JIMMA UNIVERSITY

INSTITUTE OF TECHNOLOGY

DEPARTEMENT OF COMPUTING

Data Mining Assignment one

Title:- Data mining in Search Engine

Group Name	**ID**
1. Addisalem Mekonnen……	RU0572/08
2. Lencho Kebata……	RU0516/08
3. Milkessa Abera……..……....	Ru0539/08
4. Bilisuma Shune……	RU4485/07

Introducation

World Wide Web provides us with huge amount of necessary data digitally available as hypertext Data may be Web Pages, images, information and other type. This hypertext pool is dynamically changing due to this reason it is more difficult to find useful information.

So Web Crawler for automatic Data and Web Mining is Useful to Us. The economic importance of web will enhance the academic interest. The Database Administrator, Management persons or others wishing to perform data mining on large number of web pages will require the services of web crawler or its based tools. For these reasons crawlers are normally multi threaded by which millions of Web Pages may be extracted parallel by only one process.

The growth of the Internet, its usage and dependency, leads towards various challenges.The Internet has opened up vast possibilities by opening the doors to Data control and access. It allows users to share their information and data through social networking site or simply by creating some Web Pages using different languages and technology.

Search engines have a unique policy for indexing information in an efficient manner, and it is essential to optimize web-pages in a specific way to enhance their search ranking. Search engine optimization is about modifying the web page accordingly to different parts of the website. These small modifications when viewed individually might make exponential improvements.

Search Engine

Search Engine provides the gateway for most of the users trying to explore the huge information base of web pages.

Search engines are programs that search documents for specified keywords on search for information on the World Wide Web and returns a list of the documents where the keywords were found. A Search Engine is really a class of programs; however, the term is often used to specifically describe systems like Google, Bing and Yahoo! Search that enable users to search for documents on the World Wide Web.

Computer program that search database and internet sites for the documents containing keywords specified by a user. It is a website whose primary function is providing a search engine for gathering and reporting information available on the internet or a portion of the internet. Most popular search engines include Google, Bing, Yahoo, AOL search, Ask, Web crawler, Dog pile, Lycos Alta Vista, etc.

Goals of Search Engine

- 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.
- Speed-Means efficiency may be defined as a process queries from users as fast as possible For it specialized data structure should be used.

Types of Search Engine

According to functioning three types of search engine .

- **Crawler Based Search Engine:**

They create their listings automatically. Spider builds them. Computer algorithm ranks all pages. These types of search engines are heavy and often retrieve a lot of information. For complex search it allows to search within the results of previous search and enable you to refine search results.

- **Human Power Directories:**

These are designed by human selection means they depend on professional to create listings. These never contain full text or web page they link to.

- **Hybrid Search Engine:**

These are different from traditional text oriented search engine such as Google or directly based searched engine such as Yahoo in which each program operates by comparing a sets of meta data.

Description of Basic Algorithm

Search Engines are inefficient if they are not able to find relevant information, to make search engines efficient SEO practitioners should use optimization of web pages as well as algorithm and search engine. Optimization is the process of improving the performance with the focus on time and accuracy. There are many techniques that are been proposed for optimizing the search engines. considered the best techniques existed with certain limitations and now overcoming them with suitable solutions using parallel processing and k -means clustering. In particular, they presented the importance of parallel processing the vector space algorithm for document indexing and incremental updating of means in k means clustering algorithm in the real world Scenario.

Cluster analysis is an important data mining technique used to find data segmentation and pattern information. By clustering the data, people can obtain the data distribution, observe the character of each cluster, and make further study on particular clusters. In addition, cluster analysis usually acts as the pre-processing of other data mining operations. Therefore, cluster analysis has become a very active research topic in data mining.

The approach is divided into two sections in section one the data collection is done by different software professionals and then the sub processes like data cleaning is done .This process involves removing of irrelevant data which can be termed as noise,thereafter missing and this irrelevant data is formatted according to the required format and a database or data warehouse is evolved. Second section generates the clusters according to the choices of the attributes. In this paper we had used k-means cluster analysis for creating the clusters.

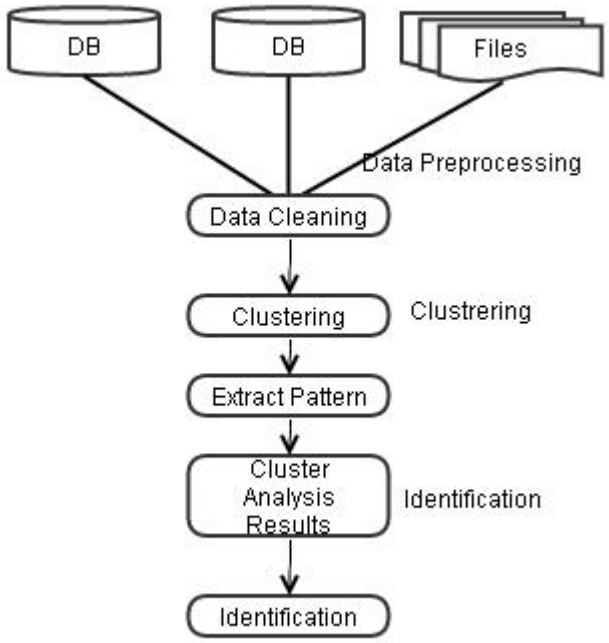

Figure 1: Displaying Clustering Process

VI. K-MEANS CLUSTER ANALYSIS

The initial step of k-means clustering is quite simple. It is based on minimum sum of squares to assign observations in a cluster .In the starting, we determine number of cluster which is represented as K and we assume the centroid of these clusters. We can take any random objects as the initial centroids or the first K objects can also serve as the initial centroids.

Then the K means algorithm will do the three steps below until convergence. Iterate until stable or no object move group:

1. Determine the centroid coordinate
2. Determine the distance of each object to the Centroids.
3. Group the object based on minimum distance (find the closest centroid).

k-means cluster analysis is used to create the clusters using a well known tool Weka 3.6. Weka provides us facilities so as to implement different data mining algorithms in very easy way.
Cluster analysis a techniques that is used to classify cases into small individual groups that are homogeneous within themselves and heterogeneous among each other, on the basis of their similar attributes and this classification is known as clusters
The clusters were created using partitioning or non-hierarchical clustering method i.e. K-means cluster analysis.

Application Area in 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

Problems Facing by Current Search Engines

- Crawlers are not able to analyze the content of keyword in web page before they download it.
- User submits his request for retrieval of information without mentioning the content in which he otherwise desire.
- Crawler treats user search request in isolation.
- There is a requirement to prepare separate files for each web document.
- Augmentation is required in HTML document.

Quality of Good Search Engine

- Ability to produce the most relevant result to any given search.
- A true search engine is an automated software program that moves around the web collecting Web Pages to include in its catalog or database.
- It searches when user requests information from a search engine has its own catalog or database of collected Web Pages, so you will get different results

Suggestion

- Different Data mining algorithm for clustering
- New and trending Search engine optimization techniques
- Strategies for Search engine optimization.

Conclusion

Data mining is implemented for finding some useful facts and patterns from different data sources. Use of data mining technique can help to understand and analyse data and information in proper way so that is will be helpful in different sectors. Search engine optimization is also a sector where there is a requirement of identifying some selective search engine optimization techniques from various SEOT so as to achieve a better rank in search engine result page.