ISSN 2395-1621

A Synonym Based and Crawling Technique using Data Mining in SEO

^{#1}Sanjivani Konapure, ^{#2}Prof. Manisha Singh

*1sanjivanikonapure14@gmail.com

^{#1}M Tech Scholar, Department of Computer Science and Engineering, ^{#2}Prof, Department of Computer Science and Engineering,

Dhole Patil Collage of Engineering, Pune, India



ABSTRACT

The processing of extracting information is too much costly in terms of processor time. A distributed design approach is proposed in this paper. Advantage, limitations are also discussed for this approach. Our main objective in this paper is to design a type of search engine architecture in which processing of a search engine may be in distributed form by which the time or processing power may be reduced. We also present a synonym based data mining approach for SEO that makes the task of improving the ranking of the website much easier way and user will get answer to their query easily through any of search engine available in market

Keywords: Web Crawler, Structure, Policies, Cluster, Data mining, SEO.

ARTICLE INFO

Article History

Received: 17th January 2018 Received in revised form:

17th January 2018

Accepted: 22nd January 2018

Published online: 22nd January 2018

1. Introduction

1.1 Data mining is performed with following tasks:

Outlier Analysis:

It is the identification of records which does not match the usual patterns. They might be interesting that require further investigation. They can assist in fraud detection in credit card transactions in banking sector.

Association rule learning:

It searches for relationships between various attributes. For example a computer sale shop may discover that printers are frequently bought along with desktops. So providing a good discount on combination of both can enhance the sales. This technique is quite useful for customer and market analysis.

Clustering:

It is the process of grouping together values in the data that have similar patterns, but these patterns are not known in advance. For example, in a bank customer data, after analyzing the data we may form clusters of customers who make more than ten transactions per week and other who make less than 10 transactions.

Classification:

It is the process of grouping the data into different classes on the basis of previously known structures. For example, a bank customer data may be classified according to region.

Regression:

Attempts to find a function which models the data with the least error. It fits the data onto the function so that one value can be derived from another.

Summarization:

It providing a more compact representation of the data set, including visualization and report generation.

1.2 Applications of Data Mining

The data mining includes applications in following areas:

- Medical and health care.
- Banking and finance
- Retail/Marketing
- Fraud detection
- Customer Management
- Search engine optimization

- Telecom industry
- Computer Security
- Education And many more

1.3 Goals of Search Engine:

- (1) Quality-Means effectiveness can be defined as to retrieve the most relevant set of document for a query. Process text and store text statistics to improve relevance be used.
- (2) Speed-Means efficiency may be defined as a process queries from users as fast as possible For it specialized data structure should be used.

2. BACKGROUD

Search Engine Optimization is the procedure of improving the visibility or traffic of a website or webpage in search engine via the natural [6] or unpaid searched results by increasing SERP (Search Engine Results Page) ranking. Optimization may target different types of search like image search, local search, video search, academic search, new search, industry specific vertical search. It can also be define as the process of affecting the visibility of a website or webpage in search engine.

2.1 Importance of Data Mining in SEO

Internet is an immense, huge and dynamic data collection that includes infinite hyperlinks and volumes of data usage information-hence requires effective data mining. But huge data is still a challenge in knowledge discovery.

- Web pages are more complex than text data: Web pages have dynamic data and do not follow any uniform structure. Web pages contains huge amount of raw data that is not indexed therefore searching in web data has become more complex; time consuming and difficult.
- The Web constitutes high quantity of dynamic information: Web not only contains static data but also data that requires timely updating such as news, stock markets, live channels etc.
- Web users include different kinds of user communities: People from different communities have different backgrounds and use internet for different usage purposes. Many have different interests and lack knowledge of internet usage. Hence user gets lost within huge amount of data.
- Only a small portion of the Web's pages contain truly relevant information: A given

user generally focuses on only a tiny portion of the Web, dismissing the rest as uninteresting data that serves only to swamp the desired search results.

2.2 Approaches for search engine

1. Keyword-based search:

This includes search which use keyword indices or manually built directories to find documents with specified keywords or topics.e.g engines such as Google or Yahoo.

2. Querying deep Web sources:

Where information such as amazon. com's book data and realtor. com's real-estate data, hides behind searchable database query forms-that, unlike the surface web, cannot be accessed through static URL links.

