Web Rank Prediction Using Naive Bayes Classification On 18 Attribute Dataset

[1]Sahil Deshmukh, [2]Akshay Magar, [3]Omkar Khandve, [4]Bhagyashree Shendkar

[1]Student, [2]Student, [3]Student, [4]Assistant Proffesor
[1]Computer Department,
[1]Sinhagad Institute of Technology and Science Narhe, Pune, India

Abstract : Search Engine Optimization (SEO) is the most widely used technique to improve the visibility of your website or web page in a Search Engine's unpaid results.[2] However, there is no clear algorithm or software, which can predict the rank of a particular web page on the Search Engine Result Pages (SERP's) for a particular keyword before the web page is created. The Naive Bayes classifier, one of the most common classification algorithm can predict the rank based on the various on-page/off-page/technical factors, which affect the rank of a website. Using the basic requirements mentioned by each of Google's algorithms, which affect the web rank of the web page as attributes of the database,we created a database by manually analyzing websites ranking at different positions for various keywords on Google. The Naive Bayes Classifier' prediction of the possibility of getting a high rank in the SERP's, will be helpful to create a practical tool to analyze and predict the rank of a particular web-page before it is posted.

***IndexTerms* - Search Engine Result Pages(SERP), Domain Authority(DA), Alexa Rank, Naive Bayes, Classification Algorithm, Accelerated Mobile Pages(AMP), Schema, Search Engine Optimization(SEO), Secure Sockets Layer(SSL) Certificate,**

I. INTRODUCTION

For Years SEO experts have been working on Google search engine algorithms and have come up with different factors that affect the rankings of a web page. According to one such professional blog backlinko.com, the Search Engine looks at 200 factors before deciding which web pages will be ranked amongst the top ten results[9]. We have used 18 most important factors amongst them as the attributes of the dataset. The Naive Bayes Classifier works on this dataset to give the rank prediction. As we have reduced the number of factors from 200 to 18, we can only predict the ranks as in top five, in top 10, and not on first page. The 18 factors are:-

Domain Authority: According to moz.com, "Domain Authority (DA) is a search engine ranking score developed by Moz that predicts how well a website will rank on search engine result pages (SERPs). A Domain Authority score ranges from one to 100, with higher scores corresponding to a greater ability to rank." [1]

Alexa: According to Alexa.com, "The traffic rank is based on three months of aggregated historical traffic data from millions of Alexa Toolbar users and data obtained from other, diverse traffic data sources, and is a combined measure of page views and users (reach). As a first step, Alexa computes the reach and number of page views for all sites on the Web on a daily basis. The main Alexa traffic rank is based on a value derived from these two quantities averaged over time (so that the rank of a site reflects both the number of users who visit that site as well as the number of pages on the site viewed by those users)." [5]

SSL Certificate: SSL Certificates are small data files that digitally bind a cryptographic key to an organization's details. When installed on a web server, it activates the padlock and the https protocol and allows secure connections from a web server to a browser. Typically, SSL is used to secure credit card transactions, data transfer and logins, and more recently is becoming the norm when securing browsing of social media sites.

Content Update Frequency: This refers to the rate at which new content is posted on a particular website. It is generally measured in number of articles per week.

Schema : Schema markup is code (semantic vocabulary) that you put on your website to help the search engines return more informative results for users.[4]

Accelerated Mobile Pages(AMP): The Accelerated Mobile Pages Project (AMP) is an open-source website publishing technology designed to improve the performance of web content and advertisements for mobile devices. It is necessary to create an AMP version of your website to improve your rankings on mobile devices.

Static Pages: You need to provide basic information about your company via static pages such as About us, Contact us etc. Also it is necessary to provide information about your privacy policy, DMCA and Terms of Use for legality issues. Google considers websites without this information as Unauthentic.

Mobile Optimization: Google updated its algorithm to provide mobile first searching. It is extremely necessary that your website is mobile optimized if you want to rank on mobile devices.

Interface And User Experience: Google aims at giving the users what they want. The Interface and user Experience are one of the major factors that affect the Website Ratings by users which directly affect your rankings. According to your inputs you have a good interface so no need to worry about that.

Ad Optimization: Ad Optimization refers to the number and placement of ads on your webpage. Ad crowding is not too user friendly and hence is a factor based on which Google evaluates the website usability which affects the rank. According to your input you have an average Ad optimization.

Wikipedia Page: Wikipedia has the highest domain authority for Google and, if there is a wikipedia page associated with the targeted keyword, then that will be the first one to be displayed on the SERP's.

Trend Level: When, a particular keyword is trending, there will be many number of websites which will target the same keyword. It is one of the most important factors as for a keyword with high competition, high search volume and high trend level, it is very difficult to predict which website will rank first.

