consequently, the clusters of the Facebook websites and their relationships and similarities of the firms are obtained.

Chen et al.\textsuperscript{101} examine the usage behavior patterns of mobile telecommunication services users using opinion leaders deemed tremendously influential on the usage behavior of other users. They examined data from one of the largest Taiwanese telecommunications databases and try to identify mobile opinion leaders and further cluster their mobile usage patterns by mining the actual data. This study exploits a combination of techniques, including statistics, data mining, and pattern recognition, in the data analysis of opinion leadership theories applied in the traditional marketplace into mobile services based on a big data system. Furthermore, they provide taxonomy to logically analyze each pattern of mobile content usage behavior gathered from mining the data to provide better planning blueprint for future mobile resource consumption.

Sunil and Doja\textsuperscript{102} discuss web data mining strategies and applications in e-services which are required for optimizing website structure that will help businesses and learning platforms to increase their revenues, attract new and retain old customers or learners, and assist developers to increase the frequency of customers/learners visits. Chajri and Fakir\textsuperscript{103} provide an introduction to the concept of data mining and the application of data mining techniques in e-commerce while Sharma and Vaisla\textsuperscript{104} provide a survey of application of data mining in e-commerce and business intelligence.

Generally, web scraping is the practice of gathering data commonly accomplished by writing an automated program that queries a web server, requests data (usually in the form of HTML and other files that compose web pages), and then parses that data to extract needed information. It encompasses a wide variety of programming techniques and technologies, such as data analysis, natural language parsing, and information security. Web scraping is highly scalable and fast as one can create massive big datasets with tens of thousands of variables, as it can also be used to create modestly sized,
more manageable datasets with tens of variables but hundreds of thousands of cases, which one can analyze within a few hours. Thus, web scraping involves the automated collection of information from web pages. Many studies have relied on web scraping lately. For example, Landers et al.\textsuperscript{105} examine web scraping as an approach to collecting and analyzing data in big data systems/environment. Additionally, Mitchell\textsuperscript{106} discusses Web Scraping with the Python programming language and provides a comprehensive guide to the automated gathering/collecting, transforming, and using data from uncooperative sources, including the Web. Also, Broucke and Baesens\textsuperscript{107} provide a concise, practical, and modern guide to web scraping using Python programming language. Similarly, Russell\textsuperscript{108} and Munzert et al.\textsuperscript{109} discuss web scraping and text mining techniques that can be implemented in Python and R, respectively, that enable one to create powerful collections of existing, but of previously unanalyzed unstructured or unsorted data at very reasonable cost.

Flory et al.\textsuperscript{110} provide an effective and efficient solution of how-to-design decision support systems to address the consumers’ need for non-burdensome sense making of online reviews through interactive web personalization artifacts and validate their superior performance for adequately validating the solutions of review quality research. Due to the increasingly high volume of such reviews, automatic analyses of their quality have become imperative. Similarly, Alkalbani et al.\textsuperscript{111} examine reviews by cloud consumers that reflect consumers’ experiences with cloud services. They analyzed the reviews of about 6000 cloud service users using sentiment analysis to identify the attitude of each review, and to determine whether the opinion expressed was positive, negative, or neutral using two data mining tools, KNIME and RapidMiner, based on four supervised machine learning algorithms: K-Nearest Neighbor (KNN), Naive Bayes, Random Tree, and Random Forest. The results show that the Random Forest predictions achieve 97.06% accuracy, which makes this model a better prediction model than the other three.

Revathy and Lawrance\textsuperscript{30} examined how data mining can help farmers to develop yield through applying data mining techniques to determine how crops can be protected from pests by predicting and enhancing crop cultivation. They compared C4.5 and C5.0 decision tree algorithms for pest data analysis with an experimental approach and found C5.0 proved its efficiency by giving more accurate result rapidly and using less memory.

Similarly, Shanmugarajeshwari and Lawrance\textsuperscript{23} and Lawrance et al.\textsuperscript{24} present classification techniques for educational data mining based on C5.0 algorithm which offers good classification accuracy. The educational data used to evaluate teachers’ performance is based on course evaluation questionnaire and the students’ perception using classification techniques based on decision tree. Also, Agaoglu\textsuperscript{25} used four different classification techniques – decision tree algorithms, support vector machines (SVMs), artificial neural networks, and discriminant analysis – to build classifier models and compared their performances over a data set composed of responses to the students’ questionnaire using accuracy, precision, recall, and specificity performance metrics. He used the C5.0 algorithm to predict and improve the instructor performance. Asanbe et al.\textsuperscript{26} present an efficient system model for evaluation and prediction of teachers’ performance in higher institutions of learning using data mining technologies. The results show that, considering the time taken to build the models and performance accuracy level, C4.5 decision tree outperformed the other two algorithms (ID3 and multilayer perceptrons [MLP]) with good performance of 83.5% accuracy level and acceptable kappa statistics of 0.743.

Ughade and Mohod\textsuperscript{27} present and compare two approaches – multiple classifier approach and single classifier approach – for relative evaluation of faculty performance. In the multiple classifier approach, KNN is used in first step and rule based classification is used in the second step of classification activity, while in single classifier approach only KNN is used in both steps of classification.

These studies show the effectiveness and expressiveness of decision tree algorithms in data mining. Decision tree is one of the predictive modeling approaches for representing data that can be visualized.

Karthika and Janet\textsuperscript{28} discuss performance analysis of several clustering algorithms and compared the various clustering algorithms on real, numerical, categorical datasets around the cluster size. Their results show that the repetition of KMeans many times does not bring better significant iterations since it starts randomly and it depends purely on the initial choice of the centroid of clusters. Furthermore, the process of finding the clusters may not converge. The expectation–maximization (EM) algorithm,\textsuperscript{17} although is more time consuming than KMeans, accommodates noisy data and missing information. Similarly, Avinash et al.\textsuperscript{29} examine SVM classification, one of the most popular supervised learning methods, and evaluate the performance of SVM classification using sequential minimal optimization (SMO) algorithm for early detection of breast cancer. Results show their proposed model performs better in terms of accuracy (94.6%), recall (89.5%), and execution time (0.085 s) on Wisconsin data set.

Kantardzic\textsuperscript{5} examines web mining and text mining and discusses some of the applicable algorithms and formalizes a text-mining framework specifying the refining and distillation phases. Also, Gorunescu\textsuperscript{9} presents a comprehensive guide of usage for an array of algorithms and techniques/methods for data mining. These algorithms and techniques/methods include: Bayesian classifier/Naive Bayes, Artificial neural networks, SVMs, Association rule mining,
Rule-based classification, KNN, Rough sets, Decision trees (which is a special classification technique), Clustering algorithms, and Genetic algorithms.

Alagukumaret al.112 adopt association rule mining, one of the most important procedures in data mining, in microarray gene expression analysis, where a large number of rules are often discovered. They proposed a novel method for clustering association rules derived from microarray gene expression data. The gene expression data are converted into gene expression intervals using discretization technique. The association rules are extracted from the maximal frequent itemsets of the gene expression based on the support and confidence. To calculate the similarity matrix for the derived association rules, they used the Euclidean distance measure, and then used the single linkage agglomerative clustering algorithm to cluster the association rules based on the similarity matrix. The results from using the proposed method show that it performs better than other methods such as complete linkage and average linkage. Similarly, Vengateshkumar et al.113 propose a novel gene association rule algorithm called Boolean association rule (BAR) mining for analyzing and extracting interesting knowledge from microarray gene expression which contains a dense amount of data. The model uses t-test to filter the non-informative genes, k-means clustering to discretize the gene expressions, and the Boolean association technique to generate frequent gene expressions.

The efficiency of a data mining algorithm generally depends on the computation time and memory utilization which is affected by the data structure used for storing the itemsets and its input/output (I/O) complexity. Pandey et al.114 use one shared near neighbor based algorithm by minimizing its I/O complexity to make it suitable for big data in external memory model. Implementation results of the algorithm show its efficacy for big data sets. The computational steps remain unchanged, thus maintaining the same cluster quality. According to Kharkongor and Nath,115 one of the main challenges in association rule mining is mining frequent itemset because the efficiency of frequent itemsets depends on the computation time and the data structure used for storing the itemsets. The data structure significantly influences the memory and space requirement of the algorithms. Most of the association rule mining algorithms work well for a sparse dataset. However, with large dataset, it becomes computationally difficult and expensive, which eventually increases the execution time. This deficiency consequently affects the scalability of the algorithm. Therefore, a compact and concise representation of the itemsets is vital to enable the itemsets fit into memory and thus bypasses the need for any I/O operations. The array, tree, and trie are the mostly used data structures.

Zhang116 explores the idea of using visualization techniques and data mining to help deal with the increasing volume of information generated by e-commerce applications. Visual data exploration can easily deal with highly non-homogeneous and noisy data and the user is directly involved in the data mining process. He uses an application frame of e-commerce data mining combined with some visualization methods that can be applied on the e-commerce data mining. Similarly, Sever117 investigates the extent to which a product’s average user rating can be predicted, using a manageable subset of a data set on a linearization-algorithm based prediction model. Experiments show that the method’s accuracy is reasonable for reconstructing volatility of user ratings, a useful property both in accurate user predictions and sensitivity computation.

Jiang and Yu118 use K-means algorithm to cluster transactions data based on customer usage data of various e-commerce Websites. The statistical results show segmentation of the data into clusters and that there was a clear distinction between the segments in terms of customer behavior. Similarly, Tang and Peng119 use clustering analysis algorithm for customer grouping based on key customers purchase information.

Yin and Pan120 examine precision marketing campaigns of B2C e-commerce enterprises using data mining to analyze valuable information from consumers shopping on Jingdong Mall in China. Similarly, Xu and Chen121 analyze the application of big data technology in B2C e-commerce precision marketing pattern using China Amazon B2C electronic commerce as an example. Also, Erevelles, Fukawa, and Swayne122 propose a conceptual framework that builds on resource-based theory to better understand the impact of big data on various marketing activities that enable firms to take advantage of the benefits of big data. They identified three resources – physical, human, and organizational capital – as key elements/resource requirements for organizations to benefit from big data. Precision marketing strategies of e-commerce enterprises have practical value because the resolution of how enterprises can accurately match the characteristics of its products or services with consumers is easily achievable, and also help to maintain a relatively stable customer group size and structure.

Hongyan and Zhenyu123 provide an in-depth study of the progress of theoretical research services and service quality management theory necessary for the establishment of collaborative filtering, recommendation systems, predict consumer model, personalized recommendations and services to address the information overload problem arising from social integration of e-commerce systems based on large data analysis and complex network in order to predict consumer behavior.
The measurement of visitors’ website activities relies highly on data mining techniques. The ability to find clusters in data without knowing features of the data sets is quite relevant in the analysis of massive e-transactions data (in the big data and data mining domains). Big data deals with large data sets that may produce clusters with arbitrary shapes from different sources, such as geographic systems, medical systems, sensor systems, e-commerce systems, and so forth. Web mining is the application of data mining techniques to discover and extract useful and interesting information from the Web. According to Arti et al., web mining is applied to e-commerce data set to know the browsing behavior of customers, to determine the success of marketing efforts, to improve the design of e-commerce web site, and to provide personalized services.

To gain competitive advantage, businesses need to have detailed information of the activities of online customers on their websites in order to have actionable and decisive knowledge. However, to monitor and to optimize website performance, organizations need strong web analytics tools and skills. Kumar and Ogunmola present a comprehensive review, and a comparative analysis, of the most important web analytics tools and techniques, which are vital to report website performance and usage. Also, a profile of commercial web analytics software products (such as Visitor Analytics) to guide businesses is available in.

