A Survey on using Clustering to Enhance Search Engine Performance

Mennatollah Mamdouh Mahmoud
Information System department, Faculty of Commerce and Business Administration
Helwan University, Cairo, Egypt

Doaa Saad Elzanfaly
Information Systems department, Faculty of Computers and Artificial Intelligence
Helwan University, Cairo, Egypt

Ahmed El-Sayed Yacoup
Information Systems department, Faculty of Computers and Artificial Intelligence
Helwan University Cairo, Egypt

ABSTRACT
Searching the web seeking relevant documents or information becomes a difficult task. The reason for this problem is the availability of huge amounts of information on the web and poor indexing. So, it is necessary to manage the web resources to help people find the content they are interested in. As a result, the user needs a reliable information retrieval system to find and arrange the pertinent information. Many research studies are conducted to improve the performance of the web to find the most crucial information based on the query of the user. Some of them use ontology and semantic web to get the most relevant information to the user. Others use machine learning techniques, such as clustering to enhance the performance of the search engine. This paper provides a review about the search engine components, and the search engine index structure and ways to update it. This paper also reviews the clustering techniques, such as hard clustering techniques and overlapping clustering techniques, and the methods employed for labeling clusters. The different techniques that have been proposed to improved clustering techniques, cluster labeling and web search index are also discussed in this paper.

Keywords
Information Retrieval (IR), Search Engine, Clustering, Clustering Labeling, Web Document Indexing.

1. INTRODUCTION
The web serves as a platform for information sharing on a worldwide scale, making a vast amount of resources easily accessible online [1]. Retrieving information from the tremendous amounts of resources flowing because of daily technological advancements has gained popularity. This is because it helps in finding the important information stored in various types, such as structured, unstructured, or semi-structured types. Examples like databases, texts, multimedia, files, and the internet, among other things [2].
It is quite challenging to manage the features in the web due to the massive increase of the information available on the internet. The volume of the shared data keeps expanding without limit. People need tools for searching the Web to obtain information that they are interested in. As a result, for the user to locate and organize the relevant information, an effective information retrieval mechanism must be used [3]. Information retrieval is the process of providing user information needs that are written as textual queries. Document retrieval is the process of getting information from a web page, but the crucial task is getting the information the user is seeking for from a collection of documents [3]. Search engines are an application of the information retrieval paradigm [4].

A search engine is the tool that is required for retrieving information from the web [1]. It is a software program with a GUI (Graphical User Interface) that is made specifically for conducting web searches on the internet [5]. It is a web-based application that scans used keywords and phrases input by users to search across multiple websites, documents, and files before providing the results as a hyperlink to the relevant web pages. [6]. It collects, organizes, analyzes, and processes information from the web while giving the user a useful interface to access network resources.

In order to present the results and organize them by relevance, search engines scan through the billions of web pages that are available on the internet. [6]. The search engine, however, offers thousands of results that contain both irrelevant and relevant information. Therefore, based on the user's query, the search engine is utilized to find the relevant information on the internet [3].
Users encounter difficulties in finding useful information related to their information needs. This issue arises because of the search engine's exact matching of the user's search terms with the keywords found on each web page before displaying the results. Any search model will deliver search results as long lists of URLs that are exceedingly difficult to filter or get to the right pages. The user query should be used to organize and filter the output documents to prevent this issue. These have made it more difficult for data mining to automatically organize these electronic documents properly and efficiently. So that, extraction of data according to users need became a tedious task. As a result, retrieving relevant information became an essential technique to extract valuable information from the internet.

2. BACKGROUND
2.1 Model of search engine
Three elements make up a typical web search engine: a web crawler, an analyzer and indexer for documents, and a search procedure [7]. Figure 1 demonstrates the search engine components of a search engine. The following subsections discuss these three components.

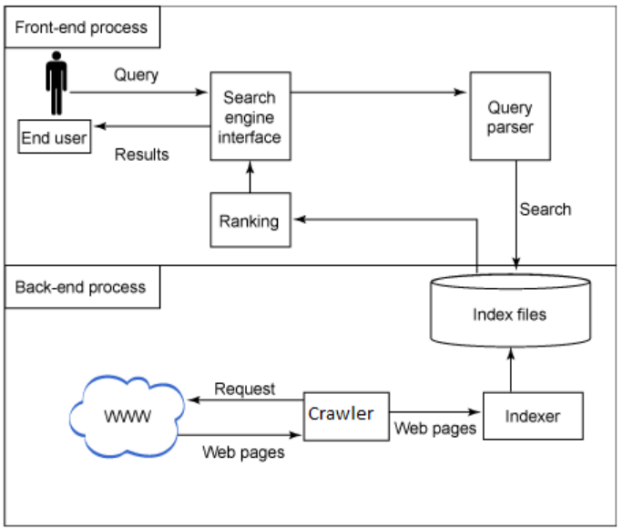

Fig 1: The architecture and key elements of a typical search engine [7]

