International Journal of Science and Research (IJSR)
ISSN: 2319-7064
ResearchGate Impact Factor (2018): 0.28 | SJIF (2019): 7.583

k-Means Clustering to Enhance SEO: A Data Driven Approach

Meghana Mukunda Joshi

Dept of Computer Science and Engineering, B.M.S. College of Engineering, Basavanagudi, Bengaluru Karnataka – 560019, India

Abstract: ***In relation to Data Mining, Search Engine Optimisation is associated with applying rigorous analysis to datasets which encompass data such as performance in terms of content and user behaviour. The goal is to convert this data into actionable insights to improve page rank and increase organic traffic. It is of utmost importance for webmasters to optimize SEO factors to satisfy the search engines and thereby attain more visibility. This paper introduces a novel approach to draw valuable, reliable and practical insights from Google Search Console using k-Means Clustering to draw focus on the influence of various SEO factors. The factors taken from Google Search Console include Page, Query, Clicks, Impressions, CTR and Position. Other parameters like length of the title and Meta Description should be included from web crawlers. The clustering technique enables one to gain insight into which parameters should be optimised for maximum impact.***

Keywords: Data Mining, Search Engine Optimisation, k-Means Clustering, Google Search Console, SEO factors

1. Introduction

1.1 Search Engine Optimization

The rapid growth of the Internet has paved for a plethora of content to be accessed with ease. It is critical to eliminate the irrelevant pages from the vast amount of information available. This hitch calls for innovation in methods to provide germane information to users. Techniques to retrieve relevant data have become extremely important. In this manner, search engines and web crawlers have gained value in determining which information is more relevant compared to others.

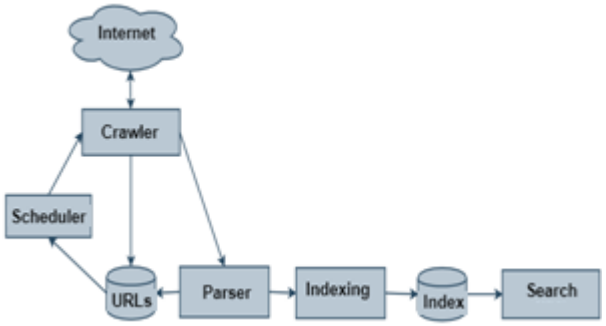

Figure 1: Working of Search Engines

Search engines employ unique guidelines in order to index webpages. The power of search engines has managed to change the ecology of the web [1]. Studies have shown that a large percentage of web page accesses are cited by search engines. It is known that the manner in which search results are displayed on the search engine results pages (SERPs) directly influences users' selection of certain results. Organic results generally take up the first few positions, these are the results which are usually preferred by the user. Furthermore, results that are "below the fold" (i.e., these results can only be found when a user scrolls down the list of displayed results) are rarely clicked on. This shows that it becomes exceedingly important to produce results sets within the visible area ("above the fold") that are appropriate and satisfy the users' needs[2].

In order to gain visibility and monetize through ads and attract more eyeballs on one's content, it is essential that webmasters optimize web-pages in a particular manner to enhance page rank. This is where Search Engine Optimisation comes into picture. This technique is related to modifying web pages according to certain SEO factors like performance, user experience, internal linking between pages and overall optimization of media attachments. All these improvements can bring about exponential growth in terms of page ranking.

SEO tools such as those provided by Google serve as a core in the SEO campaign. These tools include Google Search Console, Google Webmaster tools, Google Ad Words and Google Trends. [3] These tools give an insight into how Google interacts with the given website and obtain actionable information from Google about the site. Moreover, these tools provide data regarding the demographics of users and the keywords through which a user landed on the site. More comprehensive results can be drawn like variations in Impressions, Clicks, CTR and Position.

1.2 k-Means Clustering Technique

In this context, web mining techniques are employed by making practical and effective use of data mining methods so as to generate and extract useful information from optimization tools. In specific, clustering could be used to cluster similar click-streams to determine learning behaviors in the case of e-learning or general site access behaviors[4].

These conclusions can help webmasters make personalized decisions that can have a tremendous impact on organic traffic and ranking. An advantage of using data mining techniques for analysis is that they require minimal human interaction which allows webmasters to focus on value added tasks such as curating high-quality content and building authority.