### 2.1 Summary

While existing Web analytics studies provide a tactical approach to track key metrics, analyze visitors’ activity and traffic flow, and generate reports, our study aligns with the position held by Kimball and Merz of the immense value in clickstream analysis and reinforces the criticality of web content usage data mining process that could capture and accurately analyze the interactions between Web users and the sites they access. This information could dramatically improve an organization’s understanding of its relationship with users, and radically improve the strategic knowledge of customer activities (motives and actions) and, thus, web content usage data mining is an indispensable technique for e-commerce merchants, thereby adding to the body of knowledge in the domain. In particular, we capture and analyze activities on both static and dynamically created pages which put our work in a distinct class.

Furthermore, although many implementations of Web analytics exist, these studies provide implementations (such as various summary statistical and data applications in a wide range of domains) different from our selected algorithms (EM and the prefixspan algorithms). In addition, there is knowledge gap in the application of the EM and the prefixspan algorithms, despite being well established in the literature, to the e-commerce domain. Our study contributes to filling this gap.

We note the existence of many studies on sequential access patterns mining algorithms. However, they focus mainly on improving the efficiency of mining sequential access patterns and they did not examine their application to web content usage data mining nor focus on dynamically created web pages. We rather devote our attention to the application of prefixspan algorithm to the e-commerce domain with its peculiar characteristic of interactivity. Similarly, while many studies have examined clustering approaches in many domains, their widespread application to web content usage data mining is limited. This knowledge whet our appetite to examine how creative online merchants can exploit the EM algorithm to mine the online merchants integrated enterprise web content usage data log for the discovery and analysis of the relationship between product sales and customers’ behavior by identifying item-based patterns in the transactions during data mining; and revealing the previously unknown, hidden, actionable, useful, and interesting information/knowledge and the relationships among items presented in the dynamic Web content usage data log transactions.

We adopt the approach taken in References by relying on existing algorithms to maintain focus on our main goal of finding interesting patterns in Web content usage log data. In this paper, we exploit results from existing methodologies such as visitor path analysis, top engagement metrics, clickstream data, content usage logging techniques, and other approaches suitable for analyzing websites in order to increase revenue and customer satisfaction through careful analysis of visitor interaction with a website. Our study supports existing studies that examine consumers behavioral patterns online.

### 3 DATA WAREHOUSING WEB CONTENT USAGE DATA

The web is a major source of e-commerce data. There is fierce competition among online merchants to attract new and retain existing customers. E-commerce generates commercial data in high volumes. This situation necessitates the
The extraction of knowledge from data in e-commerce sites, using data mining techniques, to gain insights into online customers’ behaviors. However, analyzing the vast volumes of commercial data is becoming increasingly difficult because of the unprecedented volume, velocity, variety, veracity, variability, various sources, various quality, and value of e-commerce transactions data, which characterize big data. Data mining is used to extract meaningful information from a large data source using some patterns and methods. Thus, data mining offers the capability of finding hidden information and correlation between massive data set that is helpful in decision making.

A key requirement for mining Web content usage data is the creation of a data warehouse to store the huge unstructured/semi-structured Web content usage data. This requirement is fundamental because most OLTP applications are not aware of business analysis during their development. This limits the scope of later business intelligence solutions to analyzing only stored data. A typical OLTP system focuses on processing business transactions. Usually, only sales and costs data are stored and can be analyzed. It does not record customers’ interests if there is no any purchase activity involved, so it is hard to discover some of the customers’ interests as potential business opportunities. Adding extra data attributes capturing and logging functions may require huge amount of effort to modify existing programs. According to Sommerville, feature maintenance tasks cost about 60% of the total cost of software after its initial deployment.

An e-commerce system usually contains an online product catalog together with online ordering system. Analyzing how customers view the products can help find potential business opportunities. A traditional e-commerce analysis system builds on top of its product catalog and online ordering system. Instead, our Web content usage logging system provides a new approach to building OLTP applications on top of a business analysis aware platform to allow quick and flexible capture and logging of data and metadata for later business analysis purposes. The Web content usage log does not only include business transaction data, but it also contains product attributes and their values that the customers viewed, which are usually not part of data in OLTP systems. Real-time content usage data logging provides accurate data and strong content auditing ability. Using eXtensible Markup Language (XML) to build content usage data enables fast and flexible data ETL from the source Web content log to data warehouse for data mining. The little overhead of real-time Web content logging can potentially save huge amount of effort in the data ETL process for data mining.

### 3.1 Preprocessing web content usage log

The data preprocessing phase loads data from source Web log, in flat file format, into a cleaned data repository, preferably a database system for easier data retrieval by OLAP and various data mining applications. There are usually two steps: removing bad requests and integrating cleaned data with OLTP data. Cleaned data has consistent data structures and data types.

A Web server gets many kinds of requests. Apart from direct requests from users by typing the URLs in the browser, and clicking on hyperlinks on Web pages and bookmarks, there are hidden requests generated by browsers for Web page embedded script files and image files, and requests from various computer programs, such as search engine robots, computer virus and worms. All requests are recorded in the Web access log. For business analysis purpose, bad requests (i.e., requests having status code other than 200) should be removed because these requests have URLs in the format the Web server cannot recognize and no real information is returned for these requests.

Integrating cleaned data with OLTP data requires parsing the URLs to get parameter names and values. These parameters are used as key column to join OLTP data. Frequent updated data in OLTP system makes this step very challenging. The success of this step has direct impact on the success of the entire business intelligence solutions. When the OLTP system does not have strong data auditing and tracking ability, the data integration may associate historical data to wrong current data or even unable to associate any data. Although an OLTP system with strong auditing capability usually can integrate them successfully, the overhead in time comparison in a huge database is obvious.

Since the Web content usage logging system only records successful requests for dynamic and static Web pages, the data cleaning step can usually be skipped. Page embedded image requests are not logged. However, interesting image requests can be captured as content attributes in the Web page, when necessary. Therefore, it is possible to capture the image file name and path, as well as the metadata of the image. Interested readers should see for additional details.

Sometimes URL consolidation may be required. For example, requests “http://myserver/webstore”, “http://myserver/webstore/”, and “http://myserver/webstore/index.jsp” are actually pointing to the same page. Usually, there is no reason to differentiate these requests for business analysis purposes.

Each record in the Web content usage log contains a standard request header and a response body in XML format with no constraints in the data structure. An example of a Web content usage log structure is given in Figure 1. The
standard request header contains common user request information, such as client IP address, timestamp, request URL, and session identification, and so forth. The response body is basically an XML document.

Due to the variety of business information and requirements, this body does not enforce a rigid structure. XML is used to build a tree structure to represent data value, metadata and hierarchy relationships. Saving this XML data in an XML enabled database leaves extracting interesting attributes to implementation time, to be queried efficiently and to be changed with no impact on existing OLTP systems.

### 3.2 Web content usage data warehouse

Statistical reports, including OLAP reports, calculate aggregate value for user selected attributes/dimensions, usually by direct queries of the data warehouse. The Web content usage data warehouse uses a dimensional data model, which is composed of a central fact table and a set of surrounding dimension tables each corresponding to one of the components or dimensions of the fact table. The fact table contains Web entries along with session, access time, and URL and Web content attributes. These attributes serve as foreign key attributes referencing the primary keys of the constituent dimension tables.

The overall data schema has a star-like data structure, called a star schema. To support dynamic attribute hierarchies, the star schema is further normalized into a snowflake schema by allowing the dimension tables to have sub-dimension tables. The real attribute hierarchies in the snowflake schema are generated dynamically from the XML data just-in-time of data querying to calibrate the data model for best practise. Figure 2 shows the physical data schema for data population in our implementation, and Figure 3 shows a snowflake data schema with dynamically generated sub-dimensions.

In our Web content usage data warehouse schema, as shown in Figures 2 and 3, Web entries are facts along with session, access time, and URL and Web content dimensions. Attributes in the standard request header are in star schema model and can be queried by regular structured query language (SQL) statements. Attributes in the XML content builds extra dimensions on top of the URL and Web content dimension to make the overall data schema model a snowflake schema model. XML fields can be queried using the XML functions provided by the database together with regular SQL statements.
Although these XML data functions may be vendor specific, they all use XML Path Language (XPath) to specify an XML element or attribute location in an XML document. XPath is a language for addressing parts of an XML document. XML enabled database systems are usually optimized for efficient access of XML data, such as XML DTD data validate to ensure the XML data integrity, and side tables used in IBM DB2 for fast retrieval of XML data.

Figure 4 shows a sample OLAP cube. This cube has total-hits as the fact, and URL and product id as dimensions. It shows a 3D product hits bar chart for each URL based on the logged session cart data.
4 | MINING WEB CONTENT USAGE LOG

The Web content usage data warehouse simplifies data modeling procedures for data mining algorithms. This data warehouse provides great flexibility in choosing interesting data attributes for analyzing e-commerce Web site visitors’ activities and interests. The choice of techniques and algorithms depends on the discovery goals. According to Gorunescu,9 the data mining goals serve as the basis to distinguish more clearly its areas of application:

- “Predictive objectives (e.g., classification, regression, anomalies/outliers detection), achieved by using a part of the variables to predict one or more of the other variables;
- Descriptive objectives (e.g., clustering, association rule discovery, sequential pattern discovery), achieved by the identification of patterns that describe data and that can be easily understood by the user.”9

Thus, the objectives of data mining are predictive which uses some existing variables to predict future values (unknown yet) of other variables (based on methods such as classification, regression, biases/anomalies detection, and so forth); and descriptive which reveals patterns in data, easily interpreted by the user (based on methods such as clustering, association rules, sequential patterns, and so forth).

In this paper, we used two data mining techniques: visitor grouping and path analysis, which are among the most interesting subject areas for e-commerce Web sites traffic analysis. Visitor grouping uses data clustering algorithms to group visitors/sessions with similar selected attributes’ value. Path analysis usually refers to mining common visiting path sequence. Our choice of the two techniques is based on their popularity, flexibility, applicability, and capacity to handle high data dimension. The ultimate goal is to enable the merchant of the e-commerce Web site provides enhanced marketing services generally and offer differentiated marketing services to its existing and potential customers.

During this research, we noticed that most data mining algorithms are not efficient in processing strings and our Web content usage data warehouse usually has sparse data set. Attribute matrix can be huge in size. Reducing the size horizontally becomes very important to run the data mining algorithm efficiently.

4.1 Clustering

Clustering is an approach of partitioning data into groups according to some similarity criteria. Thus, clustering is a technique that partitions data into different groups in such a way that data items in a group are more similar to each other than the data items in any other group. A standard for clustering is the difference of inter-cluster distance and intra-cluster difference. Therefore, clustering techniques are used to divide instances into natural groups so the instances’ intra-group relationship is maximized while the inter-group relationship is minimized.6

Data clustering programs can build clusters based on the navigation URLs from traditional Web log. With more detailed page content information logged by the Web content usage logging system, data clustering algorithms are able to provide in-depth analysis. In real world scenario, the clustering approaches may vary in algorithms and data schemas, but they all depend on business requirements. We used the expectation–maximization (EM) algorithm,17 a popular probability-based clustering algorithm, to analyze customers’ shopping carts at the checkout point.

