International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 06 Issue: 06 | June - 2022 Impact Factor: 7.185 ISSN: 2582-3930

Blog Mining Using Search Engine Optimization and Machine learning

Dr.V. Shanmukha Rao[1], Vikram Pagadala[2], MD. Ibraheem[3], CH. Roop Kumar[4]
Associate Professor, B.Tech Students
[*] Department of Information Technology
[**] ANDHRA LOYOLA INSTITUTE OF ENGINEERING & TECHNOLOGY

Abstract- We are providing a platform to Blogger's and Reader's. Where the blogger will continually post new content or information in the Blog and readers will check them frequently and gain knowledge or information from the blogs. So, we created a website for both of them, where bloggers have a separate account in the website, they can login and update/write the content and readers can go to the home page directly to view the blogs of bloggers. Here we are creating a Techie Blog website.

Keywords- Blog, Blog-mining, Machine learning, Latent Semantic Analysis, Search Engine Optimization, Hummingbird, Panda & Hybridization.

I. INTRODUCTION

The blog is a regularly updated website or web page, typically run by an individual or small group, that is written in an informal or conversational style. Blogs consist of a series of posts where posts are archived, and are usually sorted into categories. Bloggers identify the sentiments, both positive and negative opinions about the topic to understand and present public views in detail. Readers can browse these categories through the blog to read older entries. It does typically involve searching and analyzing blogs in order to generate additional insights and acts as an information source for the user's ideas. Blog Mining provides a capability of processing large amounts of text data effectively, blog mining can be a valuable method for gaining insights into a given topic. This study of blog mining is used to analyze and search the online blog posts relevant contents in a quite simpler fashion

In this website we are providing three main features to the website, where it will most helpful to Blogger's and Reader's as well, they are:

1. Blog Mining
2. Text Summarization
3. Search Engine Optimization

II. AIM & OBJECTIVE

To provide fast & quality information to blog readers, for Blogger a safe secure login page and write/update the content.

The typical goals of the Blog mining are to collect the blog corpus and design a system for topic identification and other text processing tasks such as text summarization unit, text categorization, and information routing

III. EXISTING SYSTEM & ITS LIMITATIONS

Blogs consist of a series of posts All posts are archived, and are usually sorted into categories. Readers can browse these categories or page back through the blog to read older entries. As the overloaded content which is extracted from a huge number of posts might even result in low search results.

LIMITATIONS

- Takes a lot of time to search.
- Old search result of blog.
- Huge content, take a lot time to read.

IV. PROPOSED SYSTEM & ITS ADVANTAGES

Here The proposed system pulls out search optimization and allows users to acknowledge deluge information of blogging websites. It explains about keyword research which involves finding the keywords (or search queries) that are extracted from the blog. It emphasizes Machine learning based text processing tasks such as text summarization unit, text categorization, and information routing. Involvement of text summarization allows to primarily produce tokens, where emphasized tokens are recognized, through which arguments are assigned for preceding and following tokens. As the amount of text keeps growing, it becomes increasingly difficult for humans to process the deluge of information in the time available. It does result in consumption of a huge amount of duration. As a part of low efficiency, the subject propagated within the blogs may not be reachable to end users. Blog Mining is overcome by performing text routing techniques like text summarization. The overloaded content which is extracted from a huge number of posts provides low search results, as a process of application of summarization providing faster pace of search.

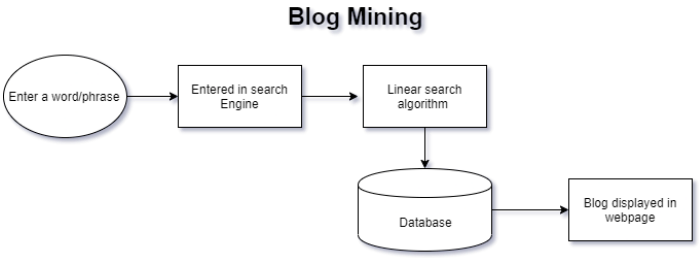
ADVANTAGES

- This process allows consumption time to be shorter, having quicker search results.
- Allowing blog users to consume quicker ways of getting understandability of information.
- Blog-Reader will receive the best results using SEO.
- Providing secure login for bloggers.