Keyword Optimization: Keyword optimization is the act of researching, analyzing, selecting and placing the best keywords to target to drive qualified traffic from search engines to your website.[6]

Internal Linking: Internal links are links that go from one page on a domain to a different page on the same domain. They are commonly used in main navigation.[7]

Grammar and Language: Grammar Language and Plagiarism refers to the Quality of Content. Without a doubt if you keep copying content word to word from different sources Google will put your website into the spam list. Also it is better to give a link back to the source from where you got the content. Lastly it goes without saying that Grammar and Language affect readability and will Lower the User Rating of your website.

Word Count: Word Count depends on the type of content you write. But if you have unique and more content compared to your competitors, you get an edge over them.

Categorization: A well categorized content helps Google to create a sitemap of your website which in turn leads to fast indexing of your article

II. PROBLEM DEFINITION:

To find out the rank of a web page on the SERP's of Google Search Engine before the page is created by using Classification Algorithm.

Objectives

- To find out the most important SEO factors that affect the rank of a web page.
- To find where the web page will stand in the Search Engine Result Pages before the page is created.

III. EQUATIONS

The Naive Bayes Classifier is based on the "Bayes' Theorem" (also known as Bayes' rule). It is a deceptively simple formula which is used to calculate the conditional probability. The Theorem was named after English mathematician Thomas Bayes (1701-1761). The formal definition for the rule is:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

IV. DESIGN OF THE STUDY

Propose Algorithm:-

Step1. Handle Data: Load the data from CSV files for training and test datasets.

Step2. Summarize: Convert the Test Data into dictionary list data structure and perform prior probability operation on training data set.

Step3. Make a Prediction: Using Conditional Probability equation calculate the posterior probability for each class. The class with the highest posterior probability is the outcome of prediction.

Step4. Evaluate Accuracy: Evaluate the accuracy of predictions made for a test dataset as the percentage correct out of all predictions made.

Tools Used:-

Python 2.x

V. RESULTS AND DISCUSSION

After testing 55 pieces of data in the training set, we find that the accuracy of training set on training set is 84.5454%. Test set has 10 pieces of data. The accuracy of test set on training set is lower than that of training set on training set which is 50%.

Table 1: Results

Accuracy		
On Training Dataset	**On Test Dataset**	**Average**
84.5454%	50%	65.2727

VI. Limitations and Future scope:

There are many factors that affect the ranking of a web-page on the Search Engine Result Pages. If all the factors are considered as individual attributes in the dataset, it will affect the accuracy of the Classifier.

The database used for verifying the working of the above algorithm was created by entering data for various keywords and websites manually. The implementation of a web crawler will help fetch the data automatically and accurately. It will also fetch the data for users website automatically which will reduce the users overhead to manually input his/her website details.

VII. Conclusion:

A classification technique is used to give the web rank prediction of a particular web-page for a particular keyword, before it is created which helps. This helps the user understand if his web-page will get visibility or not.

VIII. Acknowledgment

I take this opportunity to thank my internal guide Prof. B.D. Shendkar, for giving me guidance and support throughout the Seminar. Her valuable guidelines and suggestions were very helpful. I wish to express my thanks to Prof. G. S. Navale, Head of Computer Engineering Department, Sinhagad Institute of Technology and Science, Narhe for giving me all the help and important suggestions all over the Seminar Work. I thank all the teaching and non-teaching staff members, for their indispensable support and priceless suggestions. I also thank my friends and family for their help in collecting data without which this Seminar report not have been Completed. At the end my special thanks to Dr. R. S. Prasad, Principal Sinhagad Institute of Technology and Science, Narhe for providing ambiance in the college, which motivate us to work.

References

[1] https://moz.com/learn/seo/domain-authority

[2] Swati Gupta, Nitin Rakesh, Abha Thakral, Dev Kumar Chaudhary, "Search engine optimization: Success factors" Parallel, Distributed and Grid Computing (PDGC), 2016 Fourth International Conference.

[3] Samedin Krrabaj, Fesal Baxhaku, Dukagjin Sadrijaj, "Investigating search engine optimization techniques for effective ranking: A case study of an educational site" Embedded Computing (MECO), 2017 6th Mediterranean Conference.

[4] https://blog.kissmetrics.com/get-started-using-schema/

[5] https://trafficgenerationcafe.com/how-alexa-ranking-works/

[6] https://www.wordstream.com/blogs/ws/2010/04/14/keyword-optimization

[7] https://moz.com/learn/seo/internal-link

[8] Yuguang Huang, Lei Li, "Naive Bayes classification algorithm based on small sample set" Cloud Computing and Intelligence Systems (CCIS), 2011 IEEE International Conference.