We select the EM algorithm for the following specific reasons, which are well documented by Abbas31:

- It has a strong statistical basis.
- It is linear in database size.
- It is robust to noisy data.
- It can accept the desired number of clusters as input.
- It can handle high dimensionality.
- It converges fast given a good initialization.
The accuracy of the EM algorithm becomes very good when using huge data set. The EM algorithm is abstracted in Algorithm 1, as follows:

**Algorithm 1. Expectation–maximization method**

```plaintext
/*%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%***/
/* Input: X: A database of n data items {x1, x2, x3, ..., xn} */
/* k: the number of clusters */
/* */
/* Output: A set of k clusters with maximum log likelihood */
/*%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%***/
0 Begin
1 Do
2 Expectation step:
   For each data item x ∈ X,
      Estimate the membership probability of x in each cluster.
3 Maximization Step:
   Update the parameter which maximizes the likelihood.
4 Until
   Converge or Goto step 2
5 End
```

The EM algorithm, which uses multivariate normal distributions, is a two-phase iterative method that alternates between execution of the Expectation step (in phase 1) and the Maximization step (in phase 2).

The Expectation step creates a function for the expectation of the log-likelihood evaluated using the current estimate for the parameters. The Maximization step computes parameters maximizing the expected log-likelihood found on the Expectation step. These parameter-estimates are then used to determine the distribution of the latent variables in the next Expectation step. The EM algorithm assigns a probability distribution to each instance which indicates the probability of it belonging to each of the clusters.

The complexity of the EM algorithm is $O(dnt)$ where $d$ is the number of input features, $n$ is the number of objects, and $t$ is the number of iterations. For the mathematical details of the EM algorithm, interested readers should see the following resources.

The data initialization procedures can find session database finding every unique attribute and counting the frequency of the unique attribute in each session, and the total frequency across all the sessions. These two frequencies are used to build the underlying Poisson distribution of each attribute. Poisson distribution is widely used to model the number
of events occurring within a given time interval for various phenomena of discrete nature. For nominal attributes with
discrete values, such as product id, the probability of each attribute value is the count of that attribute value divided by
the sum of the count of all values. Thus,

\[ P[x_i] = \frac{\text{count}(x_i)}{\sum_i \text{count}(x_i)} \]

where \( P[x_i] \) denotes the possibility of an attribute having value \( x_i \), \( \text{count}(x_i) \) denotes the total number of times
an attribute’s value \( x_i \) appeared, \( \sum \text{count}(x_i) \) denotes the sum of the total number of each attribute’s value \( x_i \)
appearance.

Numeric attributes, such as the total amount of purchases made by a customer, are assumed to have Gaussian
distribution as underlying distribution model. They have continuous values. The parameters of the underlying
Gaussian distribution, the mean \( \mu \) and variance \( \sigma^2 \), are calculated using following formulas:

\[ \mu = \frac{w_1 x_1 + w_2 x_2 + \ldots + w_n x_n}{w_1 + w_2 + \ldots + w_n} \]
\[ \sigma^2 = \frac{w_1 (x_1 - \mu)^2 + w_2 (x_2 - \mu)^2 + \ldots + w_n (x_n - \mu)^2}{w_1 + w_2 + \ldots + w_n} \]

So,

\[ P(x_i) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x_i - \mu)^2}{2\sigma^2}} \]

where \( x_i \) denotes a value of the attribute, and \( w_i \) denotes the possibility of the attribute having value \( x_i \), and \( P(x_i) \) denotes
the possibility of the attribute having value \( x_i \).

As stated previously, the EM algorithm consists of four steps. We now explain in detail each of these steps and
also provide the pseudo-code algorithm to assist other researchers who may wish to replicate or implement the EM
algorithm.

Finding the best number of clusters to represent the observed data is not trivial. Generally, it is diffi-
cult to tell how many clusters best represent the observed data. Our implementation uses the likelihood to
evaluate the “goodness” of the number of clusters. First, begin with a cluster and then calculate the like-
lihood of each cluster and compares the mean of the likelihood of all clusters. Continue to increase the
number of clusters for the observed data until the likelihood decreases. The number of clusters just before
the last iteration is considered as the best number of clusters used by EM algorithm. Algorithm 2 shows
this step.

To improve the “goodness” of the likelihood for the clusters generated by the generic EM algorithm, we use a
10-fold cross validation. The 10-fold cross-validation randomly splits all instances in the observation data set into 10
subsets \( \{X_1, X_2, \ldots, X_{10}\} \) in approximately equal size. Nine of them are randomly chosen as training data \( \{Y_1, Y_2, \ldots, Y_9\} \)
and the remaining one is used as test data. The EM algorithm takes each training data set \( Y_i (1 \leq i \leq 9) \) to give a pair
of estimated parameters \( \theta_i(\mu_i, \sigma_i) \). For each pair of estimated parameters \( \theta_i \), this module uses the test data \( T \) to cal-
culate the corresponding likelihood \( L(\theta_i|T) \). Finally, the mean of all nine likelihoods for a given number of clusters
\( L_i = \left( \sum_{m=1}^{9} L(\theta_m T) \right) / 9 \) will be used to compare the previous mean of likelihood \( L_{i-1} \) to test if the best number of clusters
\( C \) for the observation data has been reached.

During the implementation test, the actual log likelihood \( L \) sometimes converges very slowly after a certain point, the
running time of this module then tends to be very long. For early convergence of the iteration to improve the efficiency
of the program after the likelihood gets stable, this module allows users to assign a number for the maximum number of
clusters (thus the maximum number of iterations) or assign the minimal difference between the current log likelihood
and the last one.
Algorithm 2. Estimate the number of clusters

```cpp
0 int function EstClusterNum ( ) {
1     boolean LkDecreased ← false;
2     numFolds ← 10;
3     num_cluster ← 1;
4     NUM_CLUSTER_LIMIT ← 0;
5     curr_likelihood ← - (MAX_DOUBLE_VALUE);
6     while (LkDecreased) {
7         LkDecreased ← false;
8         RandomShuffle (AllInstanceSet);
9         Subset[10] ← RandomSplit (AllInstanceSet);
10        testSetIndex ← Random (seed);
11        testSet ← subset[testSetIndex];
12        sum_likelihood ← 0.0;
13        for (i ← 0; i < numFolds; i ++) {
14            if (i == testSetIndex)
15                continue;
16            trainSet ← subset[i];
17            curr_likelihood ← iterateEM (trainset; num_cluster);
18            fold_likelihood ← E (testset, num_cluster);
19            sum_likelihood ← sum_likelihood + fold_likelihood;
20        }
21        mean_likelihood ← sum_likelihood / numFolds;
22        if (mean_likelihood > curr_likelihood &&
23            num_cluster < NUM_CLUSTER_LIMIT) {
24            curr_likelihood ← mean_likelihood;
25            LkDecreased ← true;
26            num_cluster ++;
27        }
28    }
29    return num_cluster - 1;
30 } end;
```

4.2 The EM iteration

The EM iteration module is the core of the EM cluster mining program. It iterates the E and M steps until the log likelihood of the data converges.

1. Our implementation uses a weighted instance approach to find the parameters to maximize the likelihood since each observed instance has a certain amount of membership for each class. So each instance correspondingly has different
weighted contribution to the statistics of every cluster. The EM iteration initially randomly guesses the weight \( w_i \) for each attribute of each instance for calculating the initial distribution models.

2. The EM iteration module utilizes Gaussian distributions in the M step (see Algorithm 5) with observed instance data to estimate parameters \( \theta_i(\mu_i, \sigma_i) \) for each attribute \( x_i \) in each cluster \( c_i \) for maximizing the overall likelihood, see Algorithm 3.

3. The E step, see Algorithm 4, calculates the probability of each instance \( x_i \) belonging to cluster \( c_i \) using the estimated parameters \( \theta_i \) from the M step. So the logarithm of overall likelihood \( \log(\text{likelihood}) = \sum_m \sum_n \log(p(x_m|c_n)) \).

4. The iteration starts step 2 again for next estimate \( \theta_i^{t+1} \) by the current \( \theta_i^t \) until reaching the maximum overall likelihood.

A maximum iteration number is also set to prevent endless loop when the likelihood does not converge.

**Algorithm 3.** The EM iteration

```cpp
/* ***************
 * random_guess() generates a random number between 0.0 and 1.0;
 * instance is a member in InstanceSet;
 * function E is shown in Algorithm 3;
 * function M is shown in Algorithm 4
 ****************/

0 function iterateEM (InstanceSet, num_cluster) {
  prev_likelihood ← 0.0;
  curr_likelihood ← 0.0;
  weight[instance][cluster] ← random_guess(seed);
  for (I ← 0; i < MAX_NUM_ITERATIONS; i++) {
    M(InstanceSet, num_cluster);
    prev_likelihood ← curr_likelihood;
    curr_likelihood ← E(InstanceSet, num_cluster);
    if((curr_likelihood - prev_likelihood) < MIN_DIFF)
      break;
  }
  return curr_likelihood;
}
```

**4.3 The E step**

The E step implementation calculates the log-likelihood. The reason for using logarithm of the likelihood instead of likelihood is that it simplifies the calculation to sum, avoiding heavy weight multiplication of double type data according to \( \log(x_1 \cdot x_2 \cdot \ldots \cdot x_n) = \sum_{i=1}^{n} \log x_i \).

1. The E step function first calculates the probability \( P(X = x_i | C = c_x) \) of each instance \( i \) belonging to a cluster \( c \) using

\[
P(X = x_i | C = c_x) = P(X = x_i) \cdot P(C = c_x), \text{ where } x_i = \{v_{1x}, v_{2y}, \ldots, v_{mx}\}
\]

2. This implementation assumes all attributes are independent of one another. So it uses equation \( P(X = \{v_{1x}, v_{2y}, \ldots, v_{mx}\}) = P(a_1 = v_{1x}) \cdot P(a_2 = v_{2y}) \cdot \ldots \cdot P(a_m = v_{mx}) \) to calculate the joint probability \( P(X = x_i) \) of instance \( x_i \). Given a value \( v_{1x} \) for attribute \( l \) of instance \( x_i \), the probability \( P(a_l = v_{1x}) = (1/\sqrt{2\pi}\sigma_l)e^{-(x_l-\mu_l)^2/2\sigma_l^2} \), where \( \mu_l \) and \( \sigma_l \) are part of the parameter set \( \theta'(\mu_l, \sigma_l) \) from the M step.

3. The probability of cluster \( C_x \) is given by \( P(C = C_x) = \sum_i P(x_i | C_x) / \sum_i \sum_j P(x_i | C_j) \) and it is implemented by function \( \text{estimate_cluster_probs()} \).

4. The probability of an instance \( x_i \) belonging to cluster \( c_i \), given by \( P(X = x_i | C = c_x) \), gets normalized and is used as the weight in the M step (see Algorithm 5) for calculating the parameters of the attributes distributions.

