An Intermediate-Level Introduction to Web Mining

Kooper Young, Nikitha Vedant Madabhushi, Divya Shree Moka, Bharath Kumar Reddy Mopuru, Susmitha Muppalla, Manohar Sri Vikram Vattikuti

Northwest Missouri State University, Maryville MO 64468, USA
S537651@nwmissouri.edu

Abstract. This paper provides an overview of web mining, which is the process of discovering patterns and extracting useful information from large data sets of web data. The paper discusses the three most common techniques of web mining: web content mining, web structure mining, and web usage mining. Web content mining involves extracting useful information from a specific web page or series of web pages, including text, images, links, videos, and audio. Web structure mining analyzes the structure of the web, focusing on how web pages relate and link to each other. Web usage mining analyzes the behavior of people who use the web, including how users navigate through websites or pages and what pages they visit. The paper also discusses the practical applications of web mining in various domains, including e-commerce, education, healthcare, and social media. The uses of web mining range from improving web surfing experiences and personalizing web content and layout to analyzing student behavior in online learning environments. The paper concludes by highlighting the potential for web mining to continue to provide valuable insights into user behavior, web content, and web structure.

Keywords: *Web Mining · Web Mining Utility · Web Scraping · O.I.E · Information Extraction ·*

1 Introduction

In recent years, the growth of the internet as a whole has led to it becoming embedded in almost all of our daily lives. We have access to more information than anyone else has ever had in history, and it is all at our fingertips. However, it is also dense, overwhelming, and complex. The internet, World Wide Web, and the information contained within needs analysis and examination to be utilized properly by the average person. Web mining, which is the process of discovering patterns and extracting useful information from the World Wide Web, has emerged as a popular research area due to its potential to provide insights into user behavior, web content, and web structure. Web mining has applications in various domains, including e-commerce, education, healthcare, and social media. The following sections of the paper provide an overview of the concept of web mining, specific uses of web mining, some practical and accessible and others more complex, as well as where web mining might take us in the future if the current trends are any indication.

2 Definition

The term "web mining" can refer to any number of specific utilizations of a broader process, but in general it is often defined as: "The process of discovering patterns and extracting useful information from large data sets of web data". This definition encompasses most of what web mining is but lacks context. Today, when someone uses the term "web mining" they are probably referring to one of

three specific areas of web mining: Web Content Mining, Web Structure Mining, or Web Usage Mining. Other areas exist, but these three are the most common.

Web content mining is probably the most common technique, and is usually what most people would think of when discussing web mining. Web content mining is the process or technique of extracting information deemed useful from a specific web page or series of web pages themselves. This data is taken from the literal elements of web pages, and can include things like text, images, links, buttons, videos, audio, and other forms of multimedia, but usually primarily consists of text, either formatted (normal) or raw (HTML). Things like natural language processing (NLP) and text mining fall within web content mining. [8]

Slightly askew from web content mining, we find web structure mining, the second technique. Web structure mining focuses on analyzing the structure of the web. In this case, structure would mostly relate to how web pages relate and link to each other, or how different pages flow and connect to each other. Web structure mining also focuses on the structure of the entire web, not just pockets and individual domains. Links and navigation are analyzed in order to build a better idea of how things are laid out, as well as understand patterns and trends. [6]

Coming to the most abstract technique, we have web usage mining. I say abstract because web usage mining, as the name implies, is more concerned with how the information on the internet is navigated and interacted with, than the actual information itself. Web usage mining typically involves analyzing the behavior of people who use the web. This can include many things such as how users might navigate through websites or pages, what pages users visit and how long they stay, as well as any actions a user might make while on a web page.

These techniques of web mining provide us with valuable information, necessary to maintaining and improving the experience of the internet's billions of daily users. It has many more practical applications such as in the healthcare industry, e-commerce, social media analysis, customer profiling, online learning, and web personalization.

3 Uses and Utility

As discussed previously, web mining has numerous applications, some of which may surprise you. Starting with something simple and benign, web mining can improve your web surfing experience and improve the efficiency of your browser. Web mining excels at identifying behavioral patterns. This strength can be applied to the internet habits of a user. By tracking pages visited, search terms used, and language and grammar, your preferenced and interests can be predicted with surprising accuracy. Taken a step further, this information can be used to personalize the content and layout of web pages. This brings a unique quirk to a web browsing experience, in that, the web can personalize and adapt itself to you over time, eventually increasing efficiency and comfortability in the best cases. [2]

A bit more significant and slightly more complex is web mining in the context of an online learning environment. In any learning environment, it is important to understand the habits and needs of the students, so that efficient learning takes place. An online learning environment opens up a new way to find, analyze, and support these needs, which is of course - web mining. With web mining, we can gain incredible insights into student behavior, assignment submissions, common trends, participation rates, completion times, forum posts, and so on. This insight is invaluable for the purpose of identifying areas of concern and areas of excellence. With that kind of information, we can make online learning and teaching better with each iteration. The sky is the limit, essentially.[1]

Clearly, web mining excels at information gathering, but the utility of such a thing isn't limited to an online environment. It can be used, to a great effectiveness in some cases, for monetary gain. Market research comes to mind. You see, aside from the fact that web mining is just good at and designed for gathering information, it specifically excels in handling large amounts of information, quickly. This means it is geared towards things like competitive analysis, customer profiling, customer satisfaction or feedback, social media reviews, the effectiveness of marketing and advertisements, the list really goes on and on. The analyzed data can be used to make decisions and affect the future of a business. One real benefit of using web mining versus traditional analysis techniques is that web mining has the ability to identify patterns and trends that may not be immediately apparent otherwise, such as trends in how users navigate a website for a business. You could then refine the layout to direct customers where you wanted them to go, thus increasing efficiency. This is similar to subliminal messaging, in the way that the customer or user is being directed without even knowing it. [7] [5]

