Appropriate Information Retrieval Tools for Efficient Data Searching

[1]EMMANUEL BOACHIE, [2]CHUNLI LI
Wuhan University of Technology
School of Computer Science and Technology, Wuhan University of Technology
Wuhan 430070, China. 008612360540137
[2]Kumasi Technical University, Kumasi, GHANA

Abstract: The retrieval of data and information housed in different media types is a day to day activity done by every technologist, scientist, and librarian from time to time in their work hours. Over the recent years, many tactical and technical progressive advances have been made that we are the witnesses to this new distributed sets of tools. Exponentially, the high growth rate in data archives has catalyzed the need for wiser techniques to handle extraction of this information. Graphical representation of gigantic databases is rapidly increasing based on the spatialized views. However, space as a data attribute has implications for using spatial concepts. This research paper covers the models and algorithms that can be used as an Information retrieval tools. Experiment is conducted to determine the one which is more appropriate for data searching. The algorithms under study are Ranking Algorithm, Cluster algorithm and Tokenization algorithms.

Keywords: Information, Data, Retrieval, Algorithm.

1. Introduction

Information recovery schemes were first developed to aid managing big scientific literature. Currently, many corporates adopt IR schemes to pave access to journals, books and generally other documents. Commercially embedded IR systems offer database solutions to millions of documents from diverse areas of specialization. The IR architecture is in two phases; storing indexed document and relevant document retrieval. It can be madly beneficial in current office work. An automated IR system must be able to 100% support some basic functionalities so as it can be termed as a success. It should have a means of entering the file, modifying them and if necessary delete them. It must also have means for searching and presenting the result to the user efficiently. This paper covers the essential algorithms and data structures used to develop IR systems and indicate the appropriate one which is more efficient for data searching.

In order to efficiently generating appropriate files by IR approaches, the files are distinctively converted into an appropriate representation. Every recovery approach integrates a particular model for its file representation objectives. The models are classified into two elements which are mathematical basis and the properties of the model. This research paper covers the models and algorithms that can be applied in Data recovery tools to ensure its effectiveness. The algorithms under study are Ranking Algorithm, Cluster algorithm and Tokenization algorithms.

2. Literature Review

Information retrieval is the science of searching for information in a document, searching for documents themselves, and also searching for metadata that define data, and for databases of texts, images or sounds. A data recovery process begins when a user keys a query into the system. Searches can rely on complete text or other indexing which is more content based. Queries are official declarations of data needs, for instance search strings in web search engines. In data recovery query does not exceptionally recognize a sole object in the gathering. Instead, many objects could match the query perhaps with varied ranks of appropriateness.

An entity is an object which is stood for by data in a content recovery or database. User queries are matched in contrast to the database information. Nonetheless, as opposed to classical SQL queries of a database in information recovery results returned might or might not be correspond with the query. So the results are typically graded. The ranking of results is a main difference of data recovery searching compared to database searching [1]. The application of the data objects may depend on text files, images, [2] audio, [3] videos or

Nevertheless, multiple researches concentrate on the anchor text which is probably the best place to find index terms. The anchor text holds higher class details of the pointed to a page [8]. This anchor text is also visible with citations. The two are both introduced aside with another descriptive identity to the cited document. It is not always the case that the terms in the document are indicators of what is important in the document cited. Retrieval of information basing on the index terms, however, raises a few eyebrows due to its oversimplification. A huge problem arises mind maps. The documents are not often saved or kept straight away in the IR system, but are instead represented in the system by document surrogates or metadata. A lot of IR systems calculate a numeric score on how well every entity in the database corresponds with the query and rank of the entities according to their values. The top ranking entities are then shown to the user. The process could then be repeated if the user wants to improve the query [4] [5] [6].

Intelligent Information recovery systems date back to early 19th century hence there have been more than enough research by scientists and technologist in this field. There are several means of cluster Algorithm implementation. Raising the standards of automatic indexing scientific papers can be done by locating index terms outside the set documents. Conventionally, perfect index terms are found by tracking the link structure between the set documents. There exit healthy literature on exhausting the link structure between the web-based documents for IR such as ‘sharing’ index terms among hyperlinked web pages [7] Extraction of Information from Public Health Emergency Web Documents

from the use of Index terms.

