Third International Conference on "Materials, Computing and Communication Technologies"
in association with International Journal of Scientific Research in Science,
Engineering and Technology
Print ISSN: 2395-1990 | Online ISSN : 2394-4099 (www.ijsrset.com)

Smart Search Engine

Dr. V. Vallinayagi M.Sc. M.Phil. Ph.D.[1], S.Vinitha[2]

[1]Associate Professor and Head, Department of Computer Science, Sri Sarada College for Women (Autonomous), Tirunelveli, Tamil Nadu, India

[2]II M.Sc. Computer Science, Department of Computer Science, Sri Sarada College for Women (Autonomous), Tirunelveli, Tamil Nadu, India

ABSTRACT

Machine learning algorithms based on expert knowledge are used to categorize web pages into three predetermined categories depending on the degree of content modification to the search engine optimization (SEO) suggestions. We used classifiers to categorize an unknown sample (web page) into one of three predetermined groups and to find crucial elements that influence the degree of page modification in this research. In the training data set, the data is manually labeled by experts in the field. Using machine learning, it is possible to forecast the level of conformity of web pages to SEO guidelines. Building software agents and expert systems that can automatically identify web pages that require change in order to comply with SEO criteria and, therefore, possibly obtain better search engine ranks is the practical importance of the suggested technique. Measuring the semantic similarity search between words is an important component in various tasks on the web such as relation extraction, community mining,clustering and automatic metadata extraction. In addition, the findings of this study contribute to the research area of determining the best values for ranking variables used by search engines to rank websites. According to the conclusions of prior study, page titles, meta descriptions, H1 tags, and body content are all crucial aspects to consider when creating a web page. As a by-product of our investigation, a new collection of manually labeled web pages has been created. Web service discovery has received considerable attention in the literature, and academics are continually working to improve the process. Using machine learning methods such as KNN (K-Nearest Neighbor) and OCR (Optical Character Recognition), this study examines the work of a number of prominent researchers in this field (Optical Character Reorganization). Researchers have a lot to look forward to in machine learning as a way to consistently deliver correct estimations. From completed project training sets, a machine learning system "learns" how to accurately predict future work. It is an aim that this publication will serve as a springboard for future research and provide researchers a sense of the direction in which they should be heading.

INDEX TERMS: Natural Language Processing, Machine Learning, K Nearest Neighbor, Optical character recognition, Lexical Pattern Extraction, Ranking, Search Engine Optimization, Semantic relations, Targeted traffic SEO, Snippets.

I. INTRODUCTION

Search Engine Optimization (SEO) is a method to get a better ranking for a website in search engines such as Google, Yahoo or Bing. A search engine optimization campaign pairs on-site optimization with off-site tactics which means that one makes changes to the site itself while they build a portfolio of natural looking back links to increase their organic rankings[1].When internet users search for the product or service related, the website relevance to specific keywords which internet users search for online. The process of optimizing search engine includes keywords, creating more content, building links and making sure that the website is visible in the search engines using Targeted Traffic SEO. Web search engines have become an important part. Searching some kind of information on the web became hectic.SEO procedure work includes two types of optimization techniques, on-page and off-page SEO optimization.

Both techniques have their personal, discrete and extensive processes to rank websites on top of search engines.[2],[3] The SEO process starts with on-page SEO optimization. Just the once the whole on-page SEO optimization is complete, the off-page SEO optimization starts. Off-page SEO includes tricks which are chosen to make relevant back links towards the website to make the web page appropriately in front of search engine spiders. Second off-page SEO is doing the responsibility to improve our site's search engine rankings outside of our site. The only thing one can do off-site to increase the rankings is building up more links. SEO Benefits: Popularity of Search engine technique popularity will increase, Increase Visibility once a website has been optimized, it will increase the visibility of a website in search engine. More people will visit the website.

All existing search engines adopt several techniques and approaches to improve the performance of the search engines but the answer is differ from one another. However, after evaluating the performance of search engines based on the retrieved web contents, it is apparent that only a few attempts were made to restructure the query, providing alternate queries or personalizing the web search. Data mining represents the integration of several fields, including machine learning, data visualization, statistics and information theory.[5]Clustering is an unsupervised algorithm, which requires a parameter that specifies the number of clusters k. Cloud Computing is a computing platform which is distributed in large-scale data centers, and can dynamically provide various server resources to meet the needs of research, e-commerce and other fields. Clustering is the task of dividing the data or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group than those in other groups.

Cloud computing is the availability of computer system resources, especially data storage (cloud storage) and computing power, without direct active management by the user. Cloud computing is the delivery of computing services—including servers, storage, databases, networking, software, analytics, and intelligence—over the Internet ("the cloud") to offer faster innovation, flexible resources, and economies of scale. Cloud computing is a virtualization-based technology that allows us to create, configure, and customize various applications via an internet connection.