5. Compute the likelihood \( \log(\text{likelihood}) = \sum_m \sum_n \log(p(x_m | c_n)) \) and \( \text{lim}(\text{likelihood}) = 1 \).
Algorithm 4. The E step

```c
/* *****************************************************************************
 * function NORMALIZE() normalizes the values of elements of the input array;
 * function LOG() returns the logarithm value of the input;
 * function NormalDens() returns the probability of normal distribution with input of a
 * value, a mean, and a standard deviation.
*******************************************************************************/

function E (InstanceSet, total_cluster) { 
  log_likelihod ← 0.0;
  estimate_cluster_probs();
  for (instance ← 0; instance < total_instance; instance++) {
    for (cluster ← 0; cluster < total_cluster; cluster++) {
      joint_prob[instance] ← 1.0;
      for (attribute ← 0; attribute < total_attribute; attribute++) {
        joint_prob[instance] ←
        NormalDens(attr_value, mean, std_dev);
      }
      weight[instance][cluster] ←
      joint_prob[instance] * cluster_probs[cluster];
    }
  } 
  likelihood ← 0.0;
  for (i ← 0; i < total_cluster; i++) {
    likelihood ← likelihood + weight[instance][i];
  }
  log_likelihod ← log_likelihod + LOG(likelihood);
  // normalize the weight of each instance across all clusters 
  NORMALIZE (weight[instance]);
}

//

/*******************************************************************************/

* function estimate_cluster_probs() defined

*******************************************************************************/

//

function estimate_cluster_probs() {
  cluster_probs[ ] ← 0;
  for (instance ← 0; instance < total_instance; instance++) {
    for (cluster ← 0; cluster < total_cluster; cluster++) {
      cluster_probs[cluster] ←
      cluster_probs[cluster] + weights[instance][cluster];
    }
  } 
  NORMALIZE (cluster_probs);
}
```
Algorithm 5. The M step

```c
/* ******************************************
 * function SQRT( ) returns the square root of the input value.
 ******************************************/

0 function M (InstanceSet, total_cluster) {
1    cluster_models[ ] ← 0.0;
2        for (cluster ← 0; cluster < total_cluster; cluster++) {
3            for (attribute ← 0; attribute < total_attribute; attribute++) {
4                for (instance ← 0; instance < total_instance; instance++) {
5                    cluster_models[cluster][attribute].mean ←
6                        cluster_models[cluster][attribute].mean
7                            + instances(instance).value(attribute)
8                        * weights[instance][cluster];
9                    cluster_models[cluster][attribute].stdDev ←
10                        (cluster_models[cluster][attribute].stdDev
11                            + instances(instance).value(attribute)
12                        * weights[instance][cluster];
13                    cluster_models[cluster][attribute].mean ←
14                        (cluster_models[cluster][attribute].mean
15                            + cluster_models[cluster][attribute].mean
16                            + instances(instance).value(attribute)
17                        * weights[instance][cluster];
18                }
19            }
20        }
21    }
22 }

4.4 The M step

The M step, shown in Algorithm 5, computes parameter set $\theta_i^{t+1}$.

1. The mean value of attribute $a$ of cluster $c$ is calculated using the equation

$$
\mu_{ca} = \frac{\sum_{i=1}^{n} v_{ia} w_{lc}}{\sum_{i=1}^{n} w_{lc}}
$$
where \( l \) is a set of all instances, and \( l = \{1, 2, \ldots, n\} \), \( v_{ia} \) is the value of attribute \( a \) of instance \( i \), and \( w_{lc} \) is the weight of instance \( i \) in cluster \( c \).

1. Similarly, we compute the variance of attribute \( a \) of cluster \( c \)

\[
\sigma_{ca}^2 = \frac{w_{c1}(x_{a1} - \mu_{ca})^2 + w_{c2}(x_{a2} - \mu_{ca})^2 + \cdots + w_{cn}(x_{an} - \mu_{ca})^2}{w_{c1} + w_{c2} + \cdots + w_{cn}}
\]

\[
\Rightarrow \sigma_{ca}^2 = \frac{(w_{c1}x_{a1}^2 - 2w_{c1}x_{a1}\mu_{ca} + w_{c1}\mu_{ca}^2) + \cdots + (w_{cn}x_{an}^2 - 2w_{cn}x_{an}\mu_{ca} + w_{cn}\mu_{ca}^2)}{\sum_{i=1}^{n} w_{ci}}
\]

\[
\Rightarrow \sigma_{ca}^2 = \frac{\sum_{i=1}^{n} w_{ci}x_{ai}^2 + \sum_{i=1}^{n} w_{ci}\mu_{ca}^2 - 2\mu_{ca} \sum_{i=1}^{n} w_{ci}x_{ai}}{\sum_{i=1}^{n} w_{ci}}
\]

\[
\Rightarrow \sigma_{ca}^2 = \frac{\sum_{i=1}^{n} w_{ci}x_{ai}^2 + \sum_{i=1}^{n} w_{ci}\mu_{ca}^2 - 2\mu_{ca} \sum_{i=1}^{n} w_{ci}x_{ai}}{\sum_{i=1}^{n} w_{ci}}
\]

\[
\Rightarrow \sigma_{ca}^2 = \frac{\sum_{i=1}^{n} w_{ci}x_{ai}^2 + \sum_{i=1}^{n} w_{ci}\mu_{ca}^2 - 2\mu_{ca} \sum_{i=1}^{n} w_{ci}x_{ai}}{\sum_{i=1}^{n} w_{ci}}
\]

\[
\Rightarrow \sigma_{ca}^2 = \frac{\sum_{i=1}^{n} w_{ci}x_{ai}^2 + \sum_{i=1}^{n} w_{ci}\mu_{ca}^2 - 2\mu_{ca} \sum_{i=1}^{n} w_{ci}x_{ai}}{\sum_{i=1}^{n} w_{ci}}
\]

where instance \( i \) from 1 to \( n \).

To improve the efficiency of the EM algorithm, the following approaches may be adopted:

1. A threshold of the observing frequency can be set to limit the total number of attributes to be further analyzed when the actual number of attributes becomes very large.
2. Theoretically, the maximum value of likelihood is 0. In practical implementations, the increase of log-likelihood tends to be negligible after a very sharp change over the first few iterations. To run the program more efficiently, a minimal difference of the log-likelihoods between successive iterations is defined for early termination of the iteration whenever this condition holds.
3. It is necessary to preset a maximum iteration number to prevent endless loop when the likelihood does not converge.

### 4.5 Mining sequential pattern from web content log

Sequential patterns mining has been used in numerous studies. The main purpose of mining sequential patterns is to find popular patterns of events in time series from a database which contains sets of time ordered events. Discovered frequent sequential patterns show the common changes of events and attributes’ values over time. For e-commerce Web sites, logged customers’ browsing sequences and purchasing behaviors are stored in a database and each entry has a timestamp. Mining sequential patterns from Web log can help in the reorganization of the Web content to be more user friendly and efficient for target customers. Mining sequential access patterns from Web content usage data can help business analysts better understand visitors’ interests.

Using page view, we can distinguish the contents of different requests for same URL. A pageview \( p \) is defined as a response page requested by a user. Each pageview \( p \) contains a URL \( l_i \) and a set of content features \( \langle a_1, a_2, \ldots, a_n \rangle \). The URL \( l_i \) is user’s request URL and the content feature set includes all interesting features contained in the response page. Content features are defined and configured by the merchant’s site administrator to indicate interesting content to log for analysis.

The XML formatted data saved in Web content usage log is usually in a tree structure. To simplify the mining procedure, each attribute value in the page is mapped to a unique content id in a global dictionary. A sample XML data to Content ID mapping is shown in Table 1. In Table 1, each \(<\text{status}>\) belongs to a different \(<\text{product}>\), so they are considered as different attributes, and thus have different Content IDs. A specific combination of attributes can also be mapped to a content ID (if it is of interest to do so). For example, we can map a combination \(<\text{main}><\text{product}><\text{price}>499.99</\text{price}></\text{product}></\text{main}>\) to Content ID 10 to analyze this particular combination among user browsing sequences.

The size of logged XML content can be huge sometimes, which can result in huge size of content IDs, and thus, causes mining procedure to be inefficient. To improve the efficiency of the mining procedure, narrowing down the size of content IDs quickly is highly recommended. Designing different XSL stylesheets according to different business purposes gives a lot of control in subtracting interesting subsets in smaller sizes from the original tree structure.

After mapping a pageview \( p \) to a URL \( l \) and a vector of content IDs \( <1, 2, \ldots, n> \), each session can be represented in a simpler form, as shown in Figure 5. However, the content ID session database is not yet ready for access pattern sequential mining because the size of the content IDs is usually still too large to run the mining algorithm efficiently.
TABLE 1 A sample XML data to content ID mapping

| XML data         | Mapped content ID |
|------------------|-------------------|
| <product>        |                   |
| <sku>40,000–02-120</sku> | 1               |
| <status>In stock</status>   | 2               |
| <price>499.99</price>      | 3               |
| </product>       |                   |
| <product>        |                   |
| <sku>40,000–02-220</sku> | 4               |
| <status>In stock</status>   | 5               |
| <price>539.99</price>      | 6               |
| </product>       |                   |
| <product>        |                   |
| <sku>40,000–02-350</sku> | 7               |
| <status>In stock</status>   | 8               |
| <price>599.99</price>      | 9               |
| </product>       |                   |

FIGURE 5 A sample user session represented by content ID

In the Web content mining domain, when a transaction contains an item, an itemset, or a sequence, we say this transaction support that item, itemset, or sequence, respectively. The support, $supp$, for an item, an itemset, or a sequence is defined as the fraction of total number of transactions which contains this item, itemset, or sequence, respectively. According to Agrawal and Srikant, any subset of a large itemset or sequence must be large too. So by eliminating items below a minimum threshold support, the sequential patterns mining algorithms can work much more efficiently without impacting on the data mining results.

The original content ID session database is scanned to build a frequent content ID session database which only contains items that have higher supports than the minimal support defined.

Mining content access sequential patterns can be formally defined as follows: Let Content Set $C = \{c_1, c_2, \ldots, c_n\}$ be a set of all contents, and $supp(c_i) (1 \leq i \leq n) \geq SUPPORT_{min}$. The content in each page $p$ is represented by a page vector $p (c_a, c_b, \ldots, c_x), 1 \leq a < b < x \leq n$. The sequence of contents in each page can be ignored, and thus they are listed in alphabetical order. Each content $c_i$ can occur at most once in a page, but can occur multiple times across multiple different pages. Each vector is considered as an element in the sequence and the length of a sequence is the total number of elements.
We use the PrefixSpan algorithm,\textsuperscript{18} a high efficient frequent sequence mining algorithm, to mine Web content usage log to find long sequential access patterns with high support. We use the PrefixSpan algorithm because its innovative approach makes the algorithm much faster and highly efficient in mining large itemsets. Specifically, the PrefixSpan algorithm offers the following benefits:\textsuperscript{145} It eliminates the need for candidate generation, uses divide-and-conquer search methodology, the frequency of local items only countable, and it is superior to generalized sequential pattern.

The PrefixSpan algorithm is shown in Algorithm 6.

**Algorithm 6. PrefixSpan algorithm\textsuperscript{18}**

```plaintext
//***************************************************************************************
/*
* The PrefixSpan Algorithm [18]. See the cited reference for the details
*/
*******************************************************************************

main ( ) {
    total ← total number of frequent 1-sequential patterns discovered in step
2.
3. a[total] ← frequent 1-sequential patterns discovered in step 2.
4. S[total] ← Projected databases for each frequent 1-sequential pattern.
5. for (i ← 0; i < total; i++)
6.    prefixSpan (α[i], I, S[i]);
}
function prefixSpan (α; l, S) {
1.   B ← freqItems (S, α);
2.   for (i ← 0; i < B's size; i++)
3.       α' ← α + B[i];
4.       construct α'-projected database S';
5.       if (S' === null)
6.           print α';
7.       else
8.           prefixSpan (α', l + I, S);
9.
}
function freqItems (S, α) {
1.   Scan S once to find frequent items F;
2.   x ← last element in α;
3.   β ← x's prefix in α;
4.   for (i ← 0; i < F's size; i++)
5.       β ← F[i];
6.       if (β, (xb)> is a sequential pattern || <a,b> is a sequential pattern )
7.       Add β into Set F';
8.   return Set F';
9.
10. end;

A typical Apriori-based algorithm usually uses multiple-pass dataset scanning for counting candidates' supports generated from shorter frequent sequences. Such algorithms usually suffer several challenges while mining a large dataset. When the variation of elements in the dataset gets bigger, the size of the candidate sequence set increases exponentially because of the exhaustive search required to find all patterns. The number of subsequences that must be examined is $N(N-1)/2$ where $N$ is the length of the overall sequence. This brute force approach is both time consuming and space
inefficient. For example, for a sequence with 1000 elements, we need to examine 499,500 subsequences. Apriori-based algorithm also results in difficulties at mining long sequential patterns.