Clusters are collections of data points that are classed together by a degree of comparability. Clustering techniques

such as k-Means cluster (an unsupervised machine learning algorithm). This algorithm creates clusters by partitioning 'n' observations into 'k' clusters. Each of these observations belongs to the cluster with the closest mean. Using these clusters, one can understand appropriate steps to enhance particular web pages. This dissertation focuses on using k-Means clustering to obtain personalized SEO recommendations.

2. Research Objectives

The purpose of this study is to explore how k-Means clustering can be used to give webmasters customized or personalised recommendations based on data from their Google Search Console. Since this data is frequently updated, which allows webmasters to monitor progress as well. Thus, the objectives of this study are as follows:

- To draw deeper and more meaningful insights from personalized Google Search Console Data
- Use clustering technique to group pages according to criteria like Impressions and CTR in order to understand which pages require additional enhancement
- Understand what values of SEO factors like meta description and tile length are more suitable

3. Literature Review and Related Studies

With the popularity of the web only increasing, millions of users make use of search engines to uncover information. However, a large portion of users are only interested in the first few results. This places pressure on webmasters to land their pages in the first few search results, which often tends to Black Hat SEO (i.e. inorganic traffic). The goal is to understand and enhance SEO factors to improve organic traffic. The authors of [5] have provided a detailed study on Search Engine algorithms and Optimization techniques like Keyword Density, Title tag, Keyword Location, URL, inbound and outbound links etc., while promoting White Hat SEO (i.e. Organic Traffic). However, it becomes difficult for webmasters to analyze all the SEO factors for each individual page while still managing to curate content.

The authors of [6] employed optimization techniques of SEO such as title length, alt images and various HTML modifications to show. In this study, results before and after implementation of these techniques are shown to throw light on the importance of SEO. This study, however, does not use any personalized recommendations to analyse which pages need to be further optimized and enhanced.

Further, in study [7], the benefits of using a data mining approach for search engine optimization is given. The study covers keyword generation, directory submission and link exchanges powered by data centered techniques. In addition, the authors of [8] propose an approach to SEO using classification and clustering techniques.

Clearly, using data driven techniques to optimise a website has gained popularity and is being studied in depth. In [8], an approach to use the Message Passing Application Programming Interface (MPAPI) technique with the K-Means clustering algorithm is proposed. This method is used in order to speed up the search process and enhance search results.

In order to find patterns and segmentation in data, cluster analysis is popularly used. The reason cluster analysis is a chief data mining technique is because one can obtain the data distribution, observe characteristics of clusters and gather deeper insights based on these clusters [9]. Additionally, visualizing clustered data can yield useful and actionable cognizance.

The authors of [10] proposed a method of using k-means to create clusters of various search engine optimization attributes such as keyword selection, meta descriptions and 21 others. Further, it was portrayed that clustering can distinguish more important attributes so as to increase traffic and page rank. However, for webmasters trying to optimise websites with many pages (such as blog portals), it is necessary that they draw insights from the performance of their own pages and derive personalized recommendations.

4. Proposed Technique for Clustering

4.1 Data Set from Google Search Console

In order to obtain valid results, data curation is a key process.
This paper proposes a simple technique for webmasters to collect and curate the data from Google search Console. This tool gives webmasters almost real-time data for each webpage. The following parameters in the dataset are taken from Google Search Console:

- **Page**: Page is a categorical label, using which data can be grouped.
- **Clicks**: The total number of clicks on the particular web page URL from Google Search results page. This attribute does not include clicks obtained via Google Ads.
- **Impressions**: Impressions represent the total number of the web page URL appeared on the Google Search results. The goal is to increase impressions, so as to increase visibility.
- **CTR:** Click through rate is calculated as:

$$Clicks/Impressions * 100 \quad (1)$$

- **Position:** Average ranking of website URL for particular queries.
- **Query:** These are the queries which generated impressions for a particular webpage URL**.** Similar to page, query is a categorical label used to group data.

The following parameters are extracted from web crawlers and merged with the dataset obtained from Google Search Console:

- **Title Length:** The title of a web page is specified by an HTML element called title tag. These titles are displayed on SERP as clickable headings. These tags should be concise as Google truncates and shows only the first 50-60 characters [12].
- **Meta Description Length:** This attribute is also an HTML element which provides a short summary of the content of a webpage. Google usually truncates this description to 155-160 characters [13].

