International Journal of Computer Applications Technology and Research
Volume 6–Issue 6, 269-272, 2017, ISSN:-2319–8656

Web Mining and Qualities of a Website Design to Be Evaluated for Customer Browsing Behavior: A Review

Sunil B. Joshi
Assistant Professor,
SIMCA,Narhe, Pune,
MH, 411041

Dr. Shivaji D. Mundhe
Professor & Director (MCA),
SIMCA, Narhe, Pune,
MH, 411041

Abstract: Mining the web is defined as discovering knowledge from hypertext and World Wide Web. The World Wide Web is one of the longest rising areas of intelligence gathering. Now a day there are billions of web pages, HTML archive accessible via the internet, and the number is still increasing. However, considering the inspiring diversity of the web, retrieving of interestingness web based content has become a very complex task. In this paper researcher has done a review of web mining concepts with techniques and the some of the qualities of a website design to be evaluated for customer browsing behavior.

Keywords: Web Mining, Web Content Mining, Web Usage Mining, Web log Mining, Browsing Behavior, website design qualities

1. INTRODUCTION

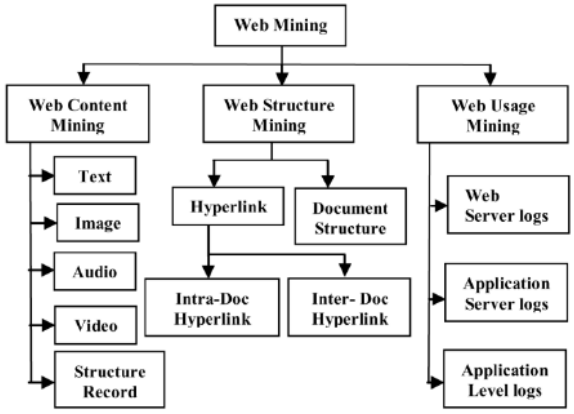
Web mining is the application of data mining techniques to extract knowledge from web data, including web documents, hyperlinks between documents, usage logs of web sites, etc. Data mining is the process which automates the extraction of predictive information, discovers the interesting knowledge from large amounts of data stored in databases, data warehouses or other information repositories. The WWW continues to grow at a wonderful rate as an information gateway and as a medium for conducting business. Web knowledge mining has been widely used in the past for analyzing huge collections of data, and is currently being applied to a variety of domains. Based on several research studies web mining can be broadly classified into three domains: content, structure and usage mining. Web content mining is the process of extracting knowledge from the content of the actual web documents (text, content, multimedia, etc.). Web structure mining is targeting knowledge from the Web structure, hyperlink references and so on. Web usage mining attempts to discover useful knowledge from the secondary data obtained from the interactions of the users with the Web.

2. WEB MINING TAXONOMY

Web mining can be broadly divided into three distinct categories, according to the kinds of data to be mined.

2.1 Web Content Mining

Web content mining is the process of extracting useful information from the contents of web documents. Content data is the collection of facts a web page is designed to contain. It may consist of text, images, audio, video, or structured records such as lists and tables. Application of text mining to web content has been the most widely researched. Issues addressed in text mining include topic discovery and tracking, extracting association patterns, clustering of web documents and classification of web pages. Research activities on this topic have drawn heavily on techniques developed in other disciplines such as Information Retrieval (IR) and Natural Language Processing (NLP). While there exists a significant body of work in extracting knowledge from images in the fields of image processing and computer vision, the application of these techniques to web content mining has been limited.



Web Structure Mining

The structure of a typical web graph consists of web pages as nodes, and hyperlinks as edges connecting related pages. Web structure mining is the process of discovering structure information from the web. This can be further divided into two kinds based on the kind of structure information used.

Hyperlinks

A hyperlink is a structural unit that connects a location in a web page to a different location, either within the same web page or on a different web page. A hyperlink that connects to a different part of the same page is called an intra-document hyperlink, and a hyperlink that connects two different pages is called an inter-document hyperlink.

Document Structure

In addition, the content within a Web page can also be organized in a tree structured format, based on the various HTML and XML tags within the page. Mining efforts here have focused on automatically extracting document object model (DOM) structures out of documents.

2.2 Web Usage Mining

Web usage mining is the application of data mining techniques to discover interesting usage patterns from web usage data, in order to understand and better serve the needs of web-based applications. Usage data captures the identity or origin of web

users along with their browsing behavior at a web site. Web usage mining itself can be classified further depending on the kind of usage data considered:

Web Server Data

User logs are collected by the web server and typically include IP address, page reference and access time.

Application Server Data

Commercial application servers such as Weblogic, StoryServer, etc have significant features to enable E-commerce applications to be built on top of them with little effort. A key feature is the ability to track various kinds of business events and log them in application server logs.