Generally, to improve the efficiency of mining large dataset, Apriori-based algorithm usually maps large itemsets (elements that contain large number of individual items) as single entities. By comparing two large itemsets for equality in constant time, it helps reduce the time required to check if a sequence is contained in a customer transaction. In real world, selecting items for each large itemset must be done very carefully if the user wants to have some level of details in the mining results while not losing much of the execution efficiency.

Recall, the PrefixSpan algorithm uses database projection for frequent sequences to make the candidates for next pass much less than the candidates generated by Apriori-based algorithms. This innovative approach makes the PrefixSpan algorithm much faster and highly efficient than the Apriori-based algorithms in mining large itemsets.

To reduce the support scanning space, the PrefixSpan algorithm divides the whole database into small sets which may lead to longer sequences by constructing frequent sequence prefix projected databases. Unlike the Apriori-like algorithms, the PrefixSpan algorithm does not generate any candidate sequence. Rather, PrefixSpan grows longer sequential patterns from the shorter frequent ones. Projected databases keep shrinking while the length of prefix growing. As the length of a prefix increases, the number of distinct frequent patterns with same prefix decreases dramatically.

The frequent content IDs are tokenized in the session access sequence database. Table 2 shows an example of Web contents in the session access sequence database.

The PrefixSpan algorithm uses database projection for frequent sequences to make the candidates for next pass much less than the candidates generated by Apriori-based algorithms. This innovative approach makes the PrefixSpan algorithm much faster and highly efficient than the Apriori-based algorithms in mining large itemsets.

To reduce the support scanning space, the PrefixSpan algorithm divides the whole database into small sets which may lead to longer sequences by constructing frequent sequence prefix projected databases. Unlike the Apriori-like algorithms, the PrefixSpan algorithm does not generate any candidate sequence. Rather, PrefixSpan grows longer sequential patterns from the shorter frequent ones. Projected databases keep shrinking while the length of prefix growing. As the length of a prefix increases, the number of distinct frequent patterns with same prefix decreases dramatically.

The frequent content IDs are tokenized in the session access sequence database. Table 2 shows an example of Web contents in the session access sequence database.

The PrefixSpan algorithm uses the following steps to mine frequent sequences.

Step 1: Find the supports of sequences with length = 1 and sorted in descending order. From the example in Table 2, a sequence only contributes once to its support, no matter how many times it appears in a whole sequence, see Table 3. By setting the $\text{Support}_{\text{min}} = 3$ as frequent sequence threshold, from Table 3 we will have frequent 1-sequences: $\langle a \rangle$, $\langle b \rangle$, $\langle c \rangle$, $\langle d \rangle$, $\langle e \rangle$, and $\langle f \rangle$.

Step 2: The 1-length frequent sequences are used as prefixes to partition the whole database into subsets. The example in Table 3 will have six subsets. Each has $\langle a \rangle$, $\langle b \rangle$, $\langle c \rangle$, $\langle d \rangle$, $\langle e \rangle$, or $\langle f \rangle$ as prefix correspondingly.

Step 3: Recursively mine projected database constructed by each partition of sequential patterns from Step 2. Find subsets of sequential patterns.

A projected database is formally defined as:

“Let $\alpha$ be a sequential pattern in a sequence database $S$. The $\alpha$-projected database, denoted as $S|_{\alpha}$, is the collection of postfix of sequences $S$ w.r.t. prefix $\alpha$”.

Let $\alpha$ and $\beta$ be two sequential patterns in sequence database $S$ such that $\alpha$ is a prefix of $\beta$. Then

1. $S|_{\beta} = (S|_{\alpha})|_{\beta}$;
2. For any sequence $\gamma$ having prefix, $\text{support}(\gamma) = \text{support}(\alpha | (\gamma))$
3. The size of $\alpha$-projected database cannot exceed that of $S$.

The PrefixSpan algorithm satisfies the assumption in the Apriori algorithm that “any subset of a large itemset must be large”.

| Seq_id | Sequence           |
|--------|--------------------|
| 1      | $\langle af \rangle(bc)(e)$ |
| 2      | $\langle eh \rangle(ab)(bh)(dhj)$ |
| 3      | $\langle ac \rangle(aef)(abc)(dhi)$ |
| 4      | $\langle bd \rangle(abd)(ac)(cf)(ac)$ |

| 1-Sequence | A | B | C | D | E | F | G | H | I | J |
|------------|---|---|---|---|---|---|---|---|---|---|
| Support    | 4 | 4 | 3 | 3 | 3 | 3 | 3 | 2 | 1 | 1 |
The main purpose of mining sequential patterns is to find popular patterns of events in time series from a database which contains sets of time ordered events. Frequent sequential patterns discovered show the common changes of events and attributes’ values over time. Finding evolving patterns of attributes among transactions in a time period and over time gives a better understanding of market trend. Sequential patterns mining can be used to analyze and predict market responses for promotions, seasonal movements, cyclic activities and other periodical patterns, and so forth. The discovery of sequential patterns also helps in target marketing aimed at groups of users based on these patterns. For e-commerce Web sites, logged customers’ browsing sequences and purchasing behaviors are saved in a database and each click has a timestamp. Mining sequential patterns from Web log can help to reorganize the Web content to be more user friendly and efficient for target customers. Web sites’ owners also can organize their business processes and logistics better to save costs once they have a better understanding of the current business processes and customers behaviors.

5 | EXPERIMENTAL RESULTS AND DISCUSSION

We used Java programming language to implement the EM clustering and the PrefixSpan sequential pattern mining algorithms, and DB2 as the backend database for the Web store, the Web content usage data raw log database, and the cleaned data repository (called data warehouse) used by the data mining programs. We used JDBC as the database driver in the Java programs to connect to the DB2 databases.

5.1 | Clustering

To demonstrate the ability of mining granular information on Web pages, the shopping cart in the checkout page is analyzed. This experiment finds direct relationship between the product unit price and quantity purchased.

The Web content log contains product number, quantity, unit price, and grand total price of each customer’s purchases. With an XML enabled DB2 database, a shopping cart DB2 document access definition (DAD) file helps parse the logged shopping cart content in XML format and builds a relational database view of the shopping cart items. Each row in this database view represents one item in a shopping cart. Each row contains a session id, a product number together with quantity purchased, unit price, and grand total price. Complex XML content parsing and data extraction reduces to very simple SQL query so as to minimize the data ETL effort.

The EM cluster mining program randomly splits the logged shopping cart data into ten data sets approximately equal in size and randomly selects one of these data sets as training data and the rest data as the testing data. The result shown in Table 4 reveals some interesting purchasing behaviors, such as a group of customers that usually purchases four units of the product priced at 649. There are about 5% such transactions among all transactions.

Table 5 shows clusters of product id and purchased quantity discovered. Cluster 2 is a cluster of products likely to be purchased four in quantity since among these products in Cluster 2, product 3030–360 has purchase price of 649.

5.2 | Web object sequential pattern mining

Our web object sequential pattern mining experiment demonstrates the ability of mining relationships in a sub-web page level. The granular information reveals the kinds of products visitors are interested in and willing to buy.

| Cluster | 1 | 2 | 3 | 4 | 5 |
|---------|---|---|---|---|---|
| Mean of unit price | 649 | 263.854 | 219 | 120.729 | 133.5844 |
| Std. of unit price | 0.0001 | 179.028 | 80 | 82.1117 | 80.4477 |
| Mean of quantity | 4 | 1 | 4 | 2.4063 | 2.4483 |
| Std. of quantity | 0 | 0 | 0 | 0.4911 | 0.4973 |
| Probability | 0.0528 | 0.7009 | 0.1877 | 0.0023 | 0.0545 |
| Support | 5% | 70% | 19% | 0% | 6% |
TABLE 5  Product ID and purchase quantity clusters

| Cluster | 1          | 2          | 3          | 4          | 5          |
|---------|------------|------------|------------|------------|------------|
| Mean of quantity | 2          | 4          | 3          | 1          | 1          |
| Std of quantity   | 0          | 0          | 0          | 0          | 0          |
| Product ID        | 3010–00005640 | 3030–315   | 3010–00005640 | 3010–00005640 | 3010–00005640 |
|                   | 3010–01640   | 3030–324   | 3010–01640   | 3030–030    |
|                   | 3030–00640   | 3030–360   | 3030–00640   | 3030–090    |
|                   | 3070–030     |           | 3030–315     | 3030–324    |
| Probability       | 0.0323      | 0.2405     | 0.0264      | 0.2312      | 0.4697      |
| Support           | 3%          | 24%        | 3%          | 0%          | 70%         |

TABLE 6  Sequential web object mining results

------- Pattern 1 -------
{[/myshop/servlet/Product:<ProductID="3010-00005640">} (0.2834) -
{[/myshop/servlet/Buy:<ProductID="3010-00005640">} (0.0472) -
{[/myshop/servlet/Update:<ProductID="3010-00005640">} (0.0472) -
{[/myshop/servlet/Bill:<ProductID="3010-00005640">} (0.0472)

------- Pattern 2 -------
{[/myshop/servlet/Product:<ProductID="3010-00005640">} (0.2834) -
{[/myshop/servlet/Buy:<ProductID="3010-00005640">} (0.0472) -
{[/myshop/servlet/Bill:<ProductID="3010-00005640">} (0.0472)

------- Pattern 3 -------
{[/myshop/servlet/Product:<ProductID="3010-00005640">} (0.2834) -
{[/myshop/servlet/Update:<ProductID="3010-00005640">} (0.0472) -
{[/myshop/servlet/Bill:<ProductID="3010-00005640">} (0.0472)

------- Pattern 4 -------
{[/myshop/servlet/Product:<ProductID="3000-1-01">} (0.1023) -
{[/myshop/servlet/Product:<ProductID="3000-1-01">}
{[/myshop/servlet/Product:<ProductID="3000-9-10">} (0.0472) -
{[/myshop/servlet/Product:<ProductID="3000-9-10">} (0.0393)

------- Pattern 5 -------
{[/myshop/servlet/Product:<ProductID="3010-00005003B">}
{[/myshop/servlet/Product:<ProductID="3000-1-01">} /{[/myshop/servlet/Product:<ProductID="3000-9-10">} (0.1338) -
{[/myshop/servlet/Product:<ProductID="3000-1-01">} /{[/myshop/servlet/Product:<ProductID="3000-9-10">} (0.0787) -
{[/myshop/servlet/Product:<ProductID="3000-1-01">} /{[/myshop/servlet/Product:<ProductID="3000-9-10">} (0.0393) -
{[/myshop/servlet/Product:<ProductID="3000-9-10">} (0.0314)

A generic product DAD file is used to extract product objects from XML formatted page content to build a product page database view. To make the sequential mining program run more efficiently, a Web object statistic program uses database queries to build a Web object statistic table, which holds URI_ID, Web Object name, Web Object ID, and occurrence information to eliminate low occurring Web objects from the sequential pattern mining. The original content log is transformed into a frequent Web object database table.

Table 6 shows the top five supported Web object sequential patterns discovered using the prefixSpan algorithm with minimal sequential pattern length of three.
TABLE 7  A web object sequential pattern sample

| Pattern | Pattern Popularity Ranking |
|---------|-----------------------------|
| /myshop/servlet/Product:<Product ID="3000-1-01"> | (0.1023) |
| /myshop/servlet/Product:<Product ID="3000-1-01"> | (0.0472) |
| /myshop/servlet/Product:<Product ID="3000-9-10"> | (0.0393) |

Each Web object, for example, /myshop/servlet/Product:<Product ID="3010-00005640">, has a URI and an object name separated by column. A pair of curly brackets {}, which contains one or more Web objects, represents a Web object set in a Web page. The number in the following parentheses is the support number of this Web object set. Table 7 illustrates these components of a Web object.

