**TOPIC: MACHINE LEARNING AND DEEP LEARNING APPROACHES IN DOMAIN SPECIFIC SEARCH ENGINE OPTIMIZATION.**

*TEAM: Priyanshi Kathuria – 18CSU161*

*Puru Behl – 18CSU162*

*Rithwick Negi – 18CSU173*

**ABSTRACT**

In this research paper we will find out the ways how SEO works and use various tools and technique to show us optimize results. The main objective of this research paper is to find how SEO can help to find desired text and image between many web pages and give best results. SEO is the art of improving a website’s visibility for the users. The art of improving a websites visibility for search engines is possible using the techniques and tools available.

**INTRODUCTION**

Search engine optimization is the process of increasing the quality and quantity of website traffic by increasing the visibility of a website or a web page to users of a web search engine. SEO refers to the improvement of results and excludes direct traffic, visitors and the purchase of paid placement.

For example, Machine learning helps Google understand what users find useful and enables the search engine to rank websites more effectively. You need to create content that is high-quality, useful, relevant, and meets user intent. If your content engages the audience, you can easily dominate the SERP.

The image data has been increasing on the net making it hard for us to sort out the similarities between so many images. Here we will try to find out the ways to find out the images by not just the text linked to the images but also by the copy of the images that are being compared to. Only Instagram daily adds 52,000,000 which is a really high amount of the data in which the searching can be a very difficult task. So here various SEO tools are used to get best results.

How does the Google image search works? Google image search tries to find the similarity in the background scene of the image tries to recognize the place in the background, tries to find the famous personality in the pictures. But if you are working with this in the incognito mode there is a high chance for inaccuracy because the Google image search uses the user history for finding the similar images.

Search engine technology has gone through several evolutions and finally reached the point where the concept of deep learning/ AI can offer tremendous help. In this paper we have review various algorithms for SEO text for giving better results and improving a website’s visibility on search. This evolution from the beginning up to now and surveyed several different techniques that have been developed to improve search engine functionality. In particular we plan to highlight some machine learning approaches to information retrieval on the web and concentrated on topic-specific search engines. Finally, we will also propose an information integration environment based for active learning. Our approach will technology in a better manner to provide appropriate results.

The methodology used in this paper goes beyond the simple website interface and try to analysis and use web crawling as a source for collecting website performance data and employed web technologies and servers. The paper complements this technical perspective with a proposed scorecard used to assess the efforts of banks in Internet presence that reflects the banks’ commitment to Internet as a distribution channel. An innovative approach based on Machine Learning Techniques, the K-Nearest Neighbor Algorithm, is proposed to estimate the Internet Popularity.

Search Engine Evaluation is a complicated task therefore many interconnected factors has to consider while determining the efficiency. It is often subjected with a set of possible actions (options/decisions) for each queries and each result has associated possible (short-term and long-term) consequences, which are uncertain. Probabilities on the extraction of the result in term of information retrieval while provide quantification of evidence to handle uncertainty in the query during the query expansion and web document exploiting. Those processes can be implemented with help of ML algorithm. The ongoing research is to evaluate various search engine which uses the data mining approaches to extract the information in the web repositories.

**SEO FOR IMAGE**

The image data have increased tremendously on the internet making it really difficult for us to sort out the similarities between various images. Here we will try to find out the ways to find out the images by not just the text linked to the images but also by the copy of the images that are being compared to. There are some ways to find the similar data on the web for now for example the Google image search, it is able to find the most similar images related to the text that was added in the search box. But it comes up with the loop hole of incognito mode where this model fails and could not show better results.

**Working of Bing image**

Bing images allows you to find sites on which the image we are looking for is present. With the task of finding sites it does a really good job and shows positive results.

**The proposed approach**

An effective solution to the image search problem requires the development of the new methods to find the similar functions and use the latest achievements in the field of engineering, deep learning and the computer vision to evolve further.