V. STUDY OF THE SYSTEM

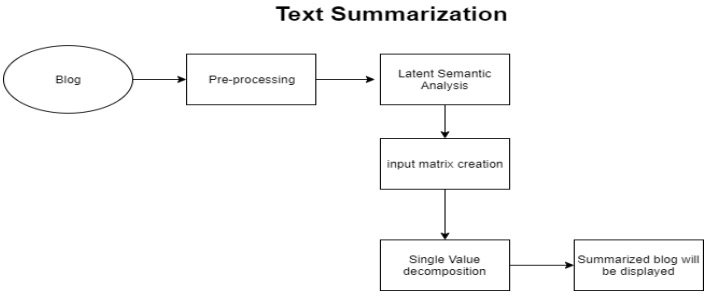
Blog-Mining

A blog is an online journal or informational website displaying information in reverse chronological order, with the latest posts appearing first, at the top. It is a platform where a writer or a group of writers share their views on an individual subject. It is an online journal where an individual, group, or corporation presents a record of activities, thoughts, or beliefs.



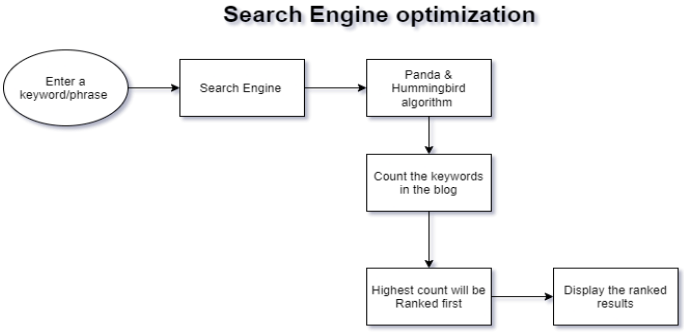
Text Summarization

Text summarization is the process of creating a short, accurate, and fluent summary of a longer text document. Automatic text summarization methods are greatly needed to address the ever-growing amount of text data available online to both better help discover relevant information and to consume relevant information faster. The data is unstructured and the best that we can do to navigate it is to use search and skim the results.



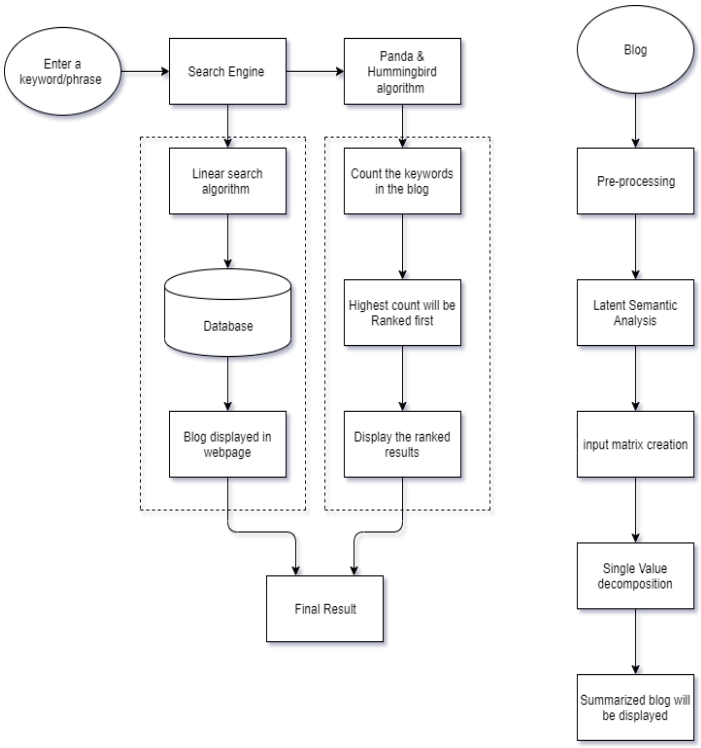
Search Engine Optimization

Search Engine Optimization (SEO) is the activity of optimizing web pages or whole sites in order to make them search engine friendly, thus getting higher positions in search results. Internet is growing at a rapid pace and it has huge impact on several businesses include on-site optimization and off-site optimization techniques. On-site optimization is optimizing your website in a way that it can rank better in search engines and improve visitor satisfaction. includes website design elements such as keyword formatting, keyword in title tag, position of keyword, keyword density, keyword in meta tag etc.



VI. SYSTEM ARCHITECTURE

The complete working procedure of the project.



VII. SYSTEM DESIGN

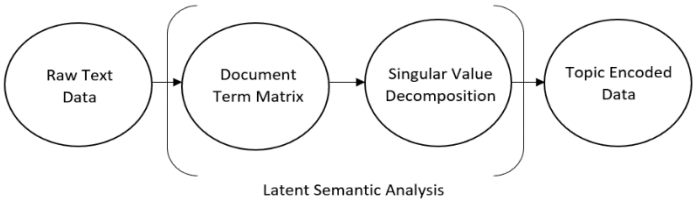
System design shows the overall design of system. In this section we discuss in detail the design aspects of the system.