Manual inspection shows that the above frequent sequential patterns are among the products purchases and browsing threads with higher frequency in the simulation plan. This result is similar to the results established in Reference 146. Compared to Web page URI-based sequential pattern mining, the result of this experiment is straightforward to business analysts and the data ETL are simple and can be implemented quickly.

The results of the above experiments provide business analysts better understanding of customers’ interests for potential opportunities to increase sales, lower costs, and efficiently manage supply chains. Web store owners can quickly and flexibly extract attributes of interest to them from the source for further analysis according to their business purposes.

Mining Web content usage log can tell what items customers are interested in instead of ambiguous URLs. The results clearly establish a direct relationship between products purchases and their unit prices (i.e., product unit price and the quantity of the product purchased), as shown in Table 4. Two groups of customers usually buy four units of a product at a price of 649 and 219, respectively; accounting for 5% and 19% support respectively. The customers often buy products priced around 263–264 in 70% of the time – making it the most popular product (although each customer usually buys only one unit of the product) – while at 6% of the time customers buy products priced around 133–134 in multiples of 2.5 (≈3) units quantity. In our example, customers were not interested in products priced around 120–121 at all. This scenario reveals a lot of information about the merchant’s customers that would potentially help the merchant to be more effective in serving its customers. For example, the merchant could put items priced 649 and 219 on promotional sales to attract more patrons of these products, while it may eliminate from its inventory/stockholding products priced 120–121. Also, the merchant could marginally vary the price of the products priced around 263–264 since these products are very popular. From Table 5, products that are often purchased in multiple units are grouped together. For example, three products are likely candidates to be purchased in multiple of 4 units by customers. Therefore, based on this knowledge and many other previously undiscovered rules, the online merchant can maximize profits by optimizing its marketing campaigns through strategic reasoning and adoption of practical mechanisms. An example of such strategy is the non-availability of discount simultaneously on such products purchased in multiple of 4 units by customers.

Sequential association rules and time series models can be used to analyze usage data from a Web site taking into account a temporal dynamics using the site. Extracting association rules from the web content access data log is useful for obtaining correlations between the various pages visited during a browsing session. Also, application of association rules provides shopping basket analysis. For example, products that are purchased together or complementary products, but other items must be placed relative to a product in such a way to improve traffic in real or virtual store, creating the possibility that a buyer would also willing to buy other products seen as he/she browses.

Click-stream is a sequence of Web pages viewed by a user; pages are displayed one by one at a time. When a visitor accesses a website, the server retains all the actions taken by the visitor in a log file. A user session describes the sequence of pages viewed by the user during a period of logging on the web pages. A session (server session) represents all visited pages from a website within a user session, known as the visit. Each click of the mouse corresponds to a web page request; the sequence of clicks corresponds to such sequence links. Successive clicks flow analysis can be used to understand the possible way to navigate a web site for online prediction of the pages a visitor will likely access knowing the sequence of links (paths) that has followed the one before.

Experiments show that customer segmentation based on clustering results can improve the analysis efficiency of the relationship between customer groups and products, identify important factors affecting product sales, and provide support basis for enterprises to carry out customer-centered precision service. This supports the position clearly established...
in Singh\textsuperscript{147} to use data analytics for better decision making. Using our analysis to support precision marketing campaigns and offer better personalized services to customers reinforces and aligns with the theoretical results established in the literature\textsuperscript{120,121,124}

6 | CONCLUSION AND FUTURE WORK

An effective Web content usage logging system captures Web site owners’ interested attributes and relationships among these attributes on the Web pages that visitors request. Storing the captured attributes in structured format simplifies the data warehousing procedure for business intelligence applications while it provides great flexibility in choosing interesting attributes for different business analyses. This paper demonstrates the adoption of two existing popular data mining algorithms to discover business-oriented intelligence from the granularized logged attributes.

This paper uses the integrated and unified Web users’ personal profiles, business transactional data, and users’ visit log enterprise data to discover useful information and knowledge so that merchants can better understand and serve the needs of e-commerce users through the discovery and analysis of the relationship between product sales and customers’ behavior by mining the integrated data using two selected algorithms. The results of our analysis provide valuable insights into customers’ behaviors and content effectiveness, thus helping online merchants to build a more effective and more profitable e-commerce system. For example, important clues for analyzing sales data and visitors’ behavior data of specific marketing campaigns for business decision by the online merchants are derivable.

However, some limitations of current data mining algorithms exist:

- There are only few mature data mining algorithms that are capable of performing analysis across a large number of attributes. A typical e-commerce Web site sells hundreds or even thousands of products. In this paper, the logged attributes are first scanned to count the occurrences in the data preparation phase. Only attributes that have higher counts than a threshold are introduced to data mining algorithms. For most data mining algorithms, this approach does not affect the mining result but greatly improves the efficiency. However, for a company/merchant having many products, each product may have a number of related attributes, such as various costs, multiple inventories, and sales data, and so forth. The total number of interested attributes can still be very big. The Hypergraph partitioning data clustering algorithm\textsuperscript{148} can be considered to discover clusters among large amount of attributes. Presenting items and frequent item sets by vertices and hyperedges respectively, the Hypergraph partition algorithm is able to discover frequent items sets from a hypergraph efficiently in a large number of items.

- Most data mining algorithms assume attributes to be independent. So mining attributes with underlying dependency may lead to biased analysis results.

We plan to extend this paper in several directions; some are listed below:

- Implement a graphical user interface (GUI) presentation of data mining results. Currently, our implementation of the data mining algorithms run in command line mode and the results are shown in text format. A graphical presentation with support for impactful visualization tools will help users understand the mining results better.

- Investigate the application of additional data mining algorithms in Web object mining domain, such as decision trees, and neural networks, and so forth. The stored Web content usage data provides a rich set of source data in XML format for easy data access by data mining algorithms. To serve various business purposes, applying additional data mining algorithms to discover different types of patterns is necessary.

- Apply the same principles and algorithms used in this paper to patients data (in the medical domain) to check if any relationship exists between patients who were treated for the malaria disease and any other tropical diseases. This extension will be topical and particularly of interest to many African nations, such as Nigeria, whose health budget over the years have been dwindling and are looking for ways to manage the health care needs of their citizens efficiently.

Additionally, it would be interesting to use our approach on real e-commerce website content usage data for analysis to test the effectiveness of our selected algorithms. Presently, we are unable to this because of the unwillingness of major online merchants [we approached] to grant third parties access to their website content usage and transaction data, perhaps due to consumer data protection compliance requirements.\textsuperscript{40-43} We will continue to intensify our efforts to source real e-commerce website content usage data for analysis.
In summary, the Web content usage logging system provides an extensible infrastructure to capture e-commerce Web site visitors’ interests. The captured data is an excellent data source for further business intelligence analyses. The flexible structured data allows the logged data to be analyzed in many diversified ways at many granular levels.

Data may be mined for two principal purposes; either to help make more efficient decisions in the future or it may be used to explore the data to find interesting associative patterns. The association rules technique helps find interesting relationships (affinities) between variables (items or events). In e-commerce transactions data, association rules help figure affinities between products in transactions, showing stronger and weaker affinities. Thus, each rule has an assigned confidence level.

Cluster analysis, a useful unsupervised learning technique, is appropriate for e-commerce transactions data because it fits in situations where there is a large variety of transactions. It is used in many business situations, such as market research, to segment data into meaningful small groups. For example, in the e-commerce domain, customers are segmented into clusters based on their characteristics such as wants and needs, geography, price sensitivity, and so forth, and thus it can help identify natural groupings of customers, products, price, location, and so forth. Therefore, it helps to provide characterization, definition, and labels for populations. In addition, cluster analysis can also help identify outliers in a specific domain and thus decrease the size and complexity of problems.

ACKNOWLEDGMENT
We thank the anonymous review panel for their many helpful suggestions that enriched this final revised version of this article.

PEER REVIEW INFORMATION
*Engineering Reports* thanks the anonymous reviewers for their contribution to the peer review of this work.

DATA AVAILABILITY STATEMENT
The data that support the findings of this study are available from the corresponding author upon reasonable request. The data are not publicly available due to privacy or ethical restrictions.

CONFLICT OF INTEREST
The authors declare that there is no conflict of interests relevant to this article.

AUTHOR CONTRIBUTIONS
Sylvanus Ehikioya: Conceptualization; data curation; formal analysis; investigation; methodology; supervision; validation; writing—original draft; writing-review & editing. Jinbo Zeng: Formal analysis; investigation; methodology; software; writing—original draft.