2.1.1. A Web Crawler

Software that methodically and mechanically searches the web is known as a web crawler. Web crawlers are sometimes referred to as web robots, web spiders, ants, automatic indexers, bots, and worms. Each spider, however, indexes a site for its own objectives. Some people utilize the keywords from the META tags, while others might use the beta description of the website, and yet more people might use the first sentence or paragraph. Consequently, a page that operates well on one search engine may not operate so well on another. Repeatedly, a crawler will take one URL from a collection of "seed" URLs, download the corresponding page, extract all the URLs from it, and then add any new URLs it finds to the seeds. To operate effectively, web search engines gather information from numerous web pages and store it in their databases. A spider, a sophisticated web browser, retrieves these pages by clicking on each link it has saved or extracted. Then, each page's content is reviewed to determine how it should be indexed; for example, keywords are taken from headings, titles, or special fields known as Meta tags [7].

2.1.2. An Analyzer and Indexer for Documents

The spider's articles are organized into an index, which is a specialized file, as part of the indexing process. As part of the indexing process, data are gathered, broken down, and stored to allow for quick and accurate information retrieval. The construction of indexes incorporates interdisciplinary concepts from the fields of languages, mathematics, informatics, physics, and computer science. The purpose of an index is to increase the speed and accuracy of finding relevant files for a search query. Without an index, the search engine would have to scan every document it could discover on the Internet, which would take a lot of time and computing power and be impractical given how large the internet is now. For instance, a 10,000-page index might be searched in just a couple of seconds as opposed to hours for a sequential scan of every word in the documents. The additional computer storage space required to hold the index and the lengthened time required for an update are swapped for the time saved during information retrieval [7].

2.1.2.1. Data Structures for Indexing

The methods used for indexing and index storage vary between search engine structures. The most common structure for indexes is the inverted index. The list of occurrences for each atomic search criterion is stored in an inverted index, which is typically a hash table or binary tree. The inverted index is a common index structure. The inverted index is used by search engines like Google and Lucene to facilitate quicker term searches [8-10].An index structure known as an "inverted index" demonstrates a relationship between various keywords taken from a document and their associated posts lists. Each document has a unique identifier that serves as a representation of it. A collection of document ids is referred to as a posts list. A postings list for a keyword implies that all of the documents in the postings list contain the keyword [9].Numerous processes have been carried out through indexing. They are described as follows: [7].

• Extract URLs.
Extracting every URL from the article being indexed is a procedure that is used to direct website crawling, perform link verification, make a site map and a table listing the page's internal and external links. [7].

• Striping Codes
Is a method of HTML removal, character references and entities used to embed special characters, as well as deleting HTML elements, scripts, and styles [7].

• Language recognition.
Is a method in which a computer program makes an automatic determination of, or classification of, the language(s) used in the text [7].

• Tokenization of documents.
Is a procedure that involves identifying the page's encoding, figuring out the content's language (some pages use multiple languages), the word, sentence, and paragraph borders, the grouping of numerous words into a phrase, and the text's case. [7].

•Syntactic analysis or document parsing.
Is a process of determining a series with tokens' grammatical structure in relation to a specified (more or less) formal grammar by analyzing them (for example, words) [7].

• Lemmatization/stemming.
It is possible to reduce inflected (or occasionally derived) words to their stem, base, or root form during the indexing and/or searching phases. It suffices that related words map to the same stem, even if that stem does not qualify as a valid root in and of itself, rather than that the stem must match the word's morphological root. The technique aids search engines in dealing with challenges with query extension, indexing, and other facets of natural language processing [7].

• The Normalization
Is the process of making text in some way different so that it may be more consistent than it was before. Text normalization is typically done before text is processed in any form, such as creating synthesized voice, automatic language translation, storing in a database, or comparing [7].

2.1.2.2. Updating Inverted Indexes

After knowing about the index structure and the processes made on it to index the data, it is important to go through updating it. Updating index includes adding new web document to the index, changing the content of a web document or delete a document from the index. There are three naive updating methods for the inverted index: the forward index, document updating, and index rebuild techniques [11].

(I) Forward Index Update Method:
A forward index is a collection of postings that are connected

