Received: 30 June 2022 | Revised: 30 September 2022 | Accepted: 25 October 2022 | Published online: 25 October 2022

REVIEW

Artificial Intelligence and Applications
2023, Vol. 1(1) 3–10
DOI: 10.47852/bonviewAIA2202290

BON VIEW PUBLISHING

Data Mining Techniques for Web Mining: A Survey

Mehdi Gheisari[1,2,3], Hooman Hamidpour[4], Yang Liu[2,*], Peyman Saedi[5], Arif Raza[6], Ahmad Jalili[7], Hamidreza Rokhsati[8] and Rashid Amin[9]

[1]Department of Computer Science and Technology, Islamic Azad University, Iran

[2]Department of Computer Science and Technology, Harbin Institute of Technology, China

[3]Department of Cognitive Computing, Institute of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, India

[4]Department of Computer Engineering and Information Technology, Shiraz University of Technology, Iran

[5]Oracle DBA at Behestan Rayan Hamrah, Iran

[6]School of Computer Science and Software Engineering, Shenzhen University, China,

[7]Department of Computer Engineering, Faculty of Basic Sciences and Engineering, Gonbad Kavous University, Iran

[8]Department of Computer, Control and Management Engineering, Sapienza University of Rome, Italy

[9]Department of Computer Science, University of Chakwal, Pakistan

Abstract: The data mining (DM) is the computational process that consists of searching, extracting, and analyzing patterns in large data sets, including methods at the intersection of artificial intelligence, machine learning, statistics, and database schemes. Specifically, its primary goal is to extract information from a raw data set and transform it into an expected structure for further use. Moreover, an evolving perspective of DM is web mining (WM), which refers to the whole of DM and related routines. It is used to discover and extract information from web records and services automatically, that is, WM's purpose is to obtain valuable data from the World Wide Web. Due to its importance, a survey about DM techniques in WM is necessary, as performed in this paper.

Keywords: social networks, web design, data mining, web mining, World Wide Web

1. Introduction

Rapid advances have been made in the procedures for data collection and generation. There was a time when data shortages were a significant issue, but that's no longer the case; the incapacity in producing data that provides valuable information is. Therefore, DM has become an increasingly important field of research (Chakrabarti et al., 2014; Piatetsky-Shapiro & Frawley, 1991; Roche, 2006) To put it simply, the process of "Mining" is extracting treasured and raw materials from the earth, such as metal and coal. The term "data mining" (DM) has been referred to in Computer Science as the primary analysis, data and database management, selecting the inference models, complexity tolerances, interesting metrics, and visualization (Medvedev et al., 2017). Data mining techniques convert bulk data sets into valuable structures for future use (Chakrabarti et al., 2014; Madni et al., 2017), that is, data are not the result of the extraction process in the case of DM. Also, DM is defined as the process of finding hidden meaning in a database. It is also known as exploratory data analysis, data-driven discovery, and deductive learning (Zhang & Segall, 2008). In contrast, DM results consist of knowledge and patterns that we derived from the extraction process (Komarek, 2004).

DM is mainly used for analyzing large volumes of data in a semi-automated/automated manner to find hidden functional patterns like unusual records (also called anomaly detection), cluster analysis (collection of records), and dependency (mining rule association). Subsequently, these patterns can work as the input data's summary (Nural et al., 2015; Tan et al., 2013). Significant amounts of data are free to use on online websites, blogs, knowledge-sharing sites, groups, and email systems (Safari et al., 2020). Each month, about 30 billion pieces of content, like blog posts, URLs, photos, and news, are shared by Facebook users. Similarly, the amount of per-day tweets of Twitter users exceeds 155 million (Statista). These businesses work as data collection platforms and create application programming interfaces to allow other organizations to access the results.

*Corresponding author: Yang Liu, Department of Computer Science and Technology, Harbin Institute of Technology, China. Email: Liu.Yang@hit.edu.cn

Business is not the only field that uses DM. The medical field is one of the fields that DM had a significant impact on by recognizing appropriate practices for helping patients (Sinha, 2013). IBM had a contribution with Mayo Clinic in the development of an online system for responding to specific treatments in the list of the last 100 Mayo Clinic patients with the same characteristic and conditions (Jalili et al., 2019).

Researchers have linked social networking data analysis at a novel scale to obtain highly accurate models in an imperative study (Sapountzi & Psannis, 2018). Deriving data from the Internet can also create problems for people, like misuse of their data. The knowledge discovery within DM is necessary for all types of applications that include devising and administration (Romero & Ventura, 2020). There are some personal privileges of web DM, such as enhancing search engine aptitude. Besides, DM performed on the web data can be applied for marketing goals by analyzing web users' habits and turning them into marketing experiences (Rabiner, 1989). Web DM is helpful to businesses for various purposes. For example, to decide who may be a new client by analyzing consumer data, government documents, and other helpful information (Al-Sultan, 1995). It can help improve values by selling more goods or services or cost optimization in a general sense. To claim this, marketing intelligence is needed. This information can concentrate on selling strategies and competitive examination (Jalili et al., 2018). Different kinds of web data related to clients are then classified and grouped to produce accurate customer profiles. This not only helps businesses keep current clients capable of providing more personalized services, but it also helps to search for potential clients. Therefore, it is exceeding discussion that DM tools can be pretty valuable for the business. There are several distinct ways to mine the web for structural analysis, and we need to differentiate between these. We can differentiate between accurate data on web pages, detailed web composition information about web documents, and web login information about users who browse web pages.