The dilemma in choosing which of the files is more important than the other. To fix this, a ranking algorithm is imposed. It operates on the basis of distinct basic premises regarding the idea of file relevance. The IR model that is adopted manages the predictions of which file is more relevant than the other. These IR models are of three types: Boolean, Vector, and Probabilistic. The Boolean model is more associated with index terms. This is because the queries and files in this model are presented as sets of index terms. In vector model, documents are represented in a T-dimensional space. The probabilistic type of model, queries, and files are presented on the basis of probability theory [9].

In the mid of 20th century, several experiments on the Information retrieval systems were done to gauge its functionalities and relevance. In the 1960s, Cranfield experiments of Information Retrieval system were done. They were done by Cyril W. Cleverdon at the University of Cranfield. His goal was to assess the functionality of the indexing system. The experiment represents a prototypical evaluation model of IR system that is on large scale use presently.

Another experiment is the Fuzzy retrieval. This experiment was based on an extended Boolean Model combined with Fuzzy set theory. Considering the earlier experiments, Mixed and Max (MMM) and the paice model, they don’t provide a way of assessing query weights. This is however undertaken by the P-norms. Fuzzy retrieval experiment was focused on the IR model. On the other hand, Cranfield’s experiment was focused on the indexing system for retrieval of information.

According to Fuzzy, documents retrieved from a form of query A or B should be in the fuzzy set linked with the union of sets A and B. documents retrieved from the query form A and B should be in the fuzzy set associated with the intersection of the two sets. From the Cranfield experiment, we can draw inference that the use of indexing system is far much faster and better than the retrieval and filtering algorithms.

In one of Luhns’s early research, he states that “it is proposed that the frequency of a word occurrence in an article furnishes a useful measurement of word significance. It is further proposed that the relative position within a sentence of words having given values of significance furnish a useful means for determining the significance of the sentences. The significance factor of a sentence is therefore based on a combination of the two measurements.” Luhn’s quote to some level summarizes his contribution to the automatic text analysis. His assumption is based on the idea that frequency data can be used in word extraction and sentences to represent the mother document.

Let f be frequency occurrence of various word types and r be their rank order. The r and f relationship is shown in the hyperbola below. Zipf’s law which states that the product of frequency and the rank order is a constant. Zipf verified his findings in an American news house. Luhn used a null hypothesis to enable him point two cut-offs (upper and lower). Words above the upper cut-off were considered as lower and those below were considered rare.

Their work is but the basic of later IR works. However, Luhn used them to device an automatic method of abstracting. He later developed numerical for significance in sentences based on the ratio of significant to non-significant words.

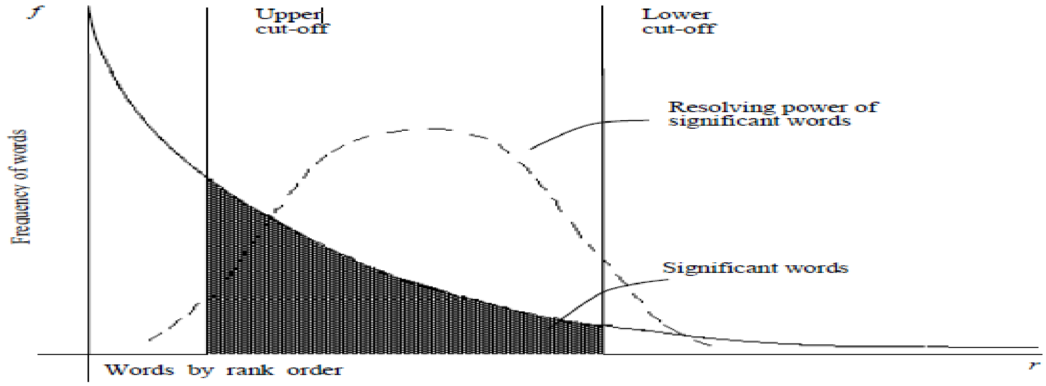


Fig 1. ***Hyperbola***

3. Mechanised Data Recovery Schemes

Mechanised information recovery schemes are applied to reduce what has been called "information overload". Many Tertiary institutions and public libraries apply IR schemes to offer access to books, journals and other files. Web search engines are the most visible IR application [10]

3.1. Connectivity Based

Every object has a relationship with the neighboring object. It uses a maximum distance limit protocol hence; the structure has a hierarchical representation

3.2. Distributed based

This algorithm is related to the pre-defined statistical models. It clusters documents basing on values with the same distribution. The WCM (Web Content Mining) is connected to the IR from the worldwide web in to a more structured form and index the data for efficient and quick retrieval. For us to be able to achieve our goal in developing an intelligent IR, we need to dig a bit deeper on counterpart related art in the whole process. These are; Citation context Analysis, Machine Learning Experiments using conditional probability Models, Semantic Web Initiatives for Scientific Discourse and Citation Context Analysis. There has been a rising interest in the use of citation context for information providing services. It is categorized into citation classification schemes, automated extraction of cited content, and using citation context to retrieve information.