Table 1: Dataset Parameters

Name	Type	Role
Page	Categorical	Meta
Clicks	Numeric	Feature
Impressions	Numeric	Feature
CTR	Numeric	Meta
Position	Numeric	Feature
Title Length	Numeric	Meta
Meta Description Length	Numeric	Meta
Query	Categorical	Meta

4.2 k-Means Clustering Algorithm

Analyzing numerous combinations of queries from hundreds of web pages, identifying those pages which require most optimization can be a tedious task. Visualizing the data in the form of clusters makes it easier to spot these pages since clusters have the ability to reveal underlying patterns in the data.

The k-Means clustering algorithm partitions a dataset into 'k' non-overlapping clusters, in such a way that a data point belongs to exactly one group or cluster. This algorithm aims to minimise the sum of squared distance between the clusters centroid and the data points.

The objective function is as below:

$$J = \sum_{i=1}^{m} \sum_{k=1}^{K} w_{ik} \, ||x^i - \mu_k||^2 \quad (2)$$

where $w_{ik}=1$, if k contains data point x^i; else $w_{ik} = 0$.

The clustering process is as follows:

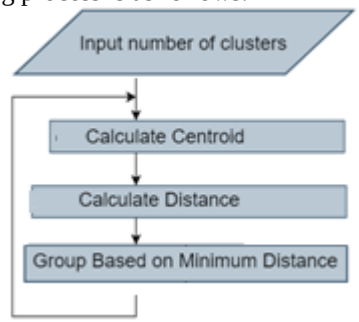

Figure 2: k-Means Algorithm

A silhouette score is used to determine the number of clusters (represented as 'k'). The silhouette is a method that validates consistency of clusters. The range is from 0 to 1, where a higher value indicates that the cluster is consistent. In this case, 3 clusters (k=3) is the best way to group this data since the silhouette score is closest to 1 (0.92).

4.3 Data Visualization

In order to visualize the gathered data, a tool called Orange can be used. The pipeline for data flow in Orange is as shown in Fig 2. After applying k-Means on the data, it is represented as a scatter plot. This scatter plot shows the different clusters, and can unveil underlying patterns in the data. .

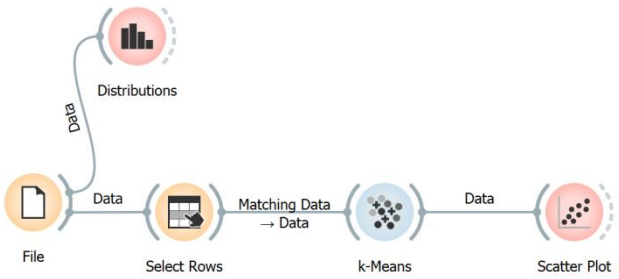

Figure 3: Pipeline for Data Flow

The file widget allows for the dataset to be imported in CSV format or excel format. Rows can be selected by applying filters (only those below a certain position), so that concentration can be drawn to pages capable of improvement. On the selected rows, k-Means is applied with k=3. This data is then represented as a scatter plot, where x and y axes can be chosen accordingly, and appropriate inferences can be drawn.

5. Results

5.1 Impressions vs CTR

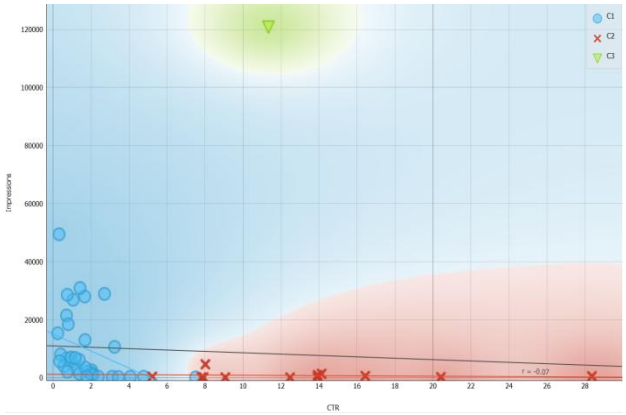

Figure 4: Impressions vs CTR Cluster Analysis

The goal is to identify web pages with high impressions and low CTR. As shown in Fig 4, there is no advantage of working on Cluster 2, as these web pages are already receiving a high CTR. Similarly, Cluster 3 is receiving both high impressions and CTR. However, in Cluster 1, though some of these pages have higher impressions, CTR is very low. This shows that attention should be given to pages grouped in this cluster. In this manner, webmasters can analyze clusters against these parameters and gain personalized insights.