to a single document and contain its terms. The word id, document id, and word location id are all present in the list of postings in the forward index, unlike the inverted index. The word and location ids are sorted after the document id in the forward index. The building of the inverted index was sped up by using the forward index data-structure. Index updates are possible using it. A list of posting operations (insert posting and delete posting) that will change the forward index of the old document to that of the new document is prepared in response to the forward indexes of the old and new documents. To update the inverted index, two forwarded indexes are constructed for the document that has content modifications. These posting procedures are utilized to update the inverted index since the postings stored in a forward index are identical to the postings stored in the inverted index. When a content change occurs near the end of a document, postings corresponding to the words before the position of that change do not need to be deleted and reinserted, which is an advantage of using the forward index method over the document delete and insert method. The drawbacks of the forward index method include the necessity to construct an additional forward index for each document to be saved, the requirement that each forward index take up the same amount of storage as the document itself, and the fact that often there is little difference between the two documents. Take the addition of one word to the original document's commencement, for instance. All location ids after that will change by one, so all associated postings in the index must be changed [11].
(II) Document Delete and Insert Method:
In this method, the documents that have changed only are to be processed. All postings for the old versions of each document are removed from the inverted index and replaced with postings for the updated version. The benefits of this approach are that the index may be updated quickly, and online documents can be crawled at various intervals based on their pace of change. If the pace of change of the document has a wide range and the magnitude of change within an updated document is significant, this approach is advised. [11].
(III) Index Rebuild Method:
By scanning and sorting the entire updated document collection, a new inverted index is created, replacing the current one. The drawbacks of this approach include periodically traversing the whole document collection and scanning each word to create the inverted file. It is inefficient and unnecessary to scan the papers and re-index the words that did not change when the amount of change is minimal. Only when there are very few documents in the index or when a significant chunk of the document has changed or has to be added to the index is index rebuilding advised [11].

2.1.3. Searching Process

The last component of the search engine model is the search procedure, where web pages are searched in the index after constructing it.
A search engine receives a user's query (usually in the form of keywords), scans its index, and then returns a list of the best-matching web pages based on its criteria. Usually, the summary comprises the document's title and, on rare occasions, some text. In this phase, the results are ranked, with ranking defined as the relationship between a set of things such that the first item is either "ranked higher than," "ranked lower than," or "ranked equal" to the second item. In mathematics, this is known as a weak order or complete pre-order of objects. The rating of the documents is not necessarily absolute because it is possible for two distinct documents to have the same ranking. When ranking, factors including document popularity, recentness, and query relevancy are taken into account [7].

2.2 Clustering

Due to its uses for categorization, learning, summarization, and target marketing, clustering or unsupervised classification is one of the crucial research areas in machine learning [12]. Clustering is a powerful unsupervised machine learning technique for grouping objects or data into uniform groups with similar features, or "clusters." The use of clustering can increase the similarity of data or items inside a group while reducing the similarity of things between groups. It is utilized to provide a high-quality cluster so that information and patterns from the data may be extracted. [12-15].

Document clustering is based on clustering algorithms, which divide documents into useful groupings. There are three types of clustering algorithms: hard, soft, and disjunctive. Using hard clustering methods, each document is assigned to a single cluster. Soft clustering techniques only place each document in one cluster, however there is some doubt as to which cluster is the correct one. Some texts are assigned to numerous groupings by disjunctive clustering techniques [12, 13].

There are several areas where clustering can be used, including healthcare, agriculture, image processing, market research, pattern recognition, medical science, text-mining, document clustering, etc. [15, 16].

There are many methods for clustering. Partitioning clustering is one of the clustering methods.

2.2.1. Partitioning Clustering:

The partitioning approach separates the k clusters C1, C2,..., Ck from the n items. This method iteratively compares intra-cluster points' distances from the cluster-centroid until the clusters remain the same or the required number of iterations have been made. The clusters must meet the following requirements: (1) they must not be empty or contain at least one object; (2) each object must be contained in either one cluster (for hard clustering) or more than one cluster (for soft clustering); and (3) they must satisfy both constraints mentioned above. First, the k cluster centers are selected using this procedure. The distance between a data point and all the centers is then determined using a specific metric. After that, the object is placed in the cluster whose centroid is closest to it. The cluster centers are then moved towards an ideal outcome, or the target function is minimized using a relocation approach. For instance, in the K-Means method, the new cluster center is determined using the average value of all the cluster's objects, whereas in the K-Medoids approach, a cluster is represented by an item that is close to the cluster's center. The objective function that aims to attain high intra-cluster similarity and low inter-cluster similarity is used to assess the quality of clustering.
K-Means, K-Medoids, K-Modes, and other partitioning clustering algorithms are examples. [17, 18] .

2.2.1.1. K-Means Technique:

K-Means is considered a hard clustering technique. A dataset is divided into k non-empty clusters using the K-means method in a way that minimizes the sum of squared distances inside each cluster. There is only one cluster for each data point, and that cluster is not empty.[13, 14].

2.2.2. Overlapping Clustering

An important problem in clustering known as Overlapping Clustering is determining non-disjoint clusters [19]. A data point can be assigned to more than one cluster using the

clustering process known as overlapping clustering.[12, 20]. For several applications, including document clustering, it is advised to create overlaps between clusters to better fit hidden structures in the seen data. In addition to many other fields, categorization of images, videos, and audio also exists. Overlapping clustering methods provide a richer model for fitting existing structures in various applications necessitating a non-disjoint partitioning, whereas traditional clustering methods ignore the possibility that a data point can be assigned to several groups and produce k exhaustive and exclusive clusters representing the data [19-21]. Each cluster, for instance, corresponds to a collection of genes that are probably members of the same functional class in biological gene expression data. The underlying groups overlap naturally because genes might perform numerous roles [14, 19, 22]. Additional applications for overlapping clustering include model construction, data compression, document categorization, dynamic system identification, and document classification (documents belonging to various clusters) [12]. The currently used overlapping clustering techniques are expansions of traditional clustering models as partitional, hierarchical, generative, or graph-based models [19-21].