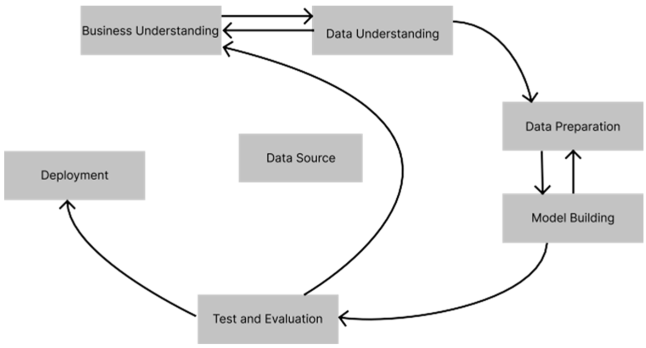
2. Overview of DM

Web DM is used by DM technology in web applications. The ever-increasing amount of data is an outstanding field for DM studies. Web mining (WM) is a combination of DM and related techniques used to automatically identify and extract information from web documents. The data center is being built into use to store vast volumes of data. The bulk of commercial databases has had a significant effect on the organization's need for DM. DM allows the company to respond to the problems that can occur in an event.

DM process has six phases:

- The first phase is business understanding in which analyzing DM goals, determining business objectives, and assessing the current situation.
- The second phase is data understanding in which, after business understanding, data are required that the concept of data understanding takes into account data requirements.
- The third step is data preparation; when initial data are recognized, the data preparation phase starts to select, clean, and integrate into the desired form and format.
- The fourth step is modeling; when a deeper understanding of the data is achieved, it becomes possible to apply more detailed models appropriate for the data type. For modeling, it is also necessary to separate training data from test data.
- Evaluation is the fifth step; during the first phase, business objectives should be established, guiding the evaluation of model results.
- Deployment is the final step of DM and can be used to either confirm or discover previously held hypotheses. Figure 1 illustrates these six steps.

Figure 1
Phases of data mining



DM is gaining popularity in both the private and public sectors. DM is widely used in industries such as accounting, insurance, and retail to cut prices, boost experimentation, and increase sales. DM use was first used in the public sector to track abuse and misuse and was extended to monitor and increase program efficiency.

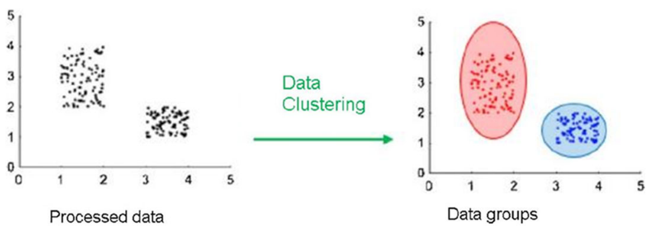
3. Information Extraction Algorithms

The different types of site data associated with clients would then be categorized and grouped to create comprehensive client profiles (Ashourian et al., 2015; Moghadasi et al., 2020; Sidana & Aggarwal, 2017). DM approaches are classified into two types: direct and indirect. The direct approach is used in prediction, and it examines the known values to predict a step with a new value. The second approach, the non-direct approach, identifies new patterns by looking at the past subjects.

(1) Cluster – The cluster approach aims to cluster data points with similar properties into clusters that could be together. It can be used to distinguish the correlation between intersex features. Once new data are introduced, its features can be charted with relationships, and it can be used to predict new data's function. Clusters can be used to detect data irregularities as well. It is generally used in fraud detection and client relation control systems. The feature of the clustering approach could start without any previous hypotheses in the data of the clusters or beginning from a particular hypothesis that was reached in past research. Figure 2, and Figure 9 illustrate how decision boundaries could divide more similar inputs into a single cluster.

Using statistical metrics such as the Bayesian information criterion, one measures how well a set of clusters fits the data compared to how much fit might be expected by chance given the number of clusters.

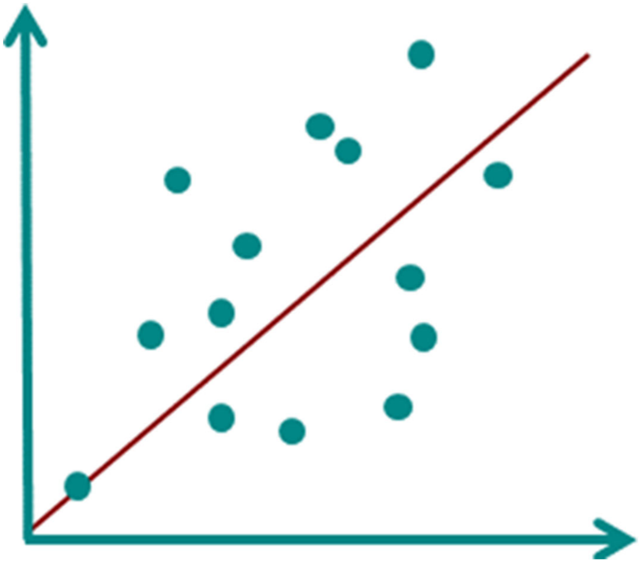
Figure 2
Schematic of clustering



(2) Regression – This algorithm has a duty to check a given data set and predict the numerical values. It is the process of identifying and analyzing the relationship between variables in relation to the other factor through the use of DM. Regression has five different types:

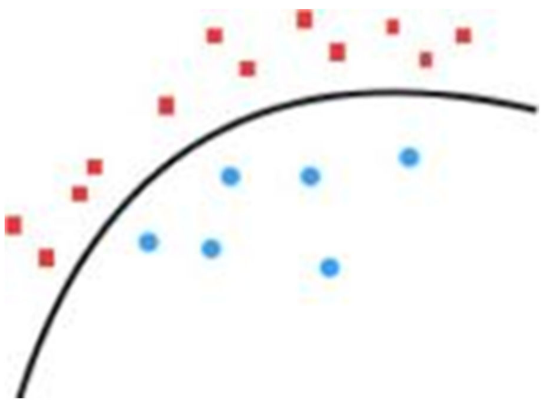
(a) Linear regression – This prediction only predicts continuous variables using the single linear regression formula and estimates a conditional value(s) based on previous observations. (Figure 3 shows how a decision boundary in regression could classify inputs.)

Figure 3
Simple linear regression with least square errors



(b) Logistic Regression – A neural network without a hidden layer is used in this algorithm. In comparison to linear regression, we can make an estimation for non-linearly separable functions (McDonald, 2009). Figure 4 illustrates the shape of a Logistic Regression which as seen could be more flexible then a Linear Regressing.

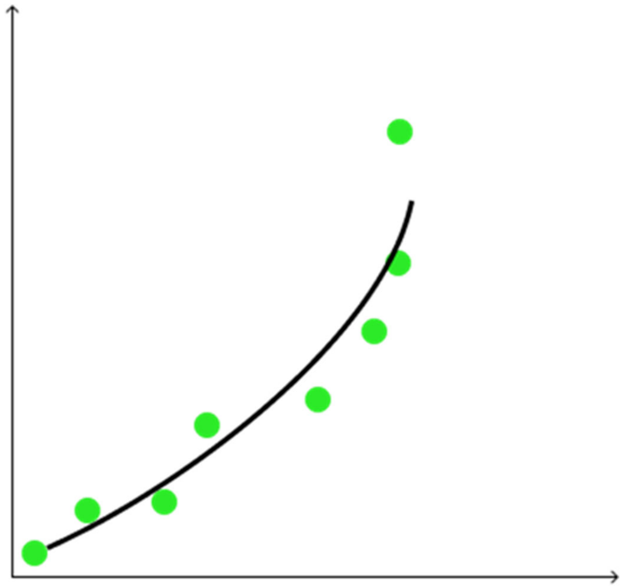
Figure 4
Schematic of logistic regression



(c) Polynomial regression – Basically, polynomial regression is a linear regression in which a relationship between independent variables x and y is modeled by a polynomial of nth order that has x as the regular attributes and y as the attribute used to predict y (Olson & Shi, 2007). (See Figure 5)

(d) Ridge regression – This type uses for analyzing issues of multicollinearity in different regression data. The occurrence of a linear correlation between two independent variables is known as multicol linearity (Dunham, 2002).

Figure 5
Schematic of polynomial regression



(e) Lasso regression – Lasso regression is a linear regression technique that employs shrinkage. All data points are reduced toward a center point, commonly known as the mean, in Lasso regression.

(3) Neural networks – This algorithm was inspired by artificial intelligence. Neural networks are one type of model for machine learning and have become relatively competitive with conventional regression and statistical models regarding usefulness (Weiss & Indurkhya, 1998). Neural networks' full applications can be evaluated with respect to data analysis factors such as accuracy, processing speed, latency, performance, fault tolerance, volume, scalability, and convergence (Jalili, 2019; Fayyad et al., 1996). Neural networks can be developed and used for image recognition, natural language processing, and so on. Nowadays, neural networks are mostly used for universal function approximation in numerical paradigms because of their excellent properties of self-learning, adaptivity, fault tolerance, nonlinearity, and advancement in input to output mapping (Piatetsky-Shapiro & Frawley, 1991). As Figure 6 abstractly illustrates, two inputs colored in green could impact the value of an output.