Application Level Data

New kinds of events can be defined in an application, and logging can be turned on for them - generating histories of these events. It must be noted, however, that many end applications require a combination of one or more of the techniques applied in the above the categories.

3. KEY CONCEPTS OF WEB MINING

3.1 Ranking Metrics—for Page Quality and Relevance Searching the web involves two main steps

Extracting the pages relevant to a query and ranking them according to their quality. Ranking is important as it helps the user look for "quality" pages that are relevant to the query. Different metrics have been proposed to rank web pages according to their quality.

We briefly discuss two of the prominent ones.

PageRank

PageRank is a metric for ranking hypertext documents based on their quality. Page, Brin, Motwani, and Winograd (1998) developed this metric for the popular search engine Google4 (Brin and Page 1998). The key idea is that a page has a high rank if it is pointed to by many highly ranked pages. So, the rank of a page depends upon the ranks of the pages pointing to it. This process is done iteratively until the rank of all pages are determined.

The rank of a page p can be written as:

$$PR(p) = d/n + (1-d)\sum_{(q,p)\epsilon G} \frac{PR(q)}{Outdegree(q)}$$

Here, n is the number of nodes in the graph and OutDegree(q) is the number of hyperlinks on page q. Intuitively, the approach can be viewed as a stochastic analysis of a random walk on the web graph. The first term in the right hand side of the equation is the probability that a random web surfer arrives at a page p by typing the URL or from a bookmark; or may have a particular page as his/her homepage. Here d is the probability that the surfer chooses a URL directly, rather than traversing a link and 1−d is the probability that a person arrives at a page by traversing a link. The second term in the right hand side of the equation is the probability of arriving at a page by traversing a link.

Hubs and Authorities

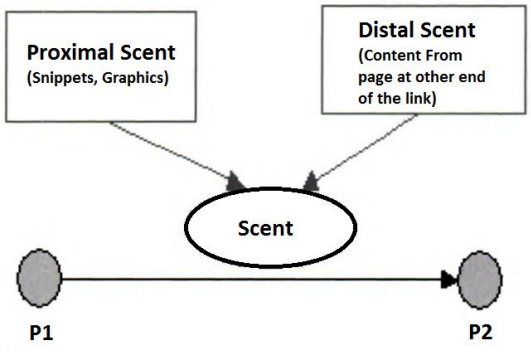
Hyperlink-Induced Topic Search (HITS; also known as hubs and authorities) is a link analysis algorithm that rates Web pages, developed by Jon Kleinberg. The idea behind Hubs and Authorities stemmed from a particular insight into the creation of web pages when the Internet was originally forming; that is, certain web pages, known as hubs, served as large directories that were not actually authoritative in the information that they held, but were used as compilations of a broad catalog of information that led users direct to other authoritative pages. In other words, a good hub represented a page that pointed to many other pages, and a good authority represented a page that was linked by many different hubs.

3.2 Robot Detection and Filtering—Separating Human and Nonhuman Web Behavior

Web robots are software programs that automatically traverse the hyperlink structure of the web to locate and retrieve information. The importance of separating robot behavior from human behavior prior to building user behavior models has been illustrated by Kohavi (2001). First, e-commerce retailers are particularly concerned about the unauthorized deployment of robots for gathering business intelligence at their web sites. Second, web robots tend to consume considerable network bandwidth at the expense of other users. Sessions due to web robots also make it difficult to perform click-stream analysis effectively on the web data. Conventional techniques for detecting web robots are based on identifying the IP address and user agent of the web clients. While these techniques are applicable to many well-known robots, they are not sufficient to detect camouflaged and previously unknown robots. Tan and Kumar (2002) proposed a classification based approach that uses the navigational patterns in click-stream data to determine if it is due to a robot. Experimental results have shown that highly accurate classification models can be built using this approach. Furthermore, these models are able to discover many camouflaged and previously unidentified robots.

3.3 Information Scent—Applying Foraging Theory to Browsing Behavior

Information scent is a concept that uses the snippets of information present around the links in a page as a "scent" to evaluate the quality of content of the page it points to, and the cost of accessing such a page (Chi, Pirolli, Chen, and Pitkow 2001). The key idea is to model a user at a given page as "foraging" for information, and following a link with a stronger "scent." The "scent" of a path depends on how likely it is to lead the user to relevant information, and is determined by a network flow algorithm called spreading activation. The snippets, graphics, and other information around a link are called "proximal cues."



The user's desired information need is expressed as a weighted keyword vector. The similarity between the proximal cues and the user's information need is computed as "proximal scent." With the proximal cues from all the links and the user's information need vector, a "proximal scent matrix" is

generated. Each element in the matrix reflects the extent of similarity between the link's proximal cues and the user's information need. If enough information is not available around the link, a "distal scent" is computed with the information about the link described by the contents of the pages it points to. The proximal scent and the distal scent are then combined to give the scent matrix. The probability that a user would follow a link is then decided by the scent or the value of the element in the scent matrix.