2.2.2.1. NEO-K-Means Algorithm

Replicated from K-Means is partitional, non-exhaustive, overlapping K-Means, or "NEO-K-Means." Assigning data points to the overlapped area of clusters and removing outliers are its key contributions [13, 14, 20, 21]. There are two steps in the NEO-K-Means algorithm. The estimated amount of data points for overlapping cluster regions comes first. Second, the appropriate clusters are assigned to the data points [14, 21]. With parameters that describe the levels of overlap and non-exhaustiveness, the objective function was a reformulation of the traditional k-means objective function [14, 22].

2.3. Cluster Labeling

After making clusters, it is important to give them labels to demonstrate the cluster topic or to give a brief explanation about the cluster content. The process of labelling groups of documents or words with representative names is known as cluster labelling. Once assigned, labels can be crucial in identifying the cluster contents, navigation, and search. The clusters' ability to be identified by sets of representative labels can increase the organization's usefulness. Cluster labelling is the process of giving a group of labels for each cluster in a document collection. It can give a helpful description of the collection as well as basic navigation and search support [23].

2.4. Using Cluster techniques to improve the search engine performance

Clustering techniques can be used to improve search engine performance. Clustering can be used to visualize the requested web pages by a users' query to makes it easier to spot the pages as in [24-28]. This is because clusters can determine underlying patterns in the data and provide similar documents in each cluster. Categorizing search results is important to the user to better visualize the result with a brief description for each cluster. Clustering techniques can also be used for indexing web pages specially in big data environment as in [29]. Using clustering in indexing web pages helps in organizing the web pages in the index, thus increasing the search speed as in [10, 30-32]. Increasing the speed of retrieving the required user query results and representing them in a form of categories could not be obtained without employing clustering techniques in the information retrieval system.

3. LITERATURE REVIEW

The most important literature and papers are discussed to get an overview about the research for improving the search engine performance. Several studies improved clustering techniques to get better clusters based on the document's similarity. Other studies proposed techniques to provide labels for the clusters to better describe them. Other studies enhanced methods of indexing documents to make it easy to search for the required document. Other studies used clustering techniques and other machine learning techniques to index the documents in a better way.

3.1. Improve Clustering Techniques

Chiheb-Eddine Ben N'Cir et al in [19] presented a description of the overlapping clustering method. They divided the overlapping clustering techniques into two groups: categories with uncertain membership and categories with hard membership. Fuzzy C-Means (FCM), Possibilistic C-Means (PCM), Evidential C-Means (ECM), and Belief C-Means (BCM) algorithms were examples of uncertain membership categories. Additive and geometrical based methods were the two different sorts of hard memberships-based approaches. Principal Cluster Analysis (PCL), Alternating Least Square Algorithms (ALS), Lowdimensional Additive Overlapping Clustering, and Bi-clustering ALS were a few examples of additive approaches. The Overlapping k-means (OKM), Overlapping k-Medoid (OKMED), Weighted Overlapping k-means (WOKM), Kernel Overlapping K-means (KOKM), and Parameterized R-OKM were examples of geometrical approaches.

Various overlapping clustering techniques were introduced by Said Baadel et al in [12]. OClustR, a graph-based technique for overlapping clustering based on relevance, was presented. Fuzzy C-Means (FCM), for example, presented probabilistic techniques where data objects were assigned membership degrees (values between 0 and 1) to clusters. Objects were given cluster membership assignments if their degree of affiliation exceeded a predetermined threshold [12, 19]. The k number of clusters and the s similarity criterion were inputs for the Overlapping Partitioning Cluster (OPC) algorithm. The MCOKE algorithm, also known as Multi-Cluster Overlapping K-Means Extension, had two steps. The usual K-means clustering was carried out as the first stage. The second phase produced a membership table that compared the matrix formed by the first K-means run step to maxdist (the greatest distance that a member might be from a cluster's centroid) [12].

Partitioning Around Medoids (PAM) algorithms, weighted K-Means (WOKM) techniques, and OKM, WOKM, and OKMED algorithms, respectively, were presented as extensions to the partitioning K-Means algorithm's overlapping case [12, 19, 33].

Joyce Jiyoung Whang et al in [14, 22] proposed an objective function that concurrently addresses the problems of overlap and non-exhaustiveness. "Non-Exhaustive, Overlapping K-Means" or simply "NEO-K-Means" was the name of the algorithm.