Figure 6
Schematic of a simple single layer neural network perceptron

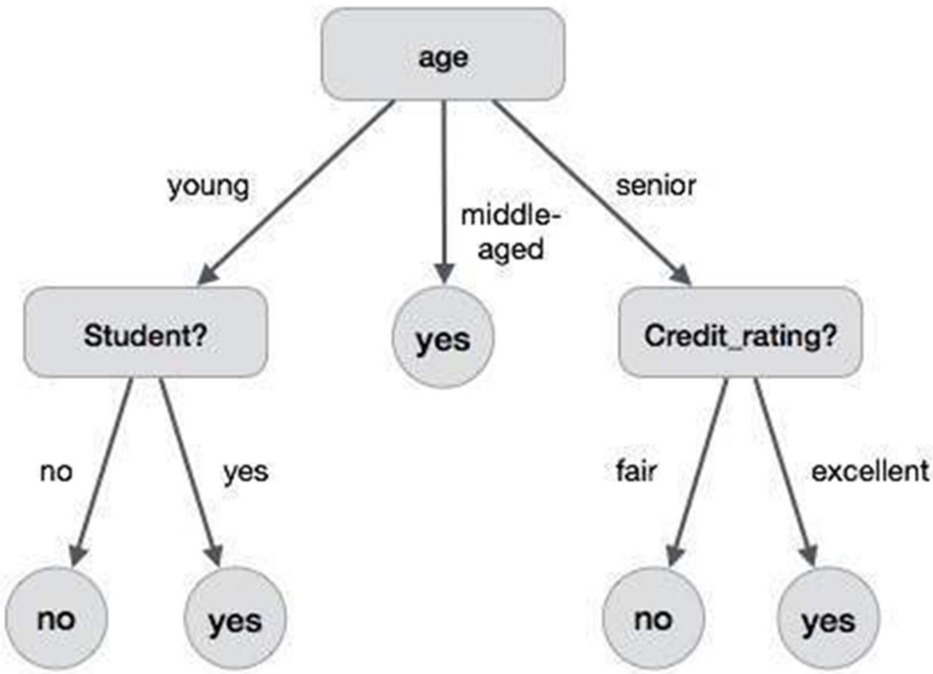
(4) Decision trees – It is a simple and widely used tool. It is used in the prediction of both discrete and continuous variables. It uses a tree data structure as a predictive model. It maps observations about an item to pre-stored or nearest neighbor. (Bonner, 1964).
(5) Naive Bayes – This classifier is a simple probabilistic classifier based on applying Bayes' theorem with strong (naive) independence assumptions. When the output of a function is estimated, it can be used to quantify probabilities for each state given an input feature. This can be used as an algorithm originating from the forecasting method (Bonner, 1964).

4. DM Techniques

Several DM approaches have recently been developed and used in DM projects, such as classification, clustering, prediction, sequential patterns, and decision trees (Kosala & Blockeel, 2000; Zhang & Segall, 2008).

(1) The Association is a well-known DM tool. A trend is defined in Association based on a similarity between particular objects in the subsequent behavior that is used to predict patterns. That is why the connection approach is often referred to as the link technique. For example, someone who goes to the store and purchases a mobile phone is predisposed to buy a Cell Phone Screen Protector.
(2) Classification is a standard machine learning-based DM technique. Classification is the method of categorizing each object in a data collection into a predefined set of groups. In classification, we develop software that can learn how to classify unseen data records so that other unclassified data in the learning set can be predicted automatically by the classification methods.
(3) Clustering is another DM approach that uses an automated method to create valuable clusters of objects with identical properties. The clustering approach represents classes and assigns objects to each category, while grouping assigns objects to predefined categories. In other words, the clustering process is unsupervised and does not need a learning set.
(4) As the title says, prediction is a DM technique that examines the relationship between dependent and independent variable features to predict output(s).
(5) Sequential pattern is a DM methodolgy, which attempts to find or classify similar patterns, routine occurrences, or variations in transaction data over a business cycle in a sequential pattern analysis. Several similar sequences can be studied at the DM stage to identify future transaction trends. When dealing with databases with time-series characteristics, this approach is helpful.
(6) Because of its simple model, the decision tree is one of the most commonly used DM techniques. The decision tree root is a basic query or situation with several responses in the decision tree system. Figure 7 illustrates how a simple trained Decision tree from a dataset could decide for any input.

Figure 7
Schematic of a generated decision tree from data source



5. A Glimpse of WM

WM is coined for the first time by Etzioni (Saura et al.). A hypothesis helped him to start this definition that most data on the web have sufficient structures. According to Etzioni, WM is a significant area of DM in the manner of explaining and discovering patterns in worldwide web data. In other words, WM techniques are used to find interesting and potentially helpful information within web data. It can be described as automatically searching for data from multiple origins because of data's heterogeneous and unstructured nature on the web, and information extraction may be challenging. Using the data mentioned above, direct extraction algorithms are not practical. Web data miner tools like web crawlers do not precisely do web extraction; they extract text and are not able to extract information or knowledge from it (Saura et al., 2017).

The considerable volume of content available on the Internet and especially on the web is a valuable knowledge source. Unfortunately, these data are usually unstructured and full of inappropriate and unusable; discovering clean and well-organized is a challenge. Web text mining (Kasemsap, 2017) extracts information from natural language origins of web data, such as web records, web users, and other web data properties, using DM techniques to purify and make a clean data set for further steps.