3.2.1. Machine Learning Experiments using conditional probability Models

Present research depicts that the process of content identification is a sequential classified problem which is achieved through using conditional probability models, Conditional Random Fields (CRFs) and Maximum Entropy Markov Models (MEMMs) for retrieving citations and bibliographic information written in documents. This method was more flexible since it did not really require 100% user interactions.

3.2.2. Semantic Web Initiatives for Scientific Discourse

Research persons from this discipline have exhibited some form of interest to model and possess scientific discourse. This field is strategically developing. There has been a proposed Semantically Annotated Latex, whose key role is developing content automatic content identification system.

3.3. Functionality and Appropriateness Procedures

The evaluation of an information recovery system is the process of evaluating how well a system meets the information requirements of its user. Traditional evaluating metrics, developed for Boolean recovery or top-k recovery, include accuracy and recollection. Many more procedures for evaluating the performance of information recovery schemes have also been suggested. In broad sense, measurement looks at the gathering of files to be searched and a search query. All collective procedures mentioned here adopt a ground factual concept of significance. Each file is considered to be whether significant or insignificant to a specific query. In practice, queries could be ill-posed and there might be different shades of significance. Almost all modern assessment metrics for instance, mean average accuracy, discounted cumulative gain are developed for ranking recovery without any clear ranking cut off, taking into consideration the relative order of the document recovered by the search engines and offering a lot of weight to documents returned at advanced ranks [11]

4. Research Questions

i. What are the most relevant and memory-economical algorithms and models that can be used to develop an efficient data searching tool in information retrieval?

ii. Does memory space utilization a key factor in determining technology adoptability?

iii. Which are the key features in determining the functionality quotient of a technology?

iv. What criteria are used on the evaluation of information retrieval tools?

v. What is the best available IR algorithm for efficient data searching?

5. Methods

5.1. The Setting and Participants

This research paper has a technological setting based in the current technological convergence era. The participants are librarians, scientists, SMART, GOOGLE and technologist. Librarians, Scientists and Technologists are favorable participants because they interact with information retrieval systems on daily basis. SMART and GOOGLE on the other hand are the common implementations of information retrieval systems used worldwide.

5.4. Quality Control Procedures

The measures used represent all the facets of Information retrieval construct. Such include; the anchor text holds higher class details of the pointed-to a page. The probabilistic type of model, queries, and documents are presented on the basis of probability theory (Foote, 1999).

6. Types of Information Retrieval Tools

6.1. Ranking Algorithm

The Boolean systems offer powerful online advantages to experienced Intermediaries such as librarians in the process of information retrieval from the system. However, this intern is not that efficient to the end user who do not interact with the system on daily basis [12] [13]. This disadvantage has led to the use of preferred ranking approach. In this approach, a user inputs any natural query without Boolean operations. The query is then processed and a list of ranked information is outputted as a response to the query. This method is more end-user oriented. It is able to produce results even if the query is wrong. The query result is often modified by statistical term-weighting. The table 1 below best explains statistical ranking.

Term	Factors	Information	Help	Human	Operation	Retrieval	Systems

Qry	Human factors in information retrieval systems						
Vtr	**1**	**1**	**0**	**1**	**0**	**1**	**1**

Rec 1.	Human , factors, information, retrieval						
Vtr.	**1**	**1**	**0**	**1**	**0**	**1**	**0**
Rec 2.	Human, factors, help, systems						
Vtr.	**1**	**0**	**1**	**1**	**0**	**0**	**1**
Rec 3.	Factors, operation, systems						
Vtr.	1	0	0	0	1	0	1

Inference

Simple match.
Query (1 1 0 1 0 1 1)
Rec 1 (1 1 0 1 0 1 0)
(1 1 0 1 0 1 0)=4
Query (1 1 0 1 0 1 1)
Rec 2 (1 0 1 1 0 0 1)
(1 0 0 1 0 0 1)= 3
Query (1 1 0 1 0 1 1)
Rec 3(1 0 0 0 1 0 1)
(1 0 0 0 0 0 1)= 2