ORCID
Sylvanus A. Ehikioya https://orcid.org/0000-0002-9743-2410

REFERENCES
1. Han J, Kamber M, Pei J. *Data Mining: Concepts and Techniques*. 3rd ed. Burlington, MA: Morgan Kaufman; 2011.
2. Witten IH, Frank E, Hall MA, Pal CJ, Mining D. *Practical Machine Learning Tools and Technologies*. 4th ed. Burlington, MA: Morgan Kaufman; 2016.
3. Mobasher B, Dai H, Luo T, Sun Y, Zhu J. Integrating web usage and content mining for more effective personalization. Paper presented at: Proceedings of the 1st International Conference on Electronic Commerce and Web Technologies, September 4–6, 2000, London - Greenwich, United Kingdom, Lecture Notes in Computer Science; Volume 1875, September 2000:165-176; Springer, Berlin, Heidelberg.
4. Kato H, Nakayama T, Yamane Y. Navigation analysis tool based on the correlation between contents distribution and access patterns. Paper presented at: Proceedings of the 6th ACM SIGKDDWebKDD Workshop on Web Mining for E-Commerce, Association for Computing Machinery, 2000:95-104; Boston, MA; New York, NY.
5. Kantardzic M. *Data Mining: Concepts, Models, Methods, and Algorithms*. 3rd ed. Hoboken, NJ: IEEE Press, John Wiley & Sons, Inc; 2020.
6. Srivastava J, Cooley R, Deshpande M, Tan P. Web usage Mining: discovery and applications of usage patterns from web data. *ACM SIGKDD Explor Newslett*. 2000;1(2):12-23.
7. S. Gomory, R Hoch, J. Lee, M. Podlaseck, E Schonberg, “E-Commerce Intelligence: Measuring, Analyzing, and Reporting on Merchandising Effectiveness of Online Stores. Technical Report”, IBM Institute for Advanced E-Commerce, IBM T.J. Watson Research Center, Yorktown Heights, NY: 2000. http://www.ibm.com/iac/papers/cabs3.pdf.
8. Liu B. Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data. 2nd ed. Berlin/Heidelberg, Germany: Springer; 2013.
9. Florin Gorunescu, Data Mining: Concepts, Models and Techniques, Intelligent Systems Reference Library, 12Springer-Verlag, Berlin/Heidelberg, Germany: 2011. DOI https://doi.org/10.1007/978-3-642-19721-5
10. Makhabel B, Mishra P, Danneman N, Heimann R. R: Mining Spatial, Text, Web, and Social Media Data: Create and Customize Data Mining Algorithms. Birmingham, UK: Packt Publishing Ltd; 2017.
11. Adiele C, Ehikioya SA. Evolving a “wise” integration system for E-commerce transactions. J Electron Commer Res Appl. 2007;6(2):219-232.
12. Ehikioya SA, Adiele C. A formal model of dynamic identification of correspondence assertions for E-commerce data integration. IJICIM. 2005;13(2):52-66.
13. Adiele C, Ehikioya SA. Algebraic signatures for scalable web data integration for electronic commerce transactions. J Electron Commer Res. 2005;6(1):56-74.
14. Ehikioya SA, Adiele C. Algebraic signatures to analyze correspondence assertions for web data integration. Int J Comput Inf Sci. 2009;10(1).
15. Hernandez M, Stolfo S. Real-world data is dirty: data cleansing and the merge/purge problem. J Data Min Knowl Discov. 1998;2(I):9-37.
16. Siepel A, Tolopko A, Farmer A, et al. An integration platform for heterogeneous bioinformatics software components. IBM Syst J. 2001;40(2):570-591.
17. Gupta MR, Chen Y. Theory and use of the EM algorithm. Found Trends Signal Process. 2011;4(3):223-296. https://doi.org/10.1561/2000000034.
18. J. Pei, J. Han, B. Mortazavi-Asl, et al. PrefixSpan: mining sequential patterns efficiently by prefix-projected pattern growth. Paper presented at: Proceedings of the 17th International Conference on Data Engineering (ICDE’01); April 2001:215-224; IEEE Computer Society, Washington, DC, Heidelberg, Germany, https://doi.org/10.1109/ICDE.2001.914830.
19. Ehikioya SA, Olukunle A. Mining of association rules in medical image data sets. J Digit Imaging Suppl. 2003;16(1):2-4.
20. Olukunle A, Ehikioya SA. A fast algorithm for mining association rules in medical image data. Paper presented at: Proceedings of the IEEE Canadian Conference on Electrical and Computer Engineering (CCECE’02) Conference Proceedings (Cat. No.02CH37373); vol. 2, 2002:1181-1187, Winnipeg, Manitoba, Canada https://doi.org/10.1109/CCECE.2002.1013116.
21. Ehikioya SA, Olukunle A. On the mining of association rules in medical image data sets. In: Callaos N, Zheng B, Kaderali F, eds. The 6th World Multiconference on Systemics, Cybernetics and Informatics (SCI-2002): Volume V – Computer Science I, July 14–18. Winter Garden, FL; Orlando, FL: International Institute of Informatics and Systemics (IIIS); 2002:17-22.
22. Mabroukeh NR, Ezeife CI. A taxonomy of sequential pattern mining algorithms. ACM Comput Surv. 2010;43(1):1-41.
23. Shanmugarajeshwari V, Lawrence R. An analysis of Teachers’ performance using classification techniques. Paper presented at: Proceedings of the 2017 IEEE International Conference on Intelligent Techniques in Control, Optimization & Signal Processing, March 23–25, 2017, Srikivilliputtur, INDIA; 2017, IEEE, Piscataway, NJ.
24. Lawrence R, Shanmugarajeshwari V. Analysis of students’ performance evaluation using classification techniques. Paper presented at: Proceedings of the IEEE International Conference on Computing Technologies and Intelligent Data Engineering (ICCTIDE’16); January 7–9, 2016:1-7; Kovilpatti, India, IEEE, Piscataway, NJ. https://doi.org/10.1109/ICCTIDE.2016.7725375,16426971.
25. Agaoglu M. Predicting instructor performance using data mining techniques in higher education. IEEE Transl Content Mining. 2016;4:2379-2387.
26. Asanbe OM, Ososihan OA, William FW. Teachers’ performance evaluation in higher educational institution using data mining technique. Int J Appl Inf Syst. 2016;10(7):10-15.
27. Ughade P, Mohod WS. A survey on analysis of faculty performance using data and opinion mining. Int J Innovat Res Comput Commun Eng. 2016;4(1):87-91.
28. Karthika N, Janet B. Clustering performance analysis. In: Chhabi BP, Panigrahi R, Buyya R, Li K-C, eds. Advanced Computing and Intelligent Engineering: Proceedings of ICACIE 2018, Volume I, Advances in Intelligent Systems and Computing. Vol 1082. Gateway East, Singapore: Springer Nature Singapore Pte Ltd; 2020. https://doi.org/10.1007/978-981-15-1081-6_3.
29. Kumar A, Bijoy MB, Jayaraj PB. Early detection of breast Cancer using support vector machine with sequential minimal optimization. In: Chhabi BP, Panigrahi R, Buyya R, Li K-C, eds. Advanced Computing and Intelligent Engineering: Proceedings of ICACIE 2018, Volume I, Advances in Intelligent Systems and Computing. Vol 1082. Gateway East, Singapore: Springer Nature Singapore Pte Ltd; 2020. https://doi.org/10.1007/978-981-15-1081-6_2.
30. Revathy R, Lawrance R. Comparative analysis of C4.5 and C5.0 algorithms on crop pest data. Int J Innovat Res Comput Commun Eng. 2017;5(1):50-58.
31. Abbas OA. Comparisons between data clustering algorithms. Int Arab J Inf Technol. 2008;5(3):320-325.
32. Rodriguez MZ, Comin CH, Casanova D, et al. Clustering algorithms: a comparative approach. PLoS One. 2019, pp. 1–34;14(1):e0210236. https://doi.org/10.1371/journal.pone.0210236.
33. Chitra K, Maheswari D. A comparative study of various clustering algorithms in data Mining. Int J Comput Sci Mob Comput. 2017;6(3):109-115.
34. Chen J. Comparison of Clustering Algorithms and Its Application to Document Clustering [PhD thesis]. Department of Computer Science, Princeton University; 2005.
35. Valarmathy N, Krishnaveni S. Performance evaluation and comparison of clustering algorithms used in educational data mining. Int J Recent Technol Eng. 2019;7(6S5):103-112.
36. Sehgal G, Garg K. Comparison of various clustering algorithms. Int J Comput Sci Inf Technol. 2014;5(3):3074-3076.
70. Montgomery AL, Li S, Srinivasan K, Liechty JC. Modeling online browsing and path analysis using clickstream data. *INFORMS Market Sci*. 2004;23(4):579-595. https://doi.org/10.1287/mksc.1040.0073.

71. Noreika A, Dražutis S. Website activity analysis model. *Inf Technol Control*. 2007;36(3):268-272.

72. Ehikioya SA, Lu S. A traffic tracking analysis model for the effective management of E-commerce transactions. *Int J Netw Distrib Comput*. 2020;8(3):171–193. https://doi.org/10.2991/ijndc.k.200515.006.

73. Fernandes RQA, Pinheiro WA, Xexéo GB, de Souza JM. Path clustering: grouping in an efficient way complex data distributions. *J Today's Ideas Tomorrow's Technol*. 2017;5(2):141-155. https://doi.org/10.15415/jottitt.2017.52004.

74. Lavanya B, Auxilia Princy A. A survey on contribution of data mining techniques and graph reading algorithms in concept map generation. *J Today's Ideas Tomorrow's Technol*. 2018;6(2):99-105. https://doi.org/10.15415/jottitt.2018.62009.

75. Ellonen H-K, Wikstrom P, Johansson A. The role of the website in a magazine business: revisiting old truths. *J Media Bus Stud*. 2015;12(4):238-249.

76. de Almeida Ribeiro JP. *The Use of Web Analytics on a Small Data Set in an Online Media Company: Shifter's Case Study* [Master's thesis]. Information Management, NOVA Information Management School, Instituto Superior de Estatística e Gestão de Informação, Universidade Nova de Lisboa; November 2016.

77. Lindén M. *Path Analysis of Online Users Using Clickstream Data: Case Online Magazine Website* [Master's thesis]. Strategy, Innovation and Sustainability, LUT School of Business and Management, Lappeenranta University of Technology; 2016.

78. Kumar Jain R, Kasana RS, Jain S. Efficient web log mining using doubly linked tree. *Int J Comput Sci Inf Secur*. 2009;3(1):40–44.

79. Pani SK, Panigrahy L, Sankar VH, Ratha BK, Mandal AK, Padhi SK. Web usage Mining: a survey on pattern extraction from web logs. *Int J Instr Control Automat*. 2011;1(1):15-23.

80. Agrawal R, Srivastava R. Mining sequential patterns. Paper presented at: Proceedings of the 11th International Conference on Data Engineering; 1995:3-14; IEEE Computer Society Press, Taipei, Taiwan; Washington, DC.

81. Agrawal R, Srivastava R. Fast algorithms for mining association rules. Paper presented at: Proceedings of the 20th International Conference on Very Large Data Bases VLDB'94, September 12–15, 1994:487-499; Santiago de Chile, Chile. Morgan Kaufmann, Burlington, MA.

82. Agrawal R, Imieliński T, Swami A. Mining association rules between sets of items in large database. Paper presented at: Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data SIGMOD’93; May 1993:207-216; Association for Computing Machinery, New York, NY, Washington, DC.

83. Agrawal R, Agrawal R. Mining sequential patterns: generalization and performance improvements. Paper presented at: Proceedings of the 5th International Conference in Extending Database Technology, Lecture Notes in Computer Science (LNCS); vol. 1057, 1996:3-17; Springer, Heidelberg, Avignon, France.

84. Jokar N, Honarvar AR, AgHamirzadeh S, Esfandiari K. Web mining and web usage mining techniques. *Bulletin de la Société Des Sciences de Liége*. 2016;85:321-328.

85. Lewis MW, White BJ. SOLO: a linear ordering approach to path analysis of web site traffic. *INFOR Inf Syst Operat Res*. 2012;50(4):186-194.

86. Ohta M, Higuchi Y. Study on the design of supermarket store layouts: the principle of “sales magnet”. *Int J Soc Behav Educ Econom Bus Ind Eng*. 2013;7(1):209-212.

87. Zheng L, Liu G, Yan C, Jiang C. Transaction fraud detection based on total order relation and behavior diversity. *IEEE Trans Comput Soc Syst*. 2018;5(3):796-806. https://doi.org/10.1109/TCSS.2018.2856910.

88. Privault N. *Understanding Markov Chains: Examples and Applications*. 2nd ed. Berlin/Heidelberg, Germany: Springer; 2018.

89. Berkhin P, Becher JD, Randall DJ. Interactive path analysis of web site traffic. Paper presented at: Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; August 2001:414-419; Association for Computing Machinery, San Francisco, CA, New York.

90. Chen M-S, Park JS, Yu PS. Efficient data mining for path traversal patterns. *IEEE Trans Knowl Data Eng*. 1998;10(2):209-221.

91. Mannila H, Toivonen H, Verkamo AI. Discovering frequent episodes in sequences. Paper presented at: Proceedings of the Eleventh ACM Conference on Recommender Systems RecSys ’17, Como, Italy, August 27–31, 2017, Association for Computing Machinery; August 2017:130-137; New York, NY. https://doi.org/10.1145/3109859.3109896.

92. Cooley R, Mobasher B, Srivastava J. Web mining: information and pattern discovery on the world wide web. Paper presented at: Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI’97); 1997:558-567; IEEE Computer Society, Washington, DC.

93. Sarwar B, Karypis G, Konstan J, Reidl J. Analysis of recommendation algorithms for E-commerce. Paper presented at: Proceedings of the 2nd ACM Conference on E-commerce (EC00), Minneapolis; October 2000; Association for Computing Machinery, New York, NY.
99. Schechter S, Krishnan M, Smith MD. Using path profiles to predict HTTP requests. *Comput Netw ISDN Syst.* 1998;30(1–7):457-467.

100. Kahya-Ozyirmidokuz E. Analyzing unstructured Facebook social network data through web text mining: a study of online shopping firms in Turkey. *Inf Dev.* 2014;32(1):70-80. https://doi.org/10.1177/026666914528523.

101. Chen CP, Weng J-Y, Yang C-S, Tseng F-M. Employing a data mining approach for identification of mobile opinion leaders and their content usage patterns in large telecommunications datasets. *Technol Forecast Soc Change.* 2018;130:88–98. https://doi.org/10.1016/j.techfore.2018.01.014.