6. DM Algorithms

DM methods are helpful in social network analysis using graph search algorithms. The World Wide Web, including social networks, is a set of interconnected hypertext documents. So the web can be seen as a bidirectional graph, where the node will be hypertext and edge documents will be links. Web structure analysis based on graph algorithms has been analyzed in many research methods in past years (Ashourian et al., 2015; Castelo-Branco et al., 2020; Fayyad et al., 1996; Jalili, 2019; Kasemsap, 2017; Moghadasi et al., 2020; Patel & Rana, 2014; Sidana & Aggarwal, 2017; Vidhya & Aghila, 2010; Weiss & Indurkhya, 1998). Patel and Rana (2014) have designed and tested linking network, semantic, data processing, transformation, and visualization-based segments. They introduced a new mining algorithm graph, the "mining period graph or discovering all the interaction patterns that occur at regular intervals," considering social networks' dynamic behavior. The algorithm is to extract repeated patterns in trading databases and graphs by periodic extraction of patterns in one-dimensional and multidimensional series.

Castelo-Branco et al. (2020) introduced a framework in discrete dynamic graphs and graph clusters. This framework can identify dynamic switching from social networks' structure, identify time dimension analysis incidents, and expose it to the command hierarchy.

DM algorithms could classify into classification, clustering, semantic web and ontology, and Markov mode algorithms.

7. Classification

The classification method is one of the many algorithms. It can be applied in WM data to classify data points such as user profiles based on display features. It is a common technique in DM to separate data points into different categories. Data sets of all sizes can be organized, including complex and large data sets as well as smaller and simpler ones.

The most basic and straightforward classification approach is user based. Unless the software is used to categorize the data, it depends on someone manually categorizing the data. It is up to the end-user to select and categorize the information while relying on their own knowledge and discretion in handling it. During data classification, training data is fed to an algorithm as the first step in the process. Following the development of classification rules, the algorithm will analyze the test data and produce new results based on the rules.

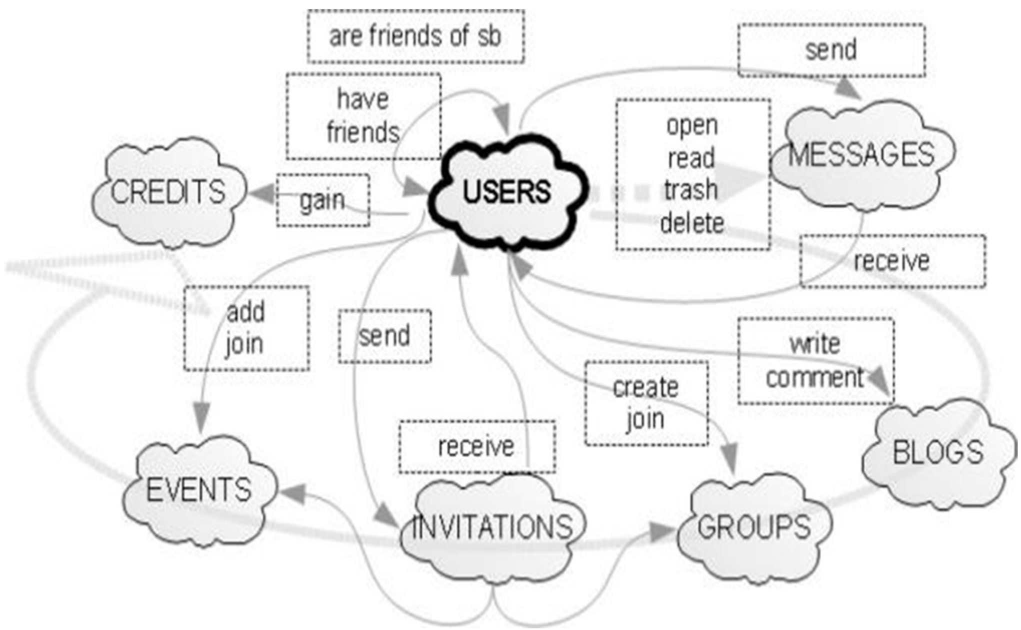
A new algorithm presented by Kasemsap (2017) is called a classification and regression tree (CART), which is a predictive model that explains how an outcome variable's value can be predicted based on the values of other variables. CART outputs are decision trees, with every fork being a predictor variable and every node containing a prediction for the outcome variable. It can be used in actual social networking data. As Figure 8 shows performed steps by an actor (user in this case) could shape a predictive model.

8. Clustering

The clustering of an object is as ancient as the human being itself. The need to describe the noticeable characteristics of humans and objects by identifying (labeling) with the purpose of categorizing unseen objects into a previously known group or recalling previously categorized instances. Therefore, it embraces various scientific disciplines: from mathematics and statistics to biology and genetics, each of which uses different terms to describe the topologies formed using this analysis. From biological "taxonomies" to medical "syndromes" and genetic "genotypes" to manufacturing "group technology," the problem is identical: forming categories of entities and assigning individuals to the proper groups within it (Ashourian et al., 2015; Azim et al., 2019; Kaur & Garg, 2018; Phyu & Thu, 2021; Sidana & Aggarwal, 2017; Wang et al., 2018;).