Tanawat Limungkura and Peerapon Vateekul in [20] improved the NEO-K-Means method for overlapping clustering described in [14, 22]. By incorporating distances between clusters into the goal function, they improved the NEO- K-Means method.

Tanawat Limungkura and Peerapon Vateekul in [21] improved the NEO-K-Means method for overlapping clustering described in[20]. They increased the amount of data points in

the overlapping area's estimation accuracy. They increased the cluster assignment's precision as well.

Eric Kerstens in [13] proposed using medoids rather than centroids, Non-Exhaustive, Overlapping K-Medoids (NEO-K-Medoids) was a novel technique based on the principles of NEO-K-Means.

3.2. Cluster Labeling

There were two approaches that can be used to label clusters based on labels source: an external source or internal source based on the documents of a particular cluster [34].

3.2.1. Internal Source or Cluster-based document-based approach

There were several ways to ensure that the cluster labels were internal and only apply to the documents in a specific cluster. [34].

M. Billah et al in [35] used derived phrases from sentences based on given unigram token as the cluster label. They used TF-IDF to get the important unigram tokens from a cluster. The problem in this method is that, taking the top 10 terms as the cluster label means that they took the top TF-IDF values. Highest TF-IDF mean that the word was the most frequently occurring in the cluster (as an English word) or it was rare. It was not reasonable to rely on rare words as the cluster representative or most frequently occurring terms as the cluster label. Also, the terms used in this approach consisted of only one word.

N. Niu et al in [36] compared three internal methods used to generate cluster labels. They compared the TF method, chi square method and Frequent & Predictive Words (FPW) method. The generated cluster label was one word term. The findings demonstrate that the TF cluster-internal technique obtained the least desirable sets of labels. However, in their trials, the chi square method rather than the hybrid method FPW yielded the best label sets.

Karel Gutiérrez-Batista Et al in [37] also used chi-square method for giving labels to clusters. The problem of using the chi-square method was that it was used in hierarchical clustering to give labels. While the used clustering method in this work was partitioning clustering method.

In [38] they called their method “The initial method”. They calculate the term frequency or TF as the past papers by using the important words within the text. But the IDF value was calculated based on the clusters number containing the word instead of the documents number. The important terms within each cluster were extracted using the normalized TD-IDF scoring approach by further discriminating the texts' cluster term content. They calculated the count vectorization of phrases with a presence in documents of less than 0.8. These top terms helped to identify the topic of each cluster. The problem with this method was losing the field or the domain of the cluster. Also, the presence of the words less than 80% in the cluster was not accurate, because low “ICF “meant that the term occurred in most clusters, and this was not discriminative to a particular cluster.

H. Kim et al in [39] proposed a fresh approach to labelling document clusters using key phrases found inside each cluster. By measuring the relative frequency of each word inside each document cluster and throughout other clusters, they were able to extract these keywords. This method's dependence on the cluster centroid's characteristics was an issue. K-means clustering, whose centroid was the mean values of the documents inside the same cluster, was the clustering technique employed in this study. In this procedure, the centroid was featureless.

3.2.2. External Methods

External methods employed external sources to generate labels for clusters. External sources include ontologies, wordnet, Wikipedia, …etc.

S. Reddivari in [40] presented creating labels using external knowledge from sources like Wikipedia, which was known as an external knowledge-based labelling approach. The five primary parts of the proposed framework were preprocessing, clustering, labelling, extracting Wikipedia labels, and label evaluation.

Sahand Vahidnia et al in [38] used Wikipedia as an external clustering labelled technique. A single 50-dimensional vector was created for each needed technique and application by parsing and vectorizing the Wikipedia pages for those topics. The labels for each cluster were then decided upon based on the top two closest methods and applications. The problem in using Wikipedia as an external way for cluster labels was that it was not an accurate ontology to depend on to get the cluster label or category.

Sujata R. Kolhe and S. D. Sawarkar in [41] used the Vector Space Model (VSM) with Singular Value Decomposition (SVD). The Latent Semantic Analysis (LSA) methods' fundamental mathematical foundation was the SVD. VSM with SVD were used to get the concepts that described the documents, and the documents were assigned to these concepts. Next, Wordnet was used to uncover further semantic clusters.

Tammishetti Vishnu and Konda Himakireeti in [42] used synonyms and hypernyms of the words to act as the cluster labels using Wordnet. The hypernym with highest frequency or top hypernyms was assigned as the label of the cluster.

Hanieh Poostchi and Massimo Piccardi in [23] provided a variety of methods to label clusters. They took advantage of WordNet lexicon synonymy and hypernymy relations as well as word embeddings. The problem of using Wordnet to construct the cluster label was that the cluster label lacked from the scientific label or the domain areas of the clusters.