II. EXISTING SYSTEM

The Existing system uses a straightforward method to calculate similarity between two words is to find the length of the shortest path connecting the two words in the taxonomy.[6] It will produce only the irrelevant

results matches to the user query. It uses the page counts to retrieve results but using page counts alone as a measure of co-occurrence between words presents several drawbacks. SEO is a long term process which means it is not fast moving. It takes lots of time and patience to see the desired results. Therefore, No guarantee exists for all the information and it is a need to measure semantic similarity between a given pair of words is contained in the top-ranking snippets.

Some of the disadvantages for this system are as follows

- Not an automatic extraction
- Very Low Efficiency
- It becomes difficult for the user to get the relevant content.
- It would take more time
- Irrelevant Results
- Page counts are unreliable.

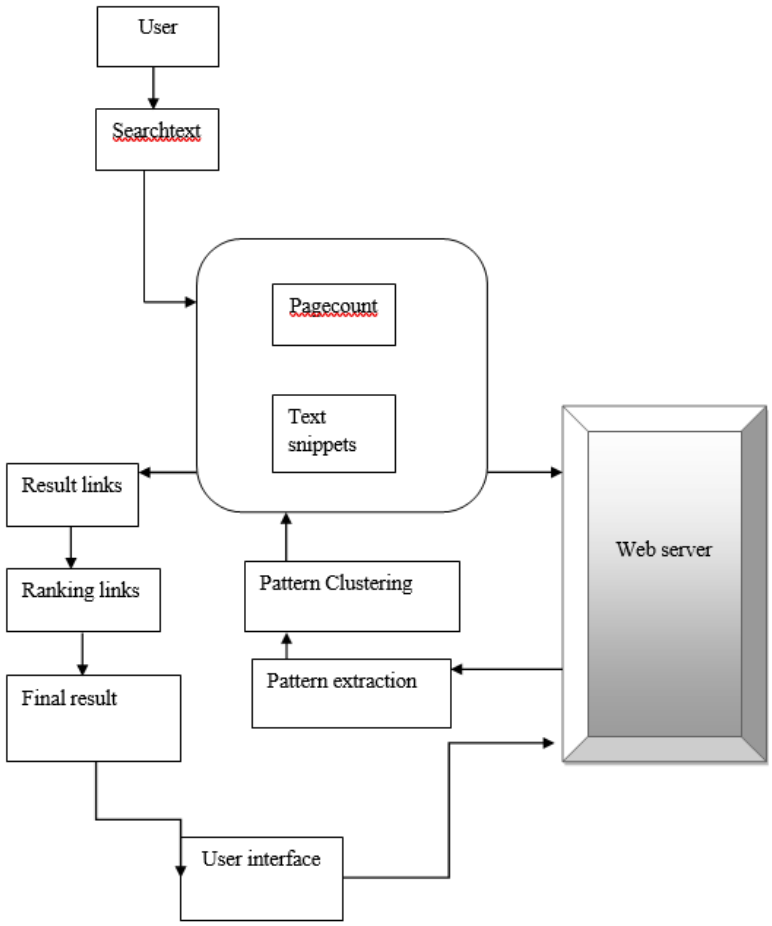
III. PROPOSED SYSTEM

This web search engine provides the most semantic relativity between the given words, and it will generate the semantic measures automatically.[7]It is time consuming to analyze each document separately. Web search engines provide an efficient interface to vast information. Page counts and snippets are two useful information sources provided by most web search engines. And then train a two-class support vector machine to classify synonymous and non-synonymous word pairs.

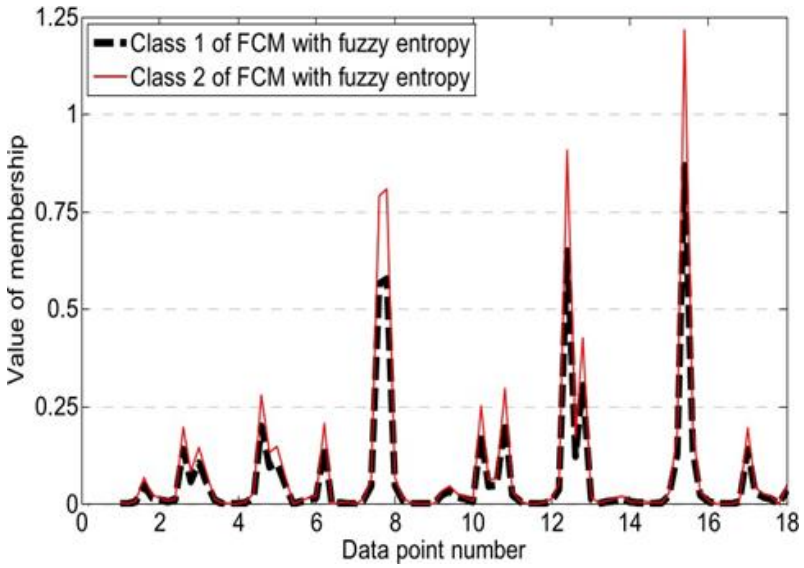
[8]A synonymous word having the same or nearly the same meaning as another in the language. Both novel pattern extraction algorithm and pattern clustering algorithm outperforms well in the case of page counts for given words with the text snippets. A search engine is an information retrieval system designed to help find information stored on a computer system. The search results are usually presented in a list and are commonly called hits.[9],[10] Search engines help to minimize the time required to find information and the amount of information which must be consulted, to other techniques for managing information overload. Search engines provide an interface to a group of items that enables users to specify criteria about an item of interest and have the engine find the matching items. The system will cover entirely any text information that can be found on the internet.