METHODOLOGY INVOLVED IN THIS PROJECT

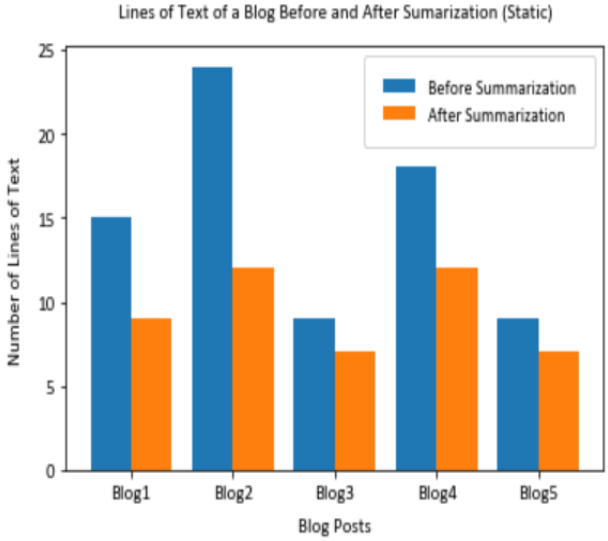
ALGORITHM – LATENT SEMANTIC ANALYSIS

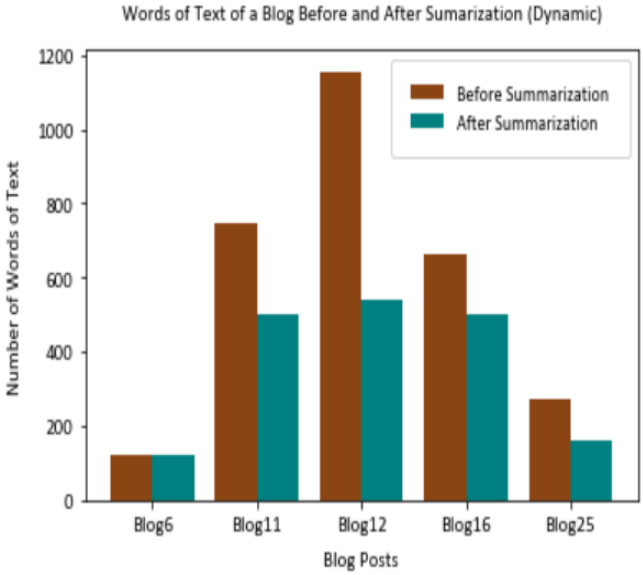
The main aim of latent semantic analysis is to create representations of text data in terms of the features and latent features. Latent semantic analysis (LSA) is a mathematical

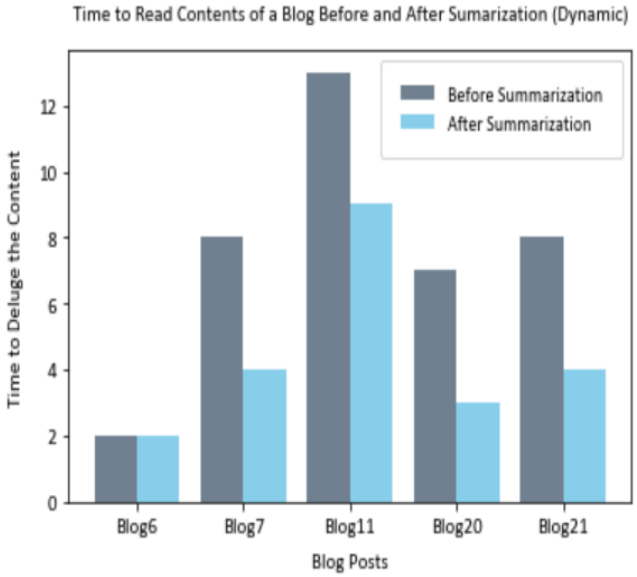
method for computer modelling and simulation of the meaning of words and passages by analysis of representative corpora of natural text. LSA closely approximates many aspects of human language learning and understanding. It supports a variety of applications in information retrieval, educational technology and other pattern recognition problems where complex wholes can be treated as additive functions of component parts. The latent sematic analysis consists of two steps:

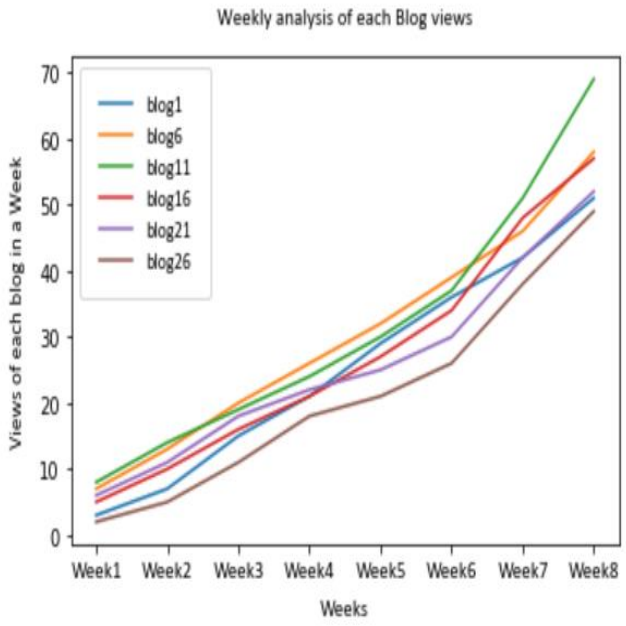


THE RESULTS





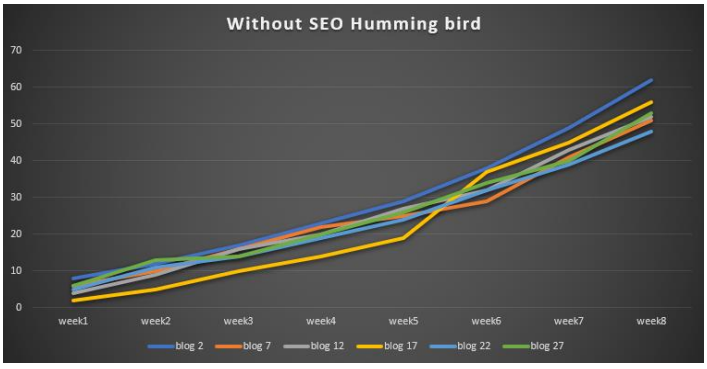


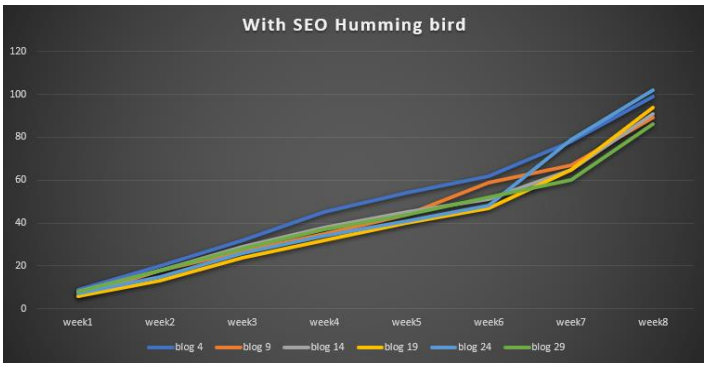


ALGORITHM – HUMMINGBIRD

Hummingbird is the codename given to a significant algorithm change in Google Search in 2013. Its name was derived from the speed and accuracy of the hummingbird. The change was announced on September 26, 2013, having already been in use for a month. "Hummingbird" places greater emphasis on natural language queries, considering context and meaning over individual keywords. It also looks deeper at content on individual pages of a website, with improved ability to lead users directly to the most appropriate page rather than just a website's homepage.