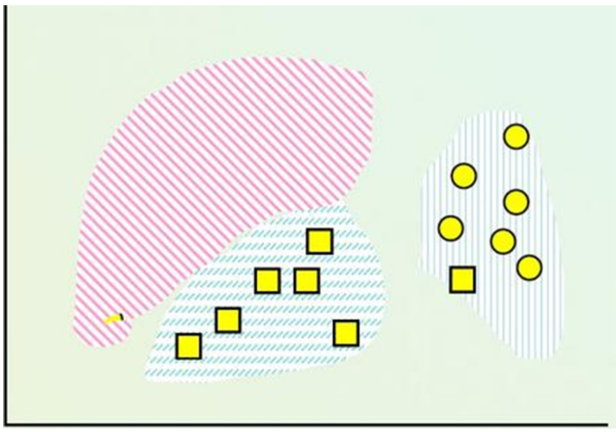
The cluster itself is a subset of records in which the elements in the same group have more similarity to one another than to those

Figure 8
A map of possible actions taken by a user (Dave & Dutta, 2014)



in other clusters. (See Figure 2, 9 as an example) Which is referred to as a cluster such that the distance between any two records in the cluster is less than the distance between any record in the cluster and any record not in it. Since clustering is the grouping of similar instances/objects, some sort of measure that can determine whether two objects are similar or dissimilar is required. There are two main types of measures used to estimate this relation for various attribute types (numeric, binary, nominal, ordinal, and mixed type), which is out of the scope of this paper: distance measures and similarity measures. Clusters are mainly used in data recovery in WM. Based on past research on clusters, productivity in data recovery increases. Clusters based on diagrams are pleasantly explained in the WM structure used as described in the initial section (Moghadasi et al., 2020).

Figure 9
Another sample of clustering



9. Semantic Web and Ontology

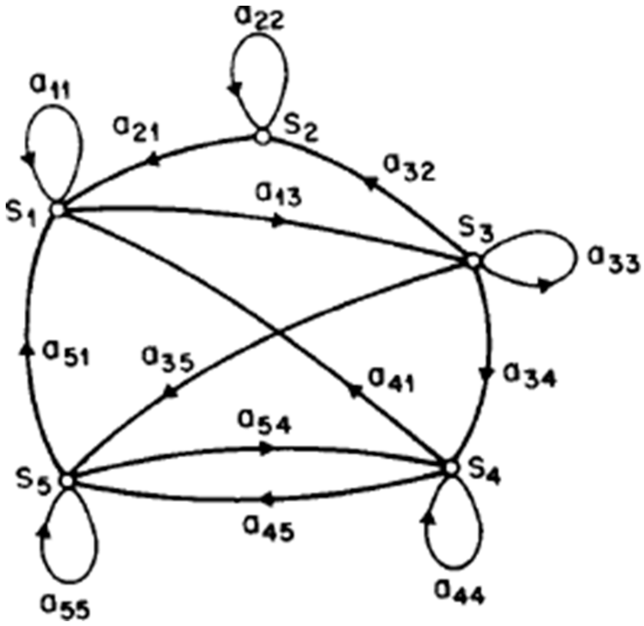
The semantic web is a recent academic area that attempts to assign context to the web. This enables intelligent communication and information sharing between computers and humans. More research has been undertaken in this field, including using geographic semantic catalogs and retrieving mental health information (Sidana & Aggarwal, 2017).

Weiss and Indurkhya (1998) proposed a new method using semantic similarity measures based on predefined ontology for classified social network data. Moshayedi et al. (2019a) have developed an algorithm for recovering data in social networks to identify trends.

10. Markov Model

The Markov chain is a mathematical algorithm that undergoes developments from one state to another, among the limited states (Gheisari & Esnaashari, 2017; Moshayedi et al., 2019b, 2022). This is a stochastic method in which the next term depends only on the current state and not on the sequence of events that have happened before. Markov models can be used to predict users' next steps. Therefore, in node information using Markov, we can predict the next site the person may visit. As Figure 10 example illustrates, in a sample Markov model with N = 5 states, Markov chain is capable of assigning a probability to each transition between states.

Figure 10
Simple visualization of Markov chain with 5 states (labeled S_1 to S_6) with selected state transitions (He & Garcia, 2009)



11. Conclusion and Future Work

The DM is fascinating, game-changing, and capable of supporting a wide variety of applications. Without gathering, purifying, and analyzing data, it is impossible to make conclusions or detect hidden patterns and relationships in unstructured and/or semi-structured problems. Various DM methodologies can generate innovation by analyzing a vast collection of data. To utilize traditional DM methods, data storage must be collected in an organized form with a minimum amount of data loss. Although DM can identify exciting patterns, traditional DM techniques are not expedient because of the lack of structure. Further research would help concentrate more on mining data where many individual behavior patterns can be identified by analyzing the social network and making use of different methods in action. We can adopt statistical methods like Markov models to solve data's temporary behavior from the web. This is helpful to the marketing and advertisement fields of marketing websites developing methods in the worlds' business to minimize their cost, redesign their marketing, sales, and customer-related service like recommendation systems, QoS, etc. In this paper, we have had a survey about DM techniques in the Web environment, its application, frameworks, and so on. We plan to have a closer look at its application on recommender systems in the future.