4. POSSIBLE WEBSITE DESIGN QUALITIES TO BE EVALUATED FOR CUSTOMER BROWSING BEHAVIOR

Everyone is talking about good website design, but do we know what it actually means? Do we know how to tell if our own web design is working? Do we know what to look for? Without having a clear metric for measuring the quality, it is hard to recognize the quality. There is more to a successful website than simply looking nice or being able to function. Ideally, a successful website would be built with a specific strategy in mind, it is focused on usability so that visitors can navigate the website with success, it incorporates a style that is pleasing to the user's eye, it is filled with content that is relevant to the users, and it is optimized for search engines. Keeping that in mind, these are a few of the key aspects of a strong website design, complemented with some questions that we may ask ourself when measuring the quality of any website.

4.1 Strategy

A Good Website Layout is backed by Good Strategy. Even the most user-friendly and attractive websites are not successful when they are not achieving the needs of the company. Ask ourself: will our visitors get a clear sense of what we offer and who we are upon arrival at our website? Will our design point our visitors in the direction that we want them to go? Is there a clear strategy that is informing of our design? If the answer to any of these questions are no, then our design is not at its full potential. In order to evaluate the effectiveness of your strategy in the website design, here is a checklist of questions to run through:

• What is the category of the business, and is this obvious on the website?
•What is the purpose of the website and is the design accomplishing this?
• Who is the target audience and how does the design take this into consideration?
•What should the audience do and is this design encouraging them to take that action?

After answering these questions, we should be able to define our brand and then set the specific website goals. This will allow us to align our design according to the goals. When the website is informed by a clear strategy the likeliness that it will succeed raises. We should be able to answer all of the previous questions with confidence that the users will be able to take the action that is intended with very little opposition from the website's overall strategy. The strategy will directly influence the design, as they work hand in hand to ensure that the website's purpose is clear, and there is no confusion about the website. Outline your target audience. Once we know who this is, approach the website as if we are them. Is it obviously apparent that the website can accomplish what the target audience needs it for? If not, take a look back and refine your strategy.

4.2 Usability

This is all about the practical consideration of what makes a good web site design, such as user-friendliness, speed, security, site maps and other technical details, etc. Many of these details are not visually apparent—we will not see a website's security when typing in the URL. Even so, usability will make or break a website. If the visitor is not able to find what they are looking for because of bad navigation, the user will generally leave. If a page takes too long to load, not only will visitors notice but search engines will notice as well. To evaluate how usable our website is ask the following questions:

• How long does it take for pages to load and will the visitors get bored waiting? There are free tools online to test page load speeds.
•Can information be found easily?
•Is there a search button available for visitors?
•Are all the links working? There are tools available online to check a website's links.
•Does the website work in different browsers? Check all of the widely used browsers.
•Does the website work on mobile devices?
•If asking for personal details for taking part in e-commerce, is the customer's information secure? Has this been communicated to the users?

Think of all the ways that will make our website as usable as possible. Imagine that we are coming to it as a visitor and we are trying to find out more information. Additionally, take the extra step in terms of security and be sure to always protect the customer's personal data. The website should be safe to use and should protect the information of the users. If this is not the case, then there will be negative repercussions. Customers will not trust the website at all, and there is a possibility that they will leave negative reviews elsewhere on the internet. The most vital components call-to-action on the website should not be any more than just a few clicks away. If a user has to hunt for the action that they need to complete, they will likely get frustrated and leave the website. They need to be able to navigate smoothly and simply, without the need to guess whether or not they are even on the correct page of the website. Accessibility also falls into the category of usability. It is vital to be sure that your website is accessible to anyone on the internet. This means that it needs to meet or be able to meet accessibility requirements. Anyone with any sort of mental, physical, or cognitive limitations should be able to use the website with little to no alteration. This means that words should be able to be read clearly with the color choices, font choices, and size choices. Creating ALT text on images will allow accessibility software to use the text to audibly describe the images to those with a hearing impairment. In addition to this, the website should be accessible with a keyboard, for those that lack the dexterity in their arms or hands.

4.3 Content and Readability

The two main things to take into consideration when regarding content are its usefulness and readability. Content needs to matter to the readers or we will lose them. Readability is vital, because if the visitors cannot make sense of our content whether it be because it is too small or in a strange color or unreadable font, there's no way to convey the message. These are some questions to ask ourselves when considering the content of the website and how to evaluate its quality:

• Are the fonts that of been chosen easy to read?
•Is there considerable contrast between the font color in the background color?
•Is the text an appropriate size?
•Will the content be relevant to the reader?
•Is the content concise yet still useful?
•Does the overall design make content easy to find?

Evaluate all of the text on the website. Is it conveying the message effectively? Will visitors be able to read the content?