Weighted match
Query (1 1 0 1 0 1 1)
Rec1 (2 3 0 5 0 3 0)
(2 3 0 5 0 3 0) =13
Query (1 1 0 1 0 1 1)
Rec 2 (2 0 4 5 0 0 1)
(2 0 0 5 0 0 1) =8
Query (1 1 0 1 0 1 1)
Rec 3 (2 0 0 0 2 0 1)
(2 0 0 0 0 0 1) = 3

Table 1. Statistical Ranking

There are two ranking models being observed, they include ranking the query input against individual documents using (vector and probability) and ranking the query against the entire groups of related documents [14] Vector space model uses cosine correlation to calculate similarity in the pool of documents. The probability based model worked in such a manner that documents appearing in the previous retrieval search logs are given a higher weight. Assuming the query terms same probability, the term weighted formulae is derived by

$$similarity_{jk} = \sum_{i=1}^{Q} (C + \log \frac{N - n_i)}{n_i})$$

Adding a within-document frequency, using a turning factor k, the similarity formula changes to the one below

$$similarity_{jk} = \sum_{i=1}^{Q} (C+IDF_i)^* f_{ij}$$

Performance increases when using a term-weighted distribution of the term within a set. It greatly increases when a within-document frequency is added to the IDF weight [15]

Creating an inverted file support structures is wildly beneficial since only the ID of the record is stored thus a smaller index is realized. The inverted file has two segments; the dictionary which stores terms and their statistics and a posting file which houses IDs and weights of the term occurrences. Storing weights in the posting files has four major options;

a. Store raw frequency b. Store normalized frequency c. Store completely weighted term. d. In case no within – record weighting is used, the posting file will not need to store weights.

$$f_{ij} = K + (1 - K) \frac{freq_{ij}}{\max\ freq_j}$$

The inverted file functions effectively when the I/O is minimized. It saves time at the expense of memory space gain. These are some of its demerits when it is used for retrieving humongous volumes of data.