3. PROPOSED SOLUTION

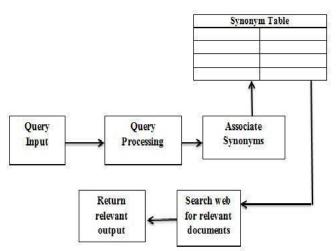


Figure 1: Synonym Based Architecture

3.1 Explain module:

1. Query Input:

User inputs the query on which he wants to perform the search. User will mention keywords relevant to his query.

2. Association of synonyms:

After the query is fed; it is then taken up for further processing. In this phase synonym table is looked up for the keyword entered and the corresponding synonyms are fetched from chained hash table. Each entry of synonym will contain the pointer to the next synonym of the keyword; search will end when null pointer is encountered. Hence the list of synonyms relevant to keyword is retrieved.

3. Matching the keywords and synonyms on web:

Web is searched for relevant documents based on keyword as well as synonym. Example-search for data mining will include search for data mining as well as knowledge discovery, classification, etc.

4. Displaying relevant results:

Documents found relevant to query are returned and displayed on SERP (Search Engine Results Page).

3.2 Crawling Techniques

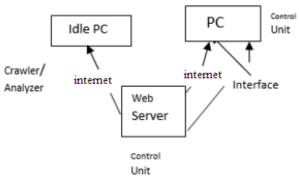


Fig 2. Crawling Techniques

This architecture should be capability of systems working in a Distributed manner. In it all processing should be processed on idle computer. The distributed architecture should not increase Network traffic. All the systems connected in a Network should be operated using some firewalls.

The module should be as much as easy to plug and play.

- (1) Crawler Unit:- It will crawl a website. It will need to use secondary memory to store the web pages downloaded before analysis. The web pages should be saved on each host system, rather than transferred to the control unit to minimize the network traffic. In crawler unit the technique of Data mining that is cluster may be applied by which similar data elements, similar URL's may be kept as a cluster. Cluster helps us to crawl the different pages. Different types of clustering algorithm may be used to crawl the useful URL's.
- (2) Control Unit:- When a crawler request a job or send some data elements the control unit will live on a web server and will be used by it. It will need to save the commands that the user like to be process. This can also be understudied by this example as doc file is saved on server.
- (3) Messaging System:- To satisfy the necessity of crawler and control unit the crawler must be able to download and process websites with less transmission with control unit-The crawler unit starting a crawl,

sending a message to control unit shows that it is ready to execute a new request. The control unit sending an instruction to the crawler showing a site to crawl and type of processing to be performed on downloaded website.

4. CONCLUSION

In this paper we study on a search based on synonyms. This approach can be further extended or improved by implementing synonym table in a more effective way which should include less space consumption and minimum access time. This SEO approach can increase the ranking of a website on SERP and benefit its owners and provide users with more accurate and relevant search results.

5. REFERENCES

- [1] Rajesh Singh and S.k Gupta, "Search engine optimization— Using Data Mining Approch", International Journal of Application or Innovation in Engineering & Management (IJAIEM) Vol 2 Issue 9 September 2013 pp 28 -32 ISSN 2319-4847.
- [2] Jiawei Han Kevin and Chen-Chuan Chang, "Data mining for web intelligence" IEEE Computers, 2002 Vol. 35 pp 64 -70.
- [3] Solihin, Niko, "Search Engine Optimization: A Survey of Current Best Practices" (2013). Technical Library. pp 151.
- [4] S. Chakrabarti et al., "Mining the Web's Link Structure," Computer, Aug. 1999, pp. 60 67.
- [5] Lalit Kumar and Naresh Kumar, "SEO Technique for a website and its effectiveness in context of Google Search Engine", International Jouranl of Computer Science and Engineering (IJCSE) Vol.2 Issuse 4, 2014. pp 113-118.
- [6] Peshave, Monica, and Kamyar Dezhgosha. "How search engines work and a web crawler application." Department of Computer Science, University of Illinois, Springfield USA (2005).
- [7]Edgar Damian Ochoa: An Analysis of the Application of selected SEO techniques and their effectiveness on Google's search ranking algorithm, California state university, Northridge, May'2012 csun-dspace.calstate.edu
- [8] Malaga, Ross A. "Worst practices in search engine optimization." Communications of the ACM 51.12 (2008) pp 147-150.

- [9] Vinit Kumar Gunjan, Pooja, Monika and Amit: Search engine optimization with Google, IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 1, No 3, January 2012, ISSN (Online): 1694-0814.
- [10] P. T. Chung, S. H. Chung and C. K. Hui, "A web server design using search engine optimization techniques for web intelligence for small organizations", LISAT IEEE Long IslandUniversity Brooklyn NY USA, (2012), pp.1-6.