Major advantages of the system are as follows

- Simplicity
- More reliable and efficiency
- Time consuming
- Most related data to the given words
- Ranking based results
- Semantic similarity based on words co-occurrences


User
Searchtext
Pagecount
Text snippets
Result links
Ranking links
Final result
Pattern Clustering
Pattern extraction
Web server
User interface

IV. RESULT


Class 1 of FCM with fuzzy entropy
Class 2 of FCM with fuzzy entropy
Value of membership
Data point number
1.25
1
0.75
0.5
0.25
0
0
2
4
6
8
10
12
14
16
18

The fuzzy entropy is applied to the seal impression problem to measure the subjective value of information. This model involves the value of membership and data point number should arrange in wave signal dedicator. The upward frequencies show the exact result of tiny Google if any user searches for anything it will give the exact result. Both vertical and horizontal scalability are related to the distributed architecture selected by the algorithm. Both vertical and horizontal scalability are related to the distributed architecture selected by the algorithm. KNN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories. Optical Character Recognition (OCR) based on AI and machine learning is a widely used to put a good visual Presentation.Making the text part more informative to the customer visiting the website with an improved website design.

V. CONCLUSION

When search engines recognized the distortive effects of keyword Meta tags, they changed their algorithms to ignore keyword Meta tags. Search result relevancy improved, and the problem was solved without regulatory intervention. Search engines naturally will continue to evolve their ranking algorithms and improve search result relevancy a process that, organically, will cause the most problematic aspects of search engine bias to largely disappear. This work proposes a semantic similarity measure using both page counts and snippets retrieved from a web search engine for two words. A lexical pattern extraction algorithm is used to extract numerous semantic relations that exist between two words. The purpose of data selection is to identify the data to be analyzed, reduce the processing scope, and improve the quality of data mining. Search Engine Optimization tools are an important consideration to help optimize a website for search engines. This algorithm tends to terminate iterative process quickly to only obtain partial optimal results.Tiny Google uses the address to retrieve information.It shows the result in ranking order.A search engine optimization campaign pairs on-site optimization with off-site tactics which means that one makes changes to the site itself while they build a portfolio of natural looking back links to increase their organic rankings. Targeted traffic Search Engine Optimization can increase the number of visitors to the website for the targeted keywords. It gives the correct result compare to existing system. Some of the most important areas to be analyzed are keywords, content, back links, domain and social media.

VI.REFERENCES

[1]. P. Resnik, “Using Information Content to Evaluate Semantic Similarity in a Taxonomy,” Proc. 14th Int’l Joint Conf. Aritificial Intelligence, 1995.

[2]. R. Rada, H. Mili, E. Bichnell, and M. Blettner, “Development and Application of a Metric on Semantic Nets,” IEEE Trans. Systems, Man and Cybernetics, vol. 19, no. 1, pp. 17-30, Jan./Feb. 1989.

[3]. G. Miller and W. Charles, “Contextual Correlates of Semantic Similarity,” Language and Cognitive Processes, vol. 6, no. 1, pp. 1-28, 1998.

[4]. M. Hearst, “Automatic Acquisition of Hyponyms from Large Text Corpora,” Proc. 14th Conf. Computational Linguistics (COLING), pp. 539-545, 1992.

[5]. D. Mclean, Y. Li, and Z.A. Bandar, "An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources," IEEE Trans. Knowledge and Data Eng., vol. 15, no. 4, pp. 871-882, July/Aug. 2003.

[6]. Kilgarriff, "Googleology Is Bad Science," Computational Linguistics, vol. 33, pp. 147-151, 2007.

[7]. M. Sahami and T. Heilman, "A Web-Based Kernel Function for Measuring the Similarity of Short Text Snippets," Proc. 15th Int'l World Wide Web Conf., 2006.

[8]. D. Bollegala, Y. Matsuo, and M. Ishizuka, "Disambiguating Personal Names on the Web Using Automatically Extracted Key Phrases," Proc. 17th European Conf. Artificial Intelligence, pp. 553- 557, 2006.

[9]. H. Chen, M. Lin, and Y. Wei, "Novel Association Measures Using Web Search with DoubleChecking," Proc. 21st Int'l Conf. Computational Linguistics and 44th Ann. Meeting of the Assoc. for Computational Linguistics (COLING/ACL '06), pp. 1009-1016, 2006.

[10].M. Pasca, D. Lin, J. Bigham, A. Lifchits, and A. Jain, "Organizing and Searching the World Wide Web of Facts - Step One: The One- Million Fact Extraction Challenge," Proc. Nat'l Conf. Artificial Intelligence (AAAI '06), 2006.