6.1.1. Searching the Inverted File

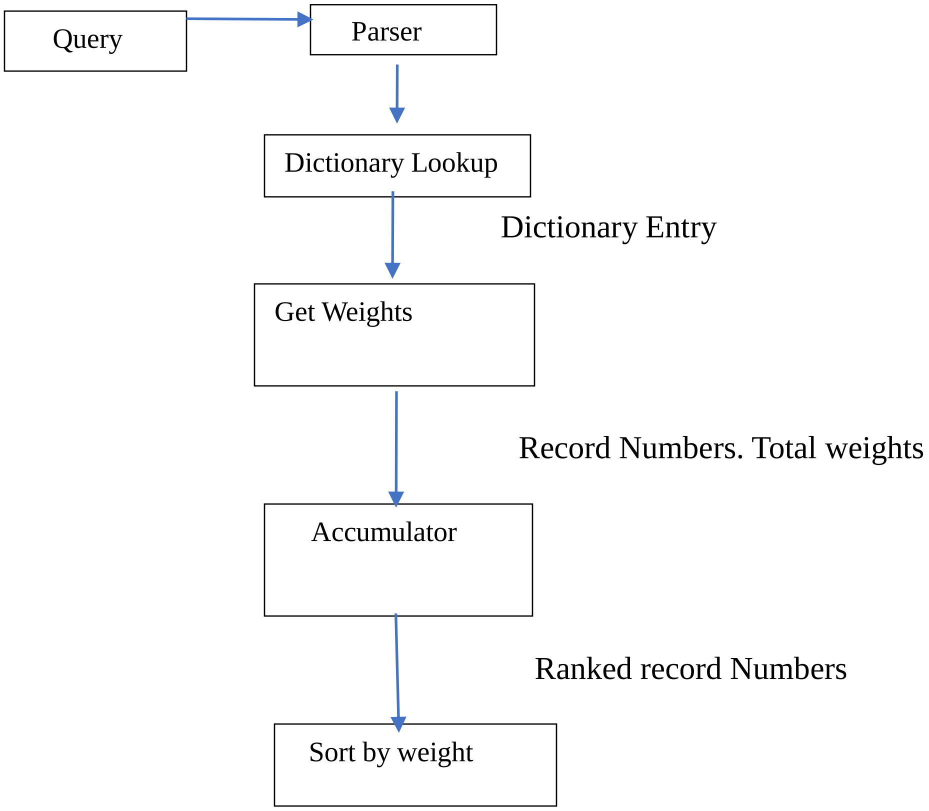


Fig 2. Searching the inverted file

6.2. Star Clustering Algorithm

This algorithm's core drivers is the cluster hypothesis ("closely associated documents relate to the same request").IR systems search as google, Inquiry, and Smart provides an automated computation of ranked list of the document which is sorted by most relevant criteria. Considering such automatic Organization of information, the star clustering algorithm.is presented to be effectively advantageous [16]. For accurate cluster computation, clustering is formalized as covering graphs by cliques. The cover is a vertex cover. Dense star shaped subgraphs are used to compute static data offline and dynamic data online. Using a vector shaped model, a collection of documents is represented by its similarity graph which is undirected, weighted graph $G = (V, E, w)$. The vertices of the graph G represents each document and each weighted edge is a similarity measure between several documents. Measuring similarity is done by use of the standard IR community metric (the cosine metric of vector space model of Smart IR system) [17] [18].

A star shaped subgraph on $m + 1$ vertices has a single star center and m vertices with edges. Through finding cliques in the similarity graph $G\sigma$, a duo similarity between two documents is guaranteed. However, no such similarity is guaranteed between the satellite vertices. Several theorems have been developed in trying to calculate the similarity between documents to be retrieved as shown below.

Consider a $G\sigma$ as a similarity graph and let S1 and S2 be satellite vertices in the same graph. The similarity between these documents must be near the result of this computation. This analysis is based on the eventuality of a worst case scenario.

$\mathrm{Cos}\ (\alpha 1 + \alpha 2) = \cos \alpha 1 \cos \alpha 2 - \sin \alpha 1 \sin \alpha 2.$

The second theorem is in line with the reasoning of the deduced similarity that is expected between two satellite vertices by putting into consideration the geometric constraints in the vector space model.

Consider letting C be the root of the star graph, S1 and S2 are satellite vertices, the similarity between the satellite vertices is computed as shown below.

$\mathrm{Cos}\ \alpha 1 \cos \alpha 2 + \cos \theta \sin \alpha 1 \sin \alpha 2.$

θ is the dihedral angle between the planes formed by S1C and S2C.

This dependent to $\cos \theta$ should be eliminated to increase accuracy and efficiency. Taking in to consideration three vertices from a pool of vertices with similarity σ, chose a random similarity among the vertices it should be cos w for w hence, $\cos w \geq \sigma$. $\mathrm{Cos}\ \omega = \cos \omega \cos \omega + \cos \theta \sin \omega \sin \omega$

From this formulae, we can proceed to;

$$\cos \theta = \frac{\cos w - \cos^2 w}{\mathrm{Sin}^2 w} = \frac{\cos w\,(1 - \cos w)}{1 - \cos^2 w} = \frac{\cos w}{1 + \cos w}$$

Substituting for $\cos \theta$ and keeping in mind that $\cos w \geq \sigma$, we get; $\cos \gamma \geq \cos \alpha 1 \cos \alpha 2 + \frac{\sigma}{1 + \sigma} \sin \alpha 1 \sin \alpha 2$

This equation formulae provides an accurate estimate of the similarity between pair of satellite vertices. After a TREC FBIS experiment was done, the mean squared (RMS) error found was as low as 0.16 in the worst case scenario. This value is negligible. The strength of an algorithm is measured in regards to the expected run time for a result to be outputted. The star algorithm is a regarded as a greedy algorithm. This is attributed to its repeated selection of unmarked vertex with the highest degree making this node and all other adjacent nodes visited or covered. For an even weaker algorithm, it is argued that the number of iterations is approximately $1 + 2\log n / \log (1/(1-p))$. This algorithm is deliberately described as weak since it selects almost any unmarked vertex randomly to the next node.