THE RESULTS





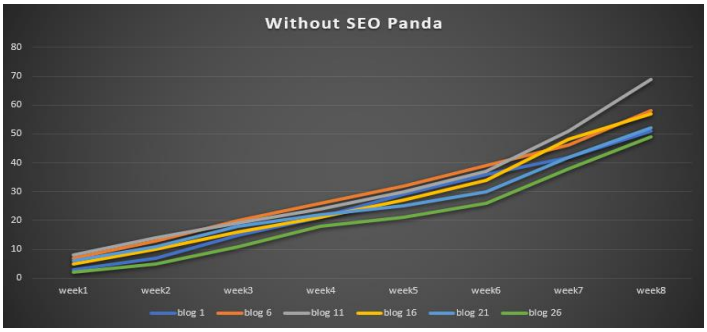
ALGORITHM – PANDA

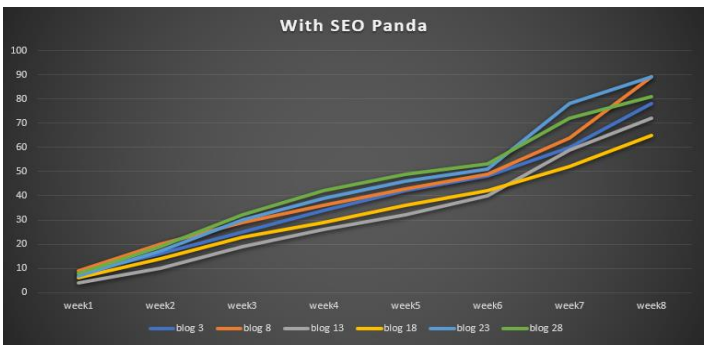
The stated purpose of the Google Panda algorithm update was to reward high-quality websites and diminish the presence of low-quality websites in Google's organic search engine results. It was also initially known as "Farmer." According to Google, Panda's initial rollout over the course of several months affected up to 12 percent of English language search results. We've tracked 28 data updates to Panda between 2011 and 2015.

THE PANDA ALGORITHM UPDATE ADDRESSED A NUMBER OF PROBLEMATIC PHENOMENA IN GOOGLE SERPS, INCLUDING:

- Thin content
- Duplicate content
- Low-quality content
- Lack of authority/trustworthiness
- Content farming

THE RESULTS





VIII. FINAL OUTPUT

Search via blogs:

1)Searched "Mobile" in search bar.

CONCLUSION

As the amount of text keeps growing, it becomes increasingly difficult for humans to process the deluge of information in the time available. It does result in consumption of a huge amount of duration. As a part of low efficiency, the subject propagated within the blogs may not be reachable to end users. Blog Mining is overcome by performing text routing techniques like text summarization and Search engine optimization. The overloaded content which is extracted from a huge number of posts provides low search results, as a process of application of summarization providing faster pace of quality search.

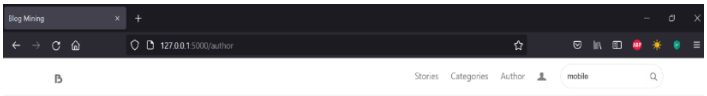
REFERENCES

[1] Olaf Sporns (2007) Complexity. Scholarpedia, 2(10):1623.

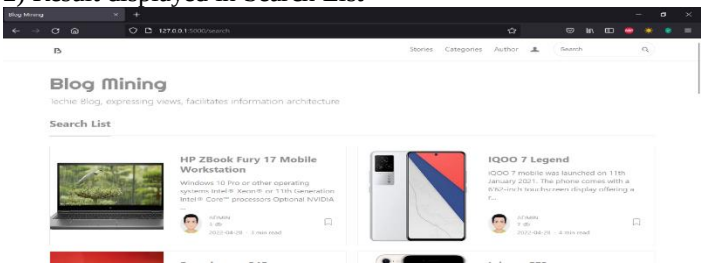
[2] Tomasz Downarowicz (2007) Entropy. Scholarpedia, 2(11):3901.

[3] Landauer, T. K., and Dumais, S. T. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. Psychological Review, 104, 211-240.

[4] Berry, M.W., and Browne, M. (2005). Understanding Search Engines: Mathematical Modeling and Text Retrieval, Second Edition, SIAM, Philadelphia.



2) Result displayed in Search List



[5] International Journal of Computer Applications (0975 – 8887) ,Volume 107 – No. 21, December 2014.

[6] Joeran Beel. Bela Gipp, Erik Wilde, Academic Search Engine Optimization ASEO. Optimizing Scholarly literature for google scholar & Co. 176-190 january 2010.

[7] Vol . 1, Issue 2, 2015, pp.1- Suresh Gyan Vihar University Journal of Engineering & Technology (An International Bi-Annual Journal) Vol . 1, Issue 2, 2015, pp.1-5 ISSN: 2395-0196 ISSN: 2395-0196

[8] IEEE TRANSACTIONS ON PROFESSIONAL COMMUNICATION, VOL. 56, NO. 1, MARCH 2013

[9] International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-0181 Vol. 3 Issue 4, April – 2014.

[10] International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-0181 www.ijert.org Vol. 2 Issue 6, June – 2013.

AUTHORS

Dr. V Shanmukha Rao[1] M.Tech, Ph.D., Associate Professor, Department of IT.
Vikram Pagadala[2] B.Tech, Andhra Loyola Institute of Engineering & Technology
MD. Ibraheem[3] B.Tech, Andhra Loyola Institute of Engineering & Technology.
CH. Roop Kumar[4] B.Tech, Andhra Loyola Institute of Engineering & Technology.