Search by image is based on the use of the computer vision. The success of such a search is based on the use of informative invariant rapidly computed features, as well as effective image classifiers.

The search effectiveness is based on the time complexity of how fast the search occurs no one wants to wait a long time for the search engine to work for. So there is a need for structure in the data this can be done by various methods additionally by adding some more information given by the user for example geo-location of the user making the search and search related to the nearby people in the area will give a more reasonable search to the image. All the values required for this shall be stored with the description of the image, this maximizes the effectiveness of the image search.

1. **Searching images using color content**

Searching the image using the color content is usually done by making the histogram of the images which gives the intensity levels of different colors used in the images. This is not the best method since there is a possibility of having the similar color intensity but in different regions making the images totally different but still advantages of this type of method is usually the speed of implementation as well as the simplicity of the method.

1. **Searching image using texture similarity**

In this method we try to compare the different textures in the images and their respective positions in the image. Textures are usually hard to describe but are the changes in the color intensity in an image. These are usually represented as the two dimensional array of brightness changes.

1. **Using the perceptive hash**

In this approach discrete cosine transform Is used to convert a huge image to a smaller 64 bit hash image now due to small size comparing the hash of 2 different images becomes really easy. There are different ways in which these hash are compared one such way is the Levenstein distance, Hamming distance.

1. **Searching images using image description and meta data**

There are many methods which tries to find similar images from the websites which provides metadata about those image. The search engine tries to find closest metadata to the search and links the image related to it and shows it the the person searching it. Example of this type of image searching is web seer which uses this kind of algorithm to search the images

**Proposed Approach**

Using information about geo-location of the user and search history information from the nearby users to that geo-location to give an appropriate search. It has been seen that people tend to be similar to their friends making them search things which are of similar interest of their friends making the search much more structured and better.

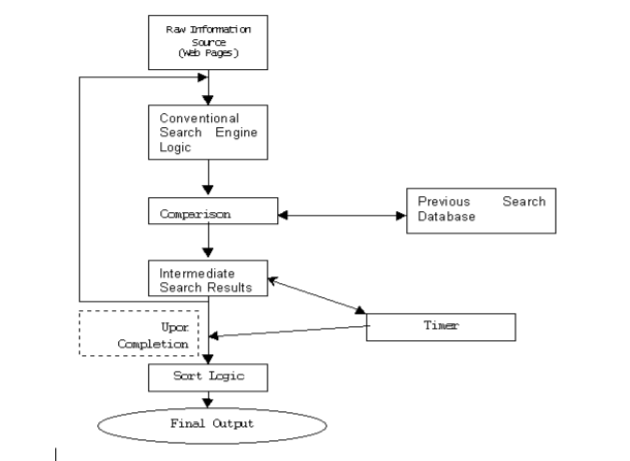
**Algorithm**

1. Find the user geo-location.
2. Find 10 users near that geo-location site.
3. Check the history of the user as well as the 10 nearest users for the similar sites or searches for previous month.
4. Convert the search to embedding using Bert model and Bert tokenize which also gets the semantic meaning of the search.
5. Try to cluster the searches using K means.
6. Find the cluster related to the search.
7. Score the similarity in the searches in the cluster using some metrics like hamming distance
8. Find the top similar search.
9. If the top searches are sites try to find the closest image using metadata of the image
10. Present the image to the user.

**SEO FOR TEXT**

Search Engine optimization is the technology of improving a website’s visibility on search. It is the art of creating fast and robust websites which appear on the top searches to a user and provides useful, important and required data. We can extract useful data only by using appropriate keywords and Meta description in our content which gets traced by the internet crawlers. This is done extensively by machine learning and deep learning methods which have gained popularity in the past 50 years. The image below explains the information retrieval architecture which makes us of keywords heavily.

Page rank is directly impacted by the number of HTML tags for a particular website, thus important concepts in this area should be examined. These concepts include Number, Proximity and Density. For extract required data for this research we design a crawler and parser. Then dataset has been learning by ANN. Then, this ANN is used to predict new websites rank that has good performance on test data.