Perhaps at the height of utility for web mining lies the healthcare field. Web mining is currently being utilized for all sorts of things related to healthcare, such as tracking disease outbreaks. Web mining can also be used to analyze patient records and medical journals. It can be applied to current medical research and improve things like patient care and disease mechanisms through analysis and reflection. While the average healthcare worker is unlikely to know about web mining's involvement in their field, they certainly feel its impact. Healthcare centers and hospitals are cauldrons of information and ripe for data gathering and analysis. Detailed patient charts that correspond to a single individual are just one of many side effects of the consolidation and analysis of information [10]

At the peak of complexity in this paper, we have the Open Information Extraction system. This is an application of web mining that currently has little to no practical application, but it easily could within the next couple of years. Simply put, the Open Information Extraction system, or OIE system is a technique of web mining that can extract structured information from unstructured text on the web. It has been argued that the OIE system has numerous practical applications, such as building knowledge bases, answering natural language queries, and automating information integration. On top of that, the OIE system can be used to extract information from a variety of sources, including news

articles, product reviews, and scientific papers. Just like web mining the OIE system can handle the vast amounts of unstructured data on the web and output a structured organized summary, for lack of a better term. You see, web mining typically relies on a schema, which is a type of template or simplified base form. The problem is that lots of text do not naturally conform to predefined schemas, or even stick to a single one consistently. The OIE system completely circumvents this issue, because it is designed to work without a schema, and handle the natural unpredictability of human-generated text. As a consequence, the OIE system can be used to identify new relationships or patterns that were already not identified, opening up a world of possibilities. [3] [4]

4 The Future of Web Mining

It is easy to see why the future of web mining is said to look so bright. Shown here, it can be used in a huge variety of fields with remarkable benefits and few drawbacks. New developments in machine learning, big data, and natural language processing are opening up new possibilities for the extraction and analysis of information from the web.

We are able to speculate on the trajectory of web mining, and as such, an area that is likely to see significant growth and improvement in the coming years is deep web mining. This is the process of extracting data from the deep web or hidden web, which is full of data that is not accessible from the typical search engines, and thus requires a mediator such as deep web mining to be utilized fully. Another area of focus is web personalization, while touched on here, 'The sky is the limit' would not be an exaggeration when discussing this topic. A website or browser learning from the users that navigate it and using that information to improve itself has an exponential curve. The more the user interacts with the system, the more the system can improve, prompting more user interaction and so on.

Another area where we are likely to see web mining be applied more is in the context of social media, where it can be used to analyze user-generated content, identify trends, and track sentiment. Social media platforms generate vast amounts of data, and web mining techniques can be used to extract valuable insights from this data. This does bring up the prospect of user security and privacy though, a hot topic in the world of social media at the moment, which is likely going to be exacerbated by the introduction of web mining. Web mining can also be used in the cybersecurity field to detect and prevent online threats. An area where it could be particularly effective is in negating malware and phishing attacks. By analyzing the form and pattern of these attacks, as well as analyzing network traffic, it's possible that these attacks could be a much lesser threat in the future than they currently are.[9]

5 Conclusion

In conclusion, web mining has emerged as a popular research area due to its potential to provide insights into user behavior, web content, and web structure. The process of discovering patterns and extracting useful information from large data sets of web data has numerous applications in various domains, including e-commerce, education, healthcare, and social media. It can also improve the web browsing experience and increase efficiency in the best cases, personalize the content and layout of web pages, and provide invaluable insights into student behavior in online learning environments. Web mining is a valuable tool and certainly has many more applications not yet discovered. As the internet continues to grow and become embedded in almost all of our daily lives, web mining is likely to become increasingly important in understanding and utilizing the wealth of information available to us.

References

1. Ai, J., Laffey, J.: Web mining as a tool for understanding online learning. MERLOT Journal of Online Learning and Teaching **3**(2), 160–169 (2007)
2. Eirinaki, M., Vazirgiannis, M.: Web mining for web personalization. ACM Transactions on Internet Technology (TOIT) **3**(1), 1–27 (2003)
3. Etzioni, O., Banko, M., Soderland, S., Weld, D.S.: Open information extraction from the web. Communications of the ACM **51**(12), 68–74 (2008)
4. Etzioni, O., Fader, A., Christensen, J., Soderland, S., et al.: Open information extraction: The second generation. In: Twenty-Second International Joint Conference on Artificial Intelligence. Citeseer (2011)
5. Khder, M.A.: Web scraping or web crawling: State of art, techniques, approaches and application. International Journal of Advances in Soft Computing & Its Applications **13**(3) (2021)
6. Kumar, A., Singh, R.K.: Web mining overview, techniques, tools and applications: A survey. International Research Journal of Engineering and Technology (IRJET) **3**(12), 1543–1547 (2016)
7. Li, Y., Zhong, N.: Web mining model and its applications for information gathering. Knowledge-Based Systems **17**(5-6), 207–217 (2004)
8. Ratnakumar, A.J.: An implementation of web personalization using web mining techniques. Journal of Theoretical and applied information technology **18**(1), 67–73 (2010)
9. Srivastava, J., Desikan, P., Kumar, V.: Web mining: Accomplishments and future directions. In: National Science Foundation Workshop on Next Generation Data Mining (NGDM'02). pp. 51–56 (2002)
10. Wang, Y.: Web mining and knowledge discovery of usage patterns. CS748T Project (Part I) Feb (2000)