Acknowledgement

This work was supported by Shenzhen Stable Supporting Program (General Project) (No. GXWD20201230155427003-20200821160539001) and Shenzhen Basic Research (General Project) (No. JCYJ20190806142601687). Moreover, it is supported by Islamic Azad University, Iran

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

References

Al-Sultan, K. S. (1995). A Tabu search approach to the clustering problem. *Pattern Recognition*, *28*, 1443–1451. https://doi.org/10.1016/0031-3203(95)00022-R

Ashourian, M., Gheisari, M., & Talkhoncheh, A. H. (2015). An improved node scheduling scheme for resilient packet ring network. *Majlesi Journal of Electrical Engineering*, *9*, 43–50. http://mjee.iaumajlesi.ac.ir/index/index.php/ee/article/view/1649

Azim, M. A., Abdelhamid, A. A., Badr, N., & Tolba, M. (2019). Large vocabulary Arabic continuous speech recognition using tied states acoustic models. *Asian Journal of Information Technology*, *18*, 49. https://doi.org/10.36478/ajit.2019.49.56.

Bonner, R. (1964). On some clustering techniques. *IBM Journal of Research and Development*, *8*, 22–32. https://doi.org/10.1147/rd.81.0022

Castelo-Branco, F., *et al.* (2020). Business intelligence and data mining to support sales in retail. In *Marketing and smart technologies* (pp. 406–419). Springer.

Chakrabarti, S., Ester, M., Fayyad, U., Gehrke, J., Han, J., Morishita, S., Piatetsky-Shapiro, G., & Wang, W. (2014) Data Mining Curriculum: A Proposal (Version 1.0). ACM SIGKDD. 30 April 2006. Retrieved 27 January 2014.

Dave, V. S., & Dutta, K. (2014). Neural network-based models for software effort estimation: A review. *Artificial Intelligence Review*, *42*, 295–307. https://doi.org/10.1007/s10462-012-9339-x

Dunham, M. H. (2002). *Data mining introductory and advanced topics*, Pearson Education.

Fayyad, U., Djorgovski, S. G., & Weir, N. (1996). *Advances in knowledge discovery and data mining*. MIT Press (pp. 471–494).

Gheisari, M., & Esnaashari, M. (2017). A survey to face recognition algorithms: advantageous and disadvantageous. *Journal of Modern Technology and Engineering*, *2*, 57–65. https://www.researchgate.net/publication/317033497_A_SURVEY_TO_FACE_RECOGNITION_ALGORITHMS_ADVANTAGEOUS_AND_DISADVANTAGEOUS

Gheisari, M., Wang, G., Bhuiyan, M. Z. A., & Zhang, W. (2017). Mapp: A modular arithmetic algorithm for privacy preserving in IoT. In *2017 IEEE international symposium on parallel and distributed processing with applications and 2017 IEEE international conference on ubiquitous computing and communications (ISPA/IUCC)* (pp. 897–903). IEEE.

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, *21*, 1263–1284. https://doi.org/10.1109/TKDE.2008.239

Jalili, A. (2019). A new SDN-based framework for wireless local area networks. *International Journal of Nonlinear Analysis and Applications*, *10*, 177–183. https://doi.org/10.22075/IJNAA.2019.4062

Jalili, A., Keshtgari, M., & Akbari, R. (2018). A new set covering controller placement problem model for large scale SDNs. *Information Systems & Telecommunication*, *25*. http://jist.ir/fa/Article/15071

Jalili, A., Keshtgari, M., Akbari, R., & Javidan, R. (2019). Multi criteria analysis of controller placement problem in software defined networks. *Computer Communications*, *133*, 115–128. https://doi.org/10.1016/j.comcom.2018.08.003

Kasemsap, K. (2017). Mastering web mining and information retrieval in the digital age. *Web usage mining techniques and applications across industries* (pp. 1–28). IGI Global.

Kaur, J., & Garg, K. (2018). Efficient management of web data by applying web mining pre-processing methodologies. In Advances in Intelligent Systems and Computing.

Komarek, P. (2004). *Logistic regression for data mining and high-dimensional classification*. Carnegie Mellon University.

Kosala, R., & Blockeel, H. (2000). Web mining research: A survey. *SIGKDD: SIGKDD Explorations: Newsletter of the Special Interest Group (SIG) on Knowledge Discovery and Data Mining*, *2*, 1–15. https://doi.org/10.1145/360402.360406

Madni, H. A., Anwar, Z., & Shah, M. A. (2017). Data mining techniques and applications—A decade review. In 2017 *23rd International conference on automation and computing (ICAC)*. IEEE.

McDonald, G. C. (2009). Wiley interdisciplinary reviews: Computational statistics. *Wiley Online Library*, *1*, 93–100.

Medvedev, V., *et al.* (2017). A new web-based solution for modeling data mining processes. *Simulation Modelling Practice and Theory*, *76*, 34–46. https://doi.org/10.1016/j.simpat.2017.03.001

Moghadasi, M. N., Safari, Z., & Zhuang, Y. (2020). A sentimental and semantical analysis on Facebook comments to detect latent patterns. In *2020 IEEE international conference on big data (Big Data)*. IEEE.