3.3. Web Documents Indexing

Abdulla Kalandar Mohideen et al in [8] and [9] a new Graph-Based Indexing (GBI) method had been presented for big information systems. It made the use of a directed graph structure to keep track of when different keywords appeared simultaneously in the same text. The goal was to make use of the correlation between the search phrases as represented by the graph structure. This made it possible to quickly receive all the results of Boolean AND queries. The problem with this strategy is that GBI provided improvements when indexing and searching were done often rather than just once, as is the case here. The index in search engine applications needed to be updated often to deliver correct search results. Additionally, compared to the Inverted Index, the GBI required more memory and took longer to index.
Dilip Kumar Jang Bahadur Saini et al in [43] suggested a method of indexing based on an inverted index. The deleted file list module, primary inverted index, and appending inverted index made up the proposed index. The primary inverted index, which stored the data since the index's development. The delete file list was used to remove files from the index, and the appended index was used to add new data to the index. The

proposed index had the advantage of addressing the issue of the index's frequent updates. Additionally, any quick changes that were performed without recreating the primary inverted index. Unlike the traditional structure, which had one inverted index module, any modification to the corpus permanently reconstructed the entire index.

Nay Nandar Linn and Thinn Thinn Win in [44] recommended creating a search over RDF Graph model that was semantically enriched. It made use of semantic web's and information retrieval's top semantic search features. It used RDF and Knowledge Base (KB) as highly formalized semantic information that was integrated into traditional IR ranking methods. They looked at the concept of an IR model built on RDF with the intention of utilizing the KB domain. They combined the advantages of semantic and keyword searches.

Traoré et al in [45] suggested an approach based on multi-label classification (ML) that predicted the categories for newly created and tagged web pages. Both the learning phase and the prediction phase made use of ontology. The ontology was utilized to create the training set during the learning phase. The new pages were tagged in the most precise categories during the prediction phase using the ontology. The proposed method improved the Binary Relevance method of ML, according to experimental results using data obtained from the biological database Uniprot. The suggested method was expected to be a viable solution to the automatic web page classification problem in a semantic web platform due to its strong performance.

3.4. Using Clustering and Machine Learning Methods to Improve Indexing.

Aditi Bankura et al in [30] have proposed a zone-based indexing for storing data in databases and retrieving relevant web pages from them. The major goal of the suggested approach was to enhance the accuracy and speed of information retrieval and storage when data was distributed over multiple databases. The suggested model was created for search queries that were related to different types of content. The problem with this model was that it needed frequent updates to place the keyword in the appropriate zone according to its context. Additionally, there were no standards for identifying the domain or context of each zone.

Madhulika Yarlagadda et al in [10] presented the Rider Spider Monkey Optimization technique (RSOA), a document retrieval technique. The index was created using clusters, and the suggested Rider Spider Monkey Optimization Algorithm was used to locate the cluster centroids. The drawback of this approach was that document retrieval was independent of keyword search context. This was because to find the necessary documents, the search phrase was compared to other keywords of the chosen cluster.

Kabil Boukhari and Mohamed Nazih Omri in [31] presented a hybrid technique based on VSM and the partial matching between documents and biomedical vocabulary for indexing biomedical documents. The suggested strategy included two approaches, one statistical and one conceptual. The suggested method involved three steps: preparation, idea extraction, and filtering to retain only the relevant concepts for indexing.

4.CONCLUSTION

Retrieving relevant information from the internet was an essential technique to extract valuable information from the internet due to the difficulty of organizing and filtering the resulting documents according to user queries. In this paper, a comprehensive survey on web search engine and clustering techniques and how the search engine performance could be improved was introduced. To better summarize the main concepts of this field, some topics were discussed, such as search engine components, clustering techniques, types, and labeling. Based on these topics, the corresponding works in detail were discussed, including the proposed techniques, results, and limitations.

5. REFERENCES

[1] Shantanu Shahi, AkhileshShukla, and S. Rastogi, "SEARCH ENGINE TECHNIQUES: A REVIEW," *Journal of Natural Remedies,* vol. 21, pp. 48-55, 2020.

[2] P. P. Joby, "Expedient Information Retrieval System for Web Pages Using the Natural Language Modeling," *Journal of Artificial Intelligence and Capsule Networks (2020),* vol. 2, pp. 100-110, 2020.

[3] Lata Jaywant Sankpal and S. H. Patil, "Rider-Rank Algorithm-Based Feature Extraction for Re-ranking the Webpages in the Search Engine," *The Computer Journal,* pp. 1-11, 2020.

[4] C. C. Aggarwal, "Information Retrieval and Search Engines," in *Machine Learning for Text*, ed Cham: Springer International Publishing, 2022, pp. 257-302.

[5] R. S. T. Lee, *Ontological-Based Search Engine*, 2020.

[6] D. Sharma, A. K. Giri, R. Shukla, and S. Kumar, "A Brief Review on Search Engine Optimization," presented at the 2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 2019.