These factors are divided into five groups of factors related to the key words, factors related to the link, factors related to site, factors related to the page and factors related to laws of specific algorithms that the factors related to keywords are the most important. So many website optimization techniques have focused on the extraction of appropriate keywords from a website to enhance the site's rank in search engines.

**Proposed method**

1. **Factors based on keywords**

In this section we shall discuss the relevance of an internet crawler and the requirement of an appropriate keyword. The concept of keyword is further broken down into 3 technical aspects of: number of keywords, density of keywords, and Prominence of keywords.

**Number of keywords**: Higher number of key words with regard to all its circumstances will increase rank.

**Density of keyword:** It is a measuring criterion that shows the percentage or actual number that a keyword is appeared compared to the total number of words of text. The more the total number of words used in the text is greater; the number of keywords can be increased.

**Prominence of keywords:** It is a measuring criterion which measures the position of the words in HTML tags. Keywords that used earlier in tags are more prominent. A formula that we will use to calculate prominence is as follows:



Identifying these keywords: In order to move to the crawler stage, it is imperative to determine the appropriate and varied keywords in several specific areas. These words can be determined without any rules and restrictions.

1. Crawler:

The concept of the crawler is represented by a set of example pages that is provided by a user to the system. In the system described there is a user brows-able topic taxonomy where the user can mark some of the documents as good and select them as the focus topic. The system has three main components: A **classifier** that makes judgments on the relevancy of crawled documents, and decides on following the links within pages. The classifier is an extended version of the Naïve Bayes classifier. The second component is a **distiller** that evaluates the centrality of each page to determine crawling priority of links within it. The distiller uses the bibliometric concepts of hub and authority pages as an approximate social judgment of web page quality.

**Problem Statement**

Let’s throw light on some of the problems that our use case incurs when we let it under SEO techniques.

• At the moment a search engine is represented as a set of intelligent agents (Internet bots, crawlers, parsers, etc.);

• Web site content is a set of keywords and files, merged semantically. The amount of keywords and their combinations are unlimited which makes the problem range massive;

• Web content is not stable. It is being changed. Thus, in order to reach the first position in the list of search engine answers we should change the web content step by step. In other words we correct the set of keywords the same way we transform the site from one state (situation) to another.

Thus, the way to solve this task is to build correct semantic net of web content of our web site. Actually this approach is well-known among SEO specialists. Based on classic learning theory such task can be solved if we have the following elements:

─ A proposed model,

─ Training samples, and

─ Learning method.

Let us consider in the capacity of the model of knowledge representation – a semantic net. Therefore, we believe that inside a search engine the semantic net is formed to describe predefined combination of question and answer. This combination leads us to the list of web sites, where our site is placed on the first position.

**PROPOSED WORK**

In order to curb the above problem(s) we use a directed Semi-supervised learning model. There we proceed to evaluate small sample data in a labeled document. Our goal is to guess the label of an unseen document. In this category we review learning from labeled and unlabeled documents

We must measure the goodness of our action for it to be evaluated in the best way possible. Thus, to enhance our domain specific research, we replace the classifier with active learner.

An Active learner is capable of predicting the hypothesis with the experiences. Semi-supervised learning is used because it is very expensive to generate labeled data for every set. We do so to improve the efficiency of the system. We use the following semi-supervised algorithms for text classification:

• Co-Training

• Semi-supervised EM (Expectation Maximization)

• Co-EM

• Co-EMT uses a multi-view active learning algorithm

Multi-view setting applies to learning problems that have a natural way to divide their features into subsets (views) each of which are sufficient to learn the target concept. Multi-view active learning maximizes the accuracy of the learned hypotheses while minimizing the amount of labeled training data.

(\*\*The EM algorithm, is a statistical method for maximum likelihood parameter estimation.)