102. Doja MN. Web data mining in E-services – concepts and applications. *IJCESE.* 2017;8(3):313-318.

103. Chajri M, Fakir M. Application of data mining in E-commerce. *Web Design and Development: Concepts, Methodologies, Tools, and Applications.* *Information Resources Management Association.* Hershey, PA: IGI Global; 2016:302-314. https://doi.org/10.4018/978-1-4666-6915-9.ch015.

104. Sharma MK, Vaisla KS. Data mining in e-commerce websites. In: Lather AS, Saini AK, Dhingra S, eds. *Business Intelligence and Data Warehousing.* New Delhi, India: Narosa Publishing House Pvt Ltd; 2012:150-153.

105. Landers RN, Brusso RC, Cavanaugh KJ, Collmus AB. A primer on theory-driven web scraping: automatic extraction of big data from the internet for use in psychological research. *Psychol Methods.* 2016;21:475-492.

106. Mitchell R. *Web Scraping with Python: Collecting More Data from the Modern Web.* 2nd ed. Sebastopol, CA: O’Reilly Media, Inc.; 2018.

107. Seppe vanden Broucke and Bart Baesens, *Practical Web Scraping for Data Science: Best Practices and Examples with Python.* Apress Media, Springer Science+Business Media (SSBM), New York, NY: 2018. https://doi.org/10.1007/978-1-4842-3582-9_1

108. Russell MA, Klassen M. *Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Google+.* 3rd ed. Sebastopol, CA: O’Reilly Media, Inc; 2018.

109. Munzert S, Rubba C, Meißner P, Nyhuis D. *Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining.* New York, NY: John Wiley & Sons; 2015.

110. Flory L, Osei-Bryson K-M, Thomas M. A new web personalization decision-support artifact for utility-sensitive customer review analysis. *J Decis Support Syst.* 2017;94(C):85-96. https://doi.org/10.1016/j.dss.2016.11.003.

111. Alkalbani AM, Gadhi L, Patel B, Hussain FK, Ghamry AM, Hussain OK. Analysing cloud services reviews using opinion mining. Paper presented at: 2017 IEEE 31st *International Conference on Advanced Information Networking and Applications* (AINA); March 27–29, 2017; Taipei, Taiwan: IEEE Computer Society, Washington, DC, USA. https://doi.org/10.1109/AINA.2017.173.

112. Alagukumar S, Venitha CDA, Lawrance R. Clustering of association rules on microarray gene expression data. In: Chhabib BP, Panigrahi R, Buyya R, Li K-C, eds. *Advanced Computing and Intelligent Engineering: Proceedings of ICACIE 2018, Volume 1, Advances in Intelligent Systems and Computing,* Vol 1082. Gateway East, Singapore: Springer Nature Singapore Pte Ltd. 2020. https://doi.org/10.1007/978-981-15-1081-6_8.

113. Vengatesh Kumar R, Alagukumar S, Lawrance R. Boolean association rule mining on microarray gene expression data. In: Chhabib BP, Panigrahi R, Buyya R, Li K-C, eds. *Advanced Computing and Intelligent Engineering: Proceedings of ICACIE 2018, Volume 1, Advances in Intelligent Systems and Computing,* Vol 1082. Gateway East, Singapore: Springer Nature Singapore Pte Ltd. 2020. https://doi.org/10.1007/978-981-15-1081-6_9.

114. Pandey S, Samal M, Mohanty SK. An SNN-DBSCAN based clustering algorithm for big data. In: Chhabib BP, Panigrahi R, Buyya R, Li K-C, eds. *Advanced Computing and Intelligent Engineering: Proceedings of ICACIE 2018 Volume 1, Advances in Intelligent Systems and Computing,* Vol 1082. Gateway East, Singapore: Springer Nature Singapore Pte Ltd.; 2020. https://doi.org/10.1007/978-981-15-1081-6_11.

115. Kharkongor C, Nath B. A survey on representation for itemsets in association rule mining. In: Chhabib BP, Panigrahi R, Buyya R, Li K-C, eds. *Advanced Computing and Intelligent Engineering: Proceedings of ICACIE 2018 Volume 1, Advances in Intelligent Systems and Computing,* Vol 1082. Gateway East, Singapore: Springer Nature Singapore Pte Ltd; 2020. https://doi.org/10.1007/978-981-15-1081-6_14.

116. Zhang F. The application of visualization technology on E-Commerce data mining Paper presented at: Proceedings of the 2008 2nd International Symposium on Intelligent Information Technology Application; December 20–22, 2008:563–566; IEEE Computer Society, Shanghai, China, Washington, DC. https://doi.org/10.1109/IITA.2008.18

117. Sever A. A problem of data mining in E-commerce. *Appl Math Comput.* 2011;217(24):9966-9970.

118. Jiang Y, Yu S. Mining E-commerce data to analyze the target customer behavior. Paper presented at: Proceedings of the 1st International Workshop on Knowledge Discovery and Data Mining. WKDD; 2008:406-409; IEEE Computer Society, Washington, DC. https://doi.org/10.1109/WKDD.2008.90.

119. Tang Y, Peng Z. A customer group mining method based on cluster analysis. In: Park J, Yang L, Jeong YS, Hao F, eds. *Advanced Multimedia and Ubiquitous Engineering* (MUE 2019), *FutureTech* 2019. *Lecture Notes in Electrical Engineering,* Vol 590. Singapore: Springer; 2020:351-357. https://doi.org/10.1007/978-981-32-9244-4_50.

120. Yim S, Pan H. Application of big data to precision marketing in B2C E-commerce. In: Atiquzzaman M, Yen N, Xu Z, eds. *Big Data Analytics for Cyber-Physical System in Smart City.* *BDCCPS* 2019. December 28–29, 2019, Shenyang, China, *Advances in Intelligent Systems and Computing,* Vol 1117. Singapore: Springer; 2020:731-738. https://doi.org/10.1007/978-981-15-2568-1_100.

121. Xu G, Shen Q. On the application of large data technology in B2C E-commerce precision marketing mode. In: Atiquzzaman M, Yen N, Xu Z, eds. *Big Data Analytics for Cyber-Physical System in Smart City.* *BDCCPS* 2019. 28–29 December 2019, Shenyang, China, *Advances in Intelligent Systems and Computing,* Vol 1117. Singapore: Springer; 2020:1197-1203. https://doi.org/10.1007/978-981-15-2568-1_166.

122. Erevelles S, Fukawa N, Swayne L. Big data consumer analytics and the transformation of marketing. *J Bus Res.* 2016;69(2):897-904.

123. Hongyan L, Zhenyu L. E-commerce consumer behavior information big data mining. *Int J Database Theory Appl.* 2016;9(7):135-146. https://doi.org/10.14257/ijdta.2016.9.7.12.

124. Arti SC, Purohit GN. Role of web mining in E-commerce. *Int J Adv Res Comput Commun Eng.* 2015;4(1):251-253.
125. Kumar V, Ogunmola GA. Web analytics for knowledge creation: a systematic review of tools, techniques, and practices. *IJCBPL*. 2020;10(1):1-14. https://doi.org/10.4018/IJCBPL.2020010101.

126. *Web Analytics Software*. Arlington, VA: Capterra Inc; 2019. https://www.capterra.com/web-analytics-software/. Accessed January 10, 2020.

127. Kimball R, Merz R. *The Data Webhouse Toolkit: Building the Web-Enabled Data Warehouse*. 1st ed. Indianapolis, IN: John Wiley & Sons, Inc; 2000.

128. *Web Design and Development: Concepts, Methodologies, Tools, and Applications*, Information Resources Management Association. Hershey, PA: IGI Global; 2016.

129. Arshdeep Bahga and Vijay Madisetti, *Big Data Analysis: A Hands-on Approach*, Bahga & Madisetti, Georgia: 2019. www.hands-on-books-series.com).

130. Pliego A, Marquez FPG. Chapter 12: big data and web intelligence: improving the efficiency on decision making process via BDD. *Big Data: Concepts, Methodologies, Tools, and Applications*, Information Resources Management Association (Editor), *Information Science Reference*. Hershey, PA: IGI Global; 2016:229-246. https://doi.org/10.4018/978-1-4666-9840-6.ch012.

131. Inmon WH. *Building the Data Warehouse*. 4th ed. Indianapolis, IN: John Wiley & Sons, Inc; 2005.

132. Kimball R, Ross M. *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling*. 3rd ed. Indianapolis, IN: John Wiley & Sons, Inc.; 2013.

133. Kimball R. *The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data*. Indianapolis, IN: John Wiley & Sons, Inc; 2004.

134. Ponniah P. *Data Warehousing Fundamentals for IT Professionals*. 2nd ed. Hoboken, NJ: John Wiley & Sons, Inc; 2010.

135. Alejandro Vaisman and Esteban Zimanyi, *Data Warehouse Systems: Design and Implementation*, Springer-Verlag. Berlin/Heidelberg, Germany: 2014. https://doi.org/10.1007/978-3-642-54655-6

136. Sommerville I. *Software Engineering*. 10th ed. New York, NY: Addison-Wesley; 2015.

137. Extensible Markup Language (XML) 1.0 (Fifth Edition). W3C recommendation 26, November 2008. World wide web consortium.Org http://www.w3.org/TR/xml/. Accessed April 29, 2018.

138. Deitel HM, Deitel PJ, Nieto TR, Lin TM, Sadhu P. *XML: How to Program*. Upper Saddle River, NJ: Prentice Hall; 2001.

139. Adamson C. *Star Schema: The Complete Reference*. New York, NY: McGraw-Hill Education; 2010.

140. Adamson C. *Mastering Data Warehouse Aggregates: Solutions for Star Schema Performance*. Indianapolis, IN: John Wiley & Sons, Inc; 2006.

141. Levene M, Loizou G. Why is the snowflake schema a good data warehouse design? *Inf Syst*. 2003;28:225-240.

142. Firdaus S, Uddin A. A survey on clustering algorithms and complexity analysis. *Int J Comput Sci Issues*. 2015;12(2):62-85.

143. Nisbet R, Miner G, Yale K. *Handbook of Statistical Analysis and Data Mining Applications*. 2nd ed. San Diego, CA: Academic Press, Elsevier; 2018.

144. Allen AO. *Probability, Statistics, and Queuing Theory with Computer Science Applications*. 2nd ed. New York, NY: Academic Press; 1990.

145. Pratik Saraf R, Sedamkar R, Rathi S. PrefixSpan algorithm for finding sequential pattern with various constraints. *Int J Appl Inf Syst*. 2015;9(3):37-41.

146. Zeng J. *Data Warehousing for Electronic Commerce* [Master thesis]. University of Manitoba, Winnipeg, MB, Canada; 2004.

147. Harshdeep Singh, “*Using Analytics for Better Decision-Making*”, Toronto, Ontario, Canada: Towards Data Science, 2018. https://towardsdatascience.com/using-analytics-for-better-decision-making-ce4f92c4a025. Accessed January 10, 2020.

148. Han E, Karypis G, Kumar V, Mobasher B. Clustering based on association rule hypergraphs. Paper presented at: Proceedings of the Workshop on Research Issues on Data Mining and Knowledge Discovery (DMKD 1997) in cooperation with ACM SIGMOD’97, Tucson, Arizona, USA, May 11, 1997:15-22; Association for Computing Machinery, New York, NY.

How to cite this article: Ehikioya SA, Zeng J. Mining web content usage patterns of electronic commerce transactions for enhanced customer services. *Engineering Reports*. 2021;e12411. https://doi.org/10.1002/eng2.12411