Moshayedi, A. J., *et al*, (2019a) Portable image based moon date detection and declaration: system and algorithm code sign. In *2019 IEEE international conference on computational intelligence and virtual environments for measurement systems and applications* (CIVEMSA) (pp. 1–6). https://doi.org/10.1109/CIVEMSA45640.2019.9071604.

Moshayedi, A. J., *et al.* (2022). Sunfa Ata Zuyan machine learning models for moon phase detection: algorithm, prototype and performance comparison. *TELKOMNIKA (Telecommunication, Computing, Electronics and Control)*, *20*, 129–140. https://doi.org/10.12928/TELKOMNIKA.v20i1.22338.

Moshayedi, A. J., *et al*, (2019b). Kinect based virtual referee for table tennis game: TTV (Table Tennis Var System). In *Proceedings - 6th IEEE international conference on control system, computing and engineering, ICCSCE 2016* (pp. 354–359). https://doi.org/10.1109/ICISCE48695.2019.00077.

Nural, M. V., Cotterell, M. E., & Miller, J. A. (2015). Using semantics in predictive big data analytics. In *2015 IEEE international congress on big data*. IEEE.

Olson, D. L., & Shi, Y. (2007). *Introduction to business data mining*, Boston: McGraw-Hill/Irwin.

Patel, B. R., & Rana, K. K. (2014). A survey on decision tree algorithm for classification. *IJEDR*, *2*. https://www.ijedr.org/papers/IJEDR1401001.pdf

Phyu, A. P., & Thu, E. E. (2021). Short survey of data mining and web mining using cloud computing. International Journal of Advanced Networking and Applications, 12, 4725–4731. https://www.ijana.in/download12-5-9.php?file=V12I5-9.pdf

Piatetsky-Shapiro, G., & Frawley, W. J. (1991). *Knowledge discovery in database*. AAAI/MIT Press.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, *77*, 257–286. https://doi.org/10.1109/5.18626.

Roche, T. G. (2006). Expect increased adoption rates of certain types of EHRs. *EMRs. Managed Healthcare Executive 16*, 58.

Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 10*, e1355. https://doi.org/10.1002/widm.1355

Safari, Z., Mursi, K. T., & Zhuang, Y. (2020). Fast automatic determination of cluster numbers for high dimensional big data. In *Proceedings of the 2020 the 4th international conference on computer and data analysis*.

Sapountzi, A., & Psannis, K. E. (2018). Social networking data analysis tools & challenges. *Future Generation Computer Systems*, *86*, 893–913. https://doi.org/10.1016/j.future.2016.10.019

Sidana, A. and Aggarwal, H. (2017). Review of web usage of data mining in web mining. *International Journal of Advanced Research in Computer Science*, *8*, 2742–2746. http://www.ijarcs.info/index.php/Ijarcs/article/view/4128

Sinha, P. (2013). Multivariate polynomial regression in data mining: methodology, problems and solutions. *International Journal of Scientific and Engineering Research*, *4*, 962–965. https://www.ijser.org/paper/Multivariate-Polynomial-Regression-in-Data-Mining-Methodology.html

Saura, J. R., Palos-Sánchez, P. R., & Suárez, L. M. C. (2017). Understanding the digital marketing environment with KPIs and web analytics. *Future Internet*, *9*, 76. https://doi.org/10.3390/fi9040076

Swartz, N. (2004). IBM, Mayo clinic to mine medical data. *The Information Management Journal 38*, 8. https://go.gale.com/ps/i.do?id=GALE%7CA125489568&sid=googleScholar&v=2.1&it=r&linkaccess=abs&issn=15352897&p=AONE&sw=w

Tan, P.-N., Steinbach, M., & Kumar, V. (2013). Data mining cluster analysis: basic concepts and algorithms. In *Introduction to data mining* (pp. 487–533).

Twitter global international mDAU 2022, Statista, https://www.statista.com/statistics/1032751/monetizable-daily-ctive-twitter-users-international/ (accessed Mar. 14, 2021)

Vidhya, K. & Aghila, G. (2010). A survey of Näive Bayes machine learning approach in text document classification. *IJIST*, *7*. https://doi.org/10.48550/arXiv.1003.1795

Wang, D., He, H., & Liu, D. (2018). Intelligent optimal control with critic learning for a nonlinear overhead crane system. *IEEE Transactions on Industrial Informatics*, *14*, 2932–2940. https://doi.org/10.1109/TII.2017.2771256

Weiss, S. H., & Indurkhya, N. (1998). *Predictive data mining: A practical guide*, Morgan Kaufmann Publishers, San Francisco, CA.

Zhang, Q., & Segall, R. S. (2008). Web mining: A survey of current research, techniques, and software. The International Journal of Information Technology and Decision Making, 7, 683–720. https://doi.org/10.1142/S0219622008003150

How to Cite: Gheisari, M., Hamidpour, H., Liu, Y., Saedi, P., Raza, A., Jalili, A., Rokhsati, H., & Amin, R. (2023). Data Mining Techniques for Web Mining: A Survey. *Artificial Intelligence and Applications* 1(1), 3–10, https://doi.org/10.47852/bonviewAIA2202290