5.2 Positions vs Impressions

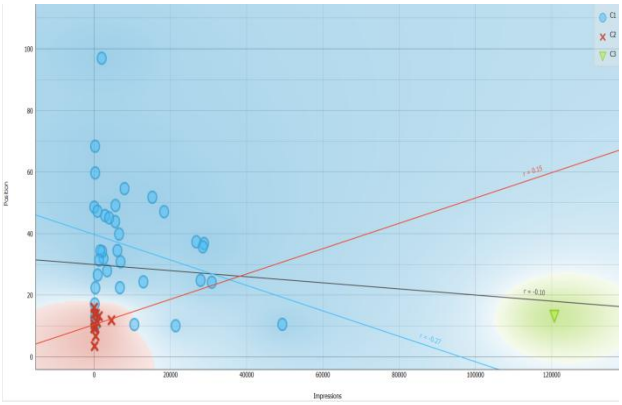

Figure 5: Position vs Impressions Cluster Analysis

Webmasters often focus on positions, however as shown in Fig 5, focus should be drawn to increasing impressions and thereby increasing organic traffic. In Cluster 3, though position is not very high, impressions are extremely high. It can be inferred that it is equally, if not more important to improve visibility by improving impressions, rather than position.

5.3 Title Length and Meta Description length

According to [12]-[13], it is recommended to keep title length between the range of 50-60 characters, and meta description length between 155-160 characters. Oftentimes, webmasters try to fit the length of these parameters within this generic range. However by clustering and visualizing length vs CTR, this range can be broadened and customized for websites.

6. Conclusion

A data driven approach to Search Engine Optimization has the ability to reveal deeper and more individualized insights. With the increase in priority given to ranking higher in SERP, there has been a consequential increase in the number of search engine optimisation factors and techniques. Webmasters may often struggle with optimization due to the plethora of factors. Analyzing customized data (i.e. data gathered from tools like Google Search Console) can make this process simpler. The data can be visualized more efficiently using cluster analysis, from k-Means clustering algorithm. Using the proposed techniques, analysis can be performed on various combinations of factors, with large amounts of data in a very short period of time. Validating and taking action on these results can result in improved page rank and overall SEO score improvement.

References

[1] Junghoo Cho and Sourashis Roy, 2004. Impact of search engines on page popularity. In Proceedings of the 13th international conference on World Wide Web
[2] (WWW '04). Association for Computing Machinery, DOI:https://doi.org/10.1145/988672.988676
[3] N. Höchstötter and D. Lewandowski, What users see – Structures in search engine results pages, Information Sciences 179 (2009)
[4] Ankalkoti, Prashant. (2017). Survey on Search Engine Optimization Tools & Techniques. "Imperial Journal of Interdisciplinary Research (IJIR). Vol-3. 40-43.
[5] Belsare Satish, Patil Sunil, *Res. J. Recent Sci.*, Volume 1, Issue (ISC-2011), Pages 375-387,(2012)
[6] Patil Swati , Pawar B.V., Patil Ajay S, "Search Engine Optimization: A Study",Research Journal of Computer and Information Technology Sciences Vol. 1(1), 10-13, February (2013)
[7] Hatab, Rayhan. (2014). Improve Website Rank Using Search Engine Optimization(SEO).
[8] P Sujatha and K Kavitha, Journal on Science Engineering & Technology Vo1ume 2, No. 03, June 2015
[9] Khorsheed, K & Madbouly, M & Khorsheed, Khattab & Madbouly, Magda & Guirguis, Shawkat. (2015). SEARCH ENGINE OPTIMIZATION USING DATA MINING APPROACH. IX. 184.
[10] S. Yong and Z. Ge, "Research on an improved algorithm for cluster analysis," 2011 International Conference on Consumer Electronics, Communications and Networks (CECNet), XianNing, 2011, pp. 598-601, doi: 10.1109/CECNET.2011.5768863.
[11] Duklan, Nitin & Mourya, Diwakar & Bahuguna, Himanshu. (2015). Classification of search engine optimization techniques: A data mining approach.
[12] https://moz.com/learn/seo/title-tag
[13] https://moz.com/learn/seo/meta-description