[7] Hussein Al-Bahadili, Saif Al-Saab, Reyadh Naoum, and S. M. Hussain, "A web search engine model based on index-query bit-level compression," *ACM International Conference Proceeding Series*, 2010.

[8] A. K. Mohideen, S. Majumdar, M. St-Hilaire, and A. El-Haraki, "A Graph-Based Indexing Technique to Enhance the Performance of Boolean AND Queries in Big Data Systems," presented at the 2020 20th IEEE/ACM International Symposium on Cluster, Cloud and Internet Computing (CCGRID), Melbourne, VIC, Australia, 2020.

[9] A. K. Mohideen, S. Majumdar, M. St-Hilaire, and A. El-Haraki, "A Data Indexing Technique to Improve the Search Latency of AND Queries for Large Scale Textual Documents," presented at the 2020 IEEE/ACM International Conference on Big Data Computing, Applications and Technologies (BDCAT), Leicester, UK, 2020.

[10] M. Yarlagadda, K. G. Rao, and A. Srikrishna, "Document Retrieval and Cluster Based Indexing using Rider Spider Monkey Optimization Algorithm," *International Journal of Recent Technology and Engineering (IJRTE),* vol. 8, pp. 1318-1327, 2020.

[11] L. Lim, M. Wang, S. Padmanabhan, J. S. Vitter, and R. Agarwal, "Efficient Update of Indexes for Dynamically Changing Web Documents," *Springer Verlag,* pp. 37-69, 2007.

[12] S. Baadel, F. Thabtah, and J. Lu, "Overlapping Clustering:

A Review," presented at the 2016 SAI Computing Conference (SAI), London, UK, 2016.

[13] E. Kerstens, "Non-Exhaustive, Overlapping k-medoids for Document Clustering," presented at the Proceedings of the 53rd Hawaii International Conference on System Sciences, 2020.

[14] J. J. Whang, Y. Hou, D. F. Gleich, and I. S. Dhillon, "Non-exhaustive, Overlapping Clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, pp. 1-14, 2019.

[15] A. Alam, M. Muqeem, and S. Ahmad, "Comprehensive review on Clustering Techniques and its application on High Dimensional Data," *IJCSNS International Journal of Computer Science & Network Security*, vol. 21, pp. 237-244, 2021.

[16] S. M. Mohammed, K. Jacksi, and S. R. M. Zeebaree, "A state-of-the-art survey on semantic similarity for document clustering using GloVe and density-based algorithms," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 22, pp. 552-562, 2021.

[17] A. N. Attri Ghosal, A. K. Das, S. Goswami, and M. Panday, "A Short Review on Different Clustering Techniques and Their Applications," presented at the Emerging Technology in Modelling and Graphics. Advances in Intelligent Systems and Computing, Singapore, 2020.

[18] V. Mehta, S. Bawa, and J. Singh, "Analytical review of clustering techniques and proximity measures," *Artificial Intelligence Review*, vol. 53, pp. 5995-6023, 2020/12/01 2020.

[19] C.-E. B. N'Cir, G. Cleuziou, and N. Essoussi, *Overview of overlapping partitional clustering methods*: © Springer International Publishing Switzerland, 2015.

[20] T. Limungkura and P. Vateekul, "Enhance Accuracy of Partition-based Overlapping Clustering by Exploiting Benefit of Distances between Clusters," presented at the 2016 Eighth International Conference on Knowledge and Systems Engineering (KSE), Hanoi, Vietnam, 2016.

[21] T. Limungkura and P. Vateekul, "Partition-based Overlapping Clustering Using Cluster's Parameters and Relations," presented at the 2017 9th International Conference on Knowledge and Smart Technology (KST), Chonburi, Thailand, 2017.

[22] J. J. Whang, I. S. Dhillon, and D. F. Gleichy, "Non-exhaustive, Overlapping k-means," *In SIAM International Conference on Data Mining (SDM)*, pp. 936-944, 2015.

[23] H. Poostchi and M. Piccardi, "Cluster Labeling by Word Embeddings and WordNet's Hypernymy," presented at the Proceedings of the Australasian Language Technology Association Workshop 2018, Dunedin, New Zealand, 2018.

[24] M. M. Joshi, "k-Means Clustering to Enhance SEO: A Data Driven Approach " *International Journal of Science and Research (IJSR)* pp. 550-553, 2020.

[25] M. SHAHROZ, M. F. MUSHTAQ, R. MAJEED, A. SAMAD, Z. MUSHTAQ, and U. AKRAM, "Feature Discrimination of News Based on Canopy and KMGC-Search Clustering," *IEEE Access*, vol. 10, pp. 26307-26319, 2022.

[26] T. Jenson and A. S. Girsang, "Performance of news clustering using ant colony optimization," *Journal of Physics: Conference Series* vol. 1566, pp. 12101-12108, 2019.