Is what they are reading important to them? Aim for a design that will make all of the content useful and readable. Aside from the aesthetic choice for the content, the quality of the content needs to pull the reader in. Speaking to the visualization that a website is a storefront, you wouldn't walk into a flower shop and find a small selection of flowers, but the walls are stocked with chocolates. This is indicative of a poor, or lack of, strategy. There needs to be some sort of passion behind the strategy to make it clear what exactly the content will be. It is fine to blur the lines of the content every now and again, but the overall message should be clear. Creating content that is consistently relevant will indicate to the visitors of the website that they can grow to expect what type of content can be found on the website, and will increase the chances that they will return to the website.

4.4 Aesthetics

Some say beauty is relative, but that does not mean that there are not defined aesthetic principles that should guide website design. The best type of design will align with the brand, create a positive impression for visitors, be clean, and it will complement the content that is being communicated. To test the effectiveness of the website's aesthetic, ask ourselves the following:

• Does the website's style align with the brand in terms of color, graphics, feel, etc.?

•Is the style consistent throughout the entire website?

• Does the style suit the target audience? (An elegant layout on the website, cartoons on a toy company website etc.)

• How do visitors view the site? Sparse or crowded, orderly or messy, formal or playful? And how does this align with the goals?

•Are there any photos or decorative touches that are getting in the way of the message?

After evaluating these questions, jettison any stylistic choices that directly contradict with the brand's message. Ensure that the logo and website design align. Consider the target audience and let that influence the style. This should not be the most important component (strategy is), but it should go hand in hand with it. The style of the website should complement the brand's goals and intentions, and should never confuse the user of the website. Think of it as the "storefront" of our website. In real life, we'd walk into a professional's office and expect it to be clean, tidy, and quiet. We'd also walk into a party store and expect bright colors, music, and lots of people talking. Imagine if we walked into a party store that was like a professional's office and vice versa. That would leave the customers very confused and they would likely choose not to do business there.

4.5 SEO and Social Networking

There are a lot of ways that the design of a website will impact search optimization. SEO and social networking starts with a strong website design. For example, does the website have a lot of graphics? If it does, remember that the search engines cannot see them. You will need to add ALT tags to the image descriptions so that the search engine will know what is being shown. Is the HTML efficient? If it is not this will hurt search rankings. Consider asking yourself the following questions to ensure that the website design is optimized:

• Are all of the images optimized with ALT tags?

•Is the coding efficient or are there extra lines that can be eliminated?

•Are relevant keywords being used in title tags, heading tags, meta-descriptions, etc.?

•Is there a site map?

A huge mistake is to think that search engine optimization and website design are two separate matters. Consider the way that the design will affect the search rankings and make any adjustments accordingly. The benefits of asking these questions and designing the website accordingly work hand in hand with one another. Not only will adding ALT tags to images aid in the search engine optimization, they will also help in making the website accessible. This is true for all of the heading tags, title tags, and other metadata. A well thought out site map will make navigating the website easy for everyone, especially those that are using the website with the aid of accessibility software.

Based upon these 5 characterizations, how well is any website design working? Think of the ways in which it can be improved. Ponder on the steps that you will take in order to make the design more effective. This could mean improving upon all 5 of the characterizations, or simply improving upon one. It all depends on how well the website was designed in the first place. If anyone wants the website to perform at its absolute best in each aspect. These aspects are all distinctly different, yet they work together to make the website operate smoothly. When all facets of the website operate smoothly, the quality of the website increases, thus bringing more traffic and eventually leads to more conversions, or completes the task intended.

5. CONCLUSION

Web mining is the search for relevant information from the World Wide Web. The found web pages in a search are relevant if they provide an accurate answer to the searcher's information need. As all users of Web search engines are aware, accurate answers are not always at the top of the result list. Web mining is the extraction of interesting and potentially useful patterns and implicit information from artifacts or activity related to the World¬ Wide Web. Before developing a website, one should take in to consideration all the possible quality dimensions, so that the output would be satisfactory.

6. ACKNOWLEDGMENTS

I would like to thank Dr. Shivaji D. Mundhe for guidingme to completethis research paper.

7. REFERENCES

[1] http://dmr.cs.umn.edu/Papers/P2004_4.pdf

[2] https://dynomapper.com/blog/19-ux/188-how-to-evaluate-the-quality-of-your-website-design

[3] https://en.wikipedia.org/wiki/HITS_algorithm

[4] B. Madasamy, Dr. J. JebmalarTamilselvi, "General Web Knowledge Mining Framework", International Journal on Computer Science and Engineering (IJCSE), ISSN : 0975-3397, Vol. 4 No. 10 Oct 2012, 1744-1750

[5] http://www.bea.com/products/weblogic/server/index.shtml

[6] http://www.bvportal.com/

[7] http://www.cio.com/sponsors/110199 vignette story2.html