The algorithm to improve the efficiency is discussed as follows:

**Given**

– A problem for learning with view

– An algorithm for learning

– Set of “U” and “T” unlabeled and labeled examples

– No. odd iterations to be performed = k

– The number of queries to be made= n

Our algorithm contains three main part:

(1) **Active learner** that checks the relevancy of crawled documents. Its task force contains of maintaining the history and makes hypothesis based on them then it allows meta- reasoning to be done.

(2) **Distiller** checks the authenticity of the webpage traversed to determine crawling priority of links within it. It looks into the web quality and judges the authority pages.

(3) **Dynamic crawler** uses the re-orderable priority queue to crawl the web.

**CONCLUSION**

Webpage rank plays an important part to search the documents for the specific keywords. We can say PageRank is a very critical segment of SEO. But page rank have certain limitations and other factors and algorithms also plays important role for giving the best results, which we have discussed in our research. Machine Learning algorithms are designed and implemented to measure the search engine efficiency on large exploring web document so that user can find relevant results quickly and efficiently. In future, we can also use some AI techniques in addition to improve the rank of web page.

**REFRENCES**

[1]. Banaei, Hengameh, and Ali Reza Honarvar. "Web page rank estimation in search engine based on SEO parameters using machine learning techniques." *Int J Comput Sci Netw Sec* 17 (2017): 95-100.

[2] Sharma, Shilpy. "Information retrieval in domain specific search engine with machine learning approaches." *World Academy of Science, Engineering and Technology* 42 (2008).

[3] Noh, Hyeonwoo, Paul Hongsuck Seo, and Bohyung Han. "Image question answering using convolutional neural network with dynamic parameter prediction." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

[4] Mishra, Richa, and Surya Prakash Tripathi. "Deep learning based search engine for biomedical images using convolutional neural networks." *Multimedia Tools and Applications* 80.10 (2021): 15057-15065.

[5] Mustak, Mekhail, et al. "Artificial intelligence in marketing: Topic modeling, scientometric analysis, and research agenda." *Journal of Business Research* 124 (2021): 389-404.

[7] Frankel, Charles, Michael J. Swain, and Vassilis Athitsos. *Webseer: An image search engine for the world wide web*. Technical Report 96-14, University of Chicago, Computer Science Department, 1996.

[8] Noh, Hyeonwoo, Paul Hongsuck Seo, and Bohyung Han. "Image question answering using convolutional neural network with dynamic parameter prediction." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

|  |
| --- |
|  |
| [9] Matošević, Goran, Jasminka Dobša, and Dunja Mladenić. "Using Machine Learning for Web Page Classification in Search Engine Optimization." *Future Internet* 13.1 (2021): 9. |

[10] Lei Ba, Jimmy, Kevin Swersky, and Sanja Fidler. "Predicting deep zero-shot convolutional neural networks using textual descriptions." *Proceedings of the IEEE International Conference on Computer Vision*. 2015.

[11] Smelyakov, Kirill, et al. "Search by image. New search engine service model." *2018 International Scientific-Practical Conference Problems of Infocommunications. Science and Technology (PIC S&T)*. IEEE, 2018.

[12] Saraç, Esra, and Selma Ayşe Özel. "An ant colony optimization based feature selection for web page classification." *The Scientific World Journal* 2014 (2014).

[13] Goel, Shikha, and Sunita Yadav. "An overview of search engine evaluation strategies." *International Journal of Applied Information Systems* 1.4 (2012): 7-10.

[14] Ren, Mengye, Ryan Kiros, and Richard Zemel. "Exploring models and data for image question answering." *arXiv preprint arXiv:1505.02074* (2015).

[15] Jensen, Richard, and Qiang Shen. "Semantics-preserving dimensionality reduction: rough and fuzzy-rough-based approaches." *IEEE Transactions on knowledge and data engineering* 16.12 (2004): 1457-1471.

[16] Ochoa, Edgar Damian. *Analysis of the application of selected search engine optimization (SEO) techniques and their effectiveness on Google's search ranking algorithm*. Diss. California State University, Northridge, 2012.