[27] Z. ZHANG, L. CHEN, F. YIN, X. ZHANG, and L. GUO, "Improving Online Clustering of Chinese Technology Web News With Bag-of-Near-Synonyms," *IEEE Access*, vol. 8, pp. 94245-94257, 2020.

[28] D. Bansal, R. Grover, and S. Saha, "A Multi-view Multiobjective Partitioning Technique for Search Results Clustering," presented at the 2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Melbourne, Australia, 2021.

[29] H. B. Abdalla, A. M. Ahmed, and M. A. A. Sibahee, "Optimization Driven MapReduce Framework for Indexing and Retrieval of Big Data," *KSII TRANSACTIONS ON INTERNET AND INFORMATION SYSTEMS*, vol. 14, 2020.

[30] A. B. A. K. S. Guha, "Zone based Indexing Model for Database Identification in Search Query Processing," presented at the 2020 IEEE 1st International Conference for Convergence in Engineering (ICCE), Kolkata, India, 2020.

[31] K. Boukhari and M. N. Omri, "Information Retrieval Approach based on Indexing Text Documents: Application to Biomedical Domain," presented at the 2017 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), Guilin, China, 2017.

[32] A. Curiel, C. Gutiérrez-Soto, P.-N. Soto-Borquez, and P. Galdames, "Measuring the Effects of Summarization in Cluster-based Information Retrieval," presented at the 2020 39th International Conference of the Chilean Computer Science Society (SCCC), Coquimbo, Chile, 2020.

[33] A. A. Aroche-Villarruel, J. A. Carrasco-Ochoa, J. F. Martínez-Trinidad, J. A. Olvera-López, and A. Pérez-Suárez, "Study of Overlapping Clustering Algorithms Based on Kmeans through FBcubed Metric," presented at the

Mexican Conference on Pattern Recognition, 2014.

[34] I. Peganova, A. Rebrova, and Y. Nedumov, "Labelling hierarchical clusters of scientific articles," *Ivannikov Memorial Workshop (IVMEM)*, pp. 26-32, 2019.

[35] M. Billah, M. Bhuiyan, and M. Akterujjaman, "The Unsupervised Method of Clustering and Labeling of the Online Product Based on Reviews," *International Journal of Modeling, Simulation, and Scientific Computing*, 2020.

[36] N. Niu, S. Reddivari, A. Mahmoud, T. Bhowmik, and S. Xu, "Automatic Labeling of Software Requirements Clusters," presented at the 2012 4th International Workshop on Search-Driven Development: Users, Infrastructure, Tools, and Evaluation (SUITE), Zurich, Switzerland, 2012.

[37] K. Gutiérrez-Batista, J. R. Campaña, M.-A. Vila, and M. J. Martin-Bautista, "An ontology-based framework for automatic topic detection in multilingual environments," *International Journal of Intelligent Systems*, pp. 1459-

1475, 2018.

[38] S. Vahidnia, A. Abbasi, and H. A. Abbass, "Document Clustering and Labeling for Research Trend Extraction and Evolution Mapping," presented at the EEKE 2020 - Workshop on Extraction and Evaluation of Knowledge Entities from Scientific Documents, Wuhan, China, 2020.

[39] H. Kim, H. K. Kim, and S. Cho, "Improving spherical k-means for document clustering: Fast initialization, sparse centroid projection, and efficient cluster labeling," *Expert Systems with Applications*, vol. 150, pp. 1-12, 2020.

[40] S. Reddivari, "Enhancing Software Requirements Cluster Labeling Using Wikipedia," presented at the 2019 IEEE 20th International Conference on Information Reuse and Integration for Data Science (IRI), Los Angeles, CA, USA, 2019.

[41] S. R. Kolhe and S. D. Sawarkar, "A Concept Driven Document Clustering Using WordNet," presented at the 2017 International Conference on Nascent Technologies in the Engineering Field (ICNTE-2017), Vashi, India, 2017.

[42] T. Vishnu and K. Himakireeti, "Automated Text Clustering and Labeling using Hypernyms " *International Journal of Applied Engineering Research*, vol. 14, pp. 447-451, 2019.

[43] D. K. J. B. Saini, PratapPatil, K. D. Gupta, S. Kumar, P. Singh, and M. Diwakar, "Optimized Web Searching Using Inverted Indexing Technique," presented at the 2022 IEEE 11th International Conference on Communication Systems and Network Technologies (CSNT), Indore, India, 2022.

[44] Nay Nandar Linn and T. T. Win, "Efficient Semantic Web Data Searching Using Virtual Documents Algorithm," *International Journal of Innovative Science and Research Technology*, vol. 5, pp. 298-303, 2020.

[45] Yaya Traoré, Sadouanouan Malo, Bassolé Didier, and S. Abdoulaye, "TOWARD MULTI-LABEL CLASSIFICATION USING AN ONTOLOGY FOR WEB PAGE CLASSIFICATION," pp. 183-191, 2019.