Taking into consideration the described weak algorithm, after *i* stars have been generated, each of their centers is marked and a few n – *i* remaining vertex. p is the probability for any none center vertex being adjacent to a particular center vertex. From this $(1 - p)^i$ is the probability that any none center vertex remains unmarked. The probability of the vertex being marked is $1 – (1 - p)^i$. The probability that any all n-*i* none center vertices are visited or marked is $(1 –(1-p)^i)^{n – i}$ .This is the probability that *i* stars are enough to cover $G_{n,p.}$ By letting X be a random variable that reflect to *i* stars required to cover $G_{n,p}$, we have;

Pr [X ≥ i +1] = $1 – (1 – (1-p)^i)^{n-i}$ Since for any discrete random variable Z whose range is {1, 2, 3…... n}, $E[Z] = \sum_{i=1}^{n} i * \Pr[Z = i\,] = \sum_{i=1}^{n} \Pr[\,Z \geq i]$ We have $E[x] = \sum_{i=1}^{n-1}[1-(1-(1-p)^i)^{n-1}]$ For any n ≥ 1and x ∈[0, 1],$(1-x)^n \geq$ 1-nx we can derive $E[x] = \sum_{i=0}^{n-1}[1-(1-(1-p)^i)^{n-1}] \leq \sum_{i=0}^{n-1}[1-(1-(1-p)^i)^n] = \sum_{i=0}^{k-1}[1-(1-(1-p)^i)^n] + \sum_{i=k}^{n-1}[1-(1-(1-p)^i)^n] \leq \sum_{i=0}^{k-1} 1 + \sum_{i=0}^{k-1} n(1-p)^i = k + \sum_{i=0}^{k-1} n(1-p)^i$ For any k, selecting k so that $n(1-p)^k = 1/n$ (implying, k=2logn/log(1/1-p))).$E[x] \leq k + \sum_{i=k}^{n-1} n(1-p)^i \leq 2 \log n/\log(1/(1-p)) + \sum_{i=k}^{n-1} 1/n \leq 2\log n/\log(1/(1-p))$ + . The expected time required to traverse this graph is approximated to; $O(np^2 \log^2 n/\log^2(1/(1-p)))$ for $0 \leq p \leq 1-\Theta(1)$.

6.3 Tokenization Algorithm

This is the process of identifying topics present in an input document text. This helps in drastically lowering the search time. This algorithm requires less memory space for storage hence very efficient when it comes to limited space. With the current state of documents enlarging in web pages from their original space, tokenization algorithm comes in handy. Ranking algorithms operate with basic principles regarding the relevance of the document. This algorithm goes to an extra mile of predicting what is more relevant to the user and what is not. This algorithm can be a critical operand in the retrieval model of the system. It simply separates all the characters, words, numbers etc. from the document. This selected character and words are the tokens. During this process of token generation a background process that evaluates token frequency value of all the present tokens in the input document text.

6.3.1. Tokenization Process

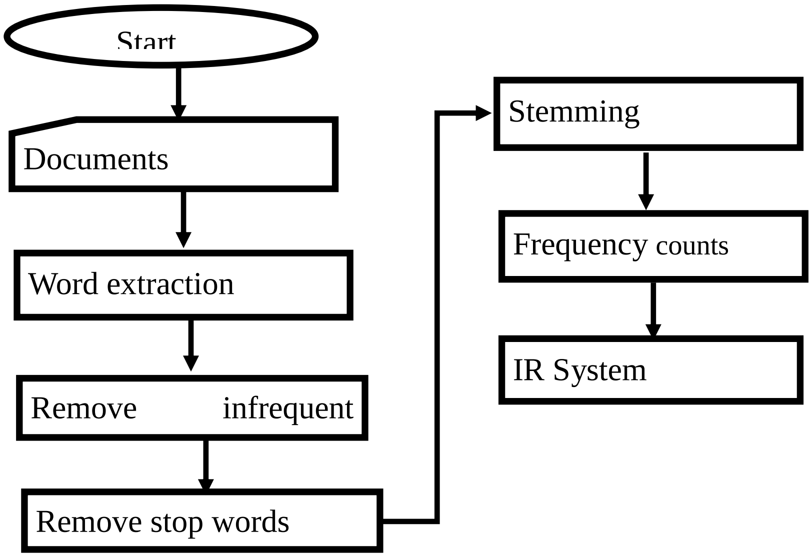


FIG 3. Tokenization Process

Obviously, the main focus of an IR system is to find a relevant document regarding the user’s query. The documents are all gathered via a preprocessor directive then passed to the word extraction face. From the wording, all the words are extracted in this phase. All the infrequent words are removed. The result is passed to the stop words removal phase where “useless words” information retrieval are removed. These words include conjunctions interjections etc. this stage is advantageous since it reduces the indexing file improving the overall general efficiency. The next procedural step is the stemming phase. This phase has the sub-part (the root or stem of a word). The aim of this phase is to remove some suffix and prefix so as to have matching stems in the words in return minimizing memory requirement [19][20]

6.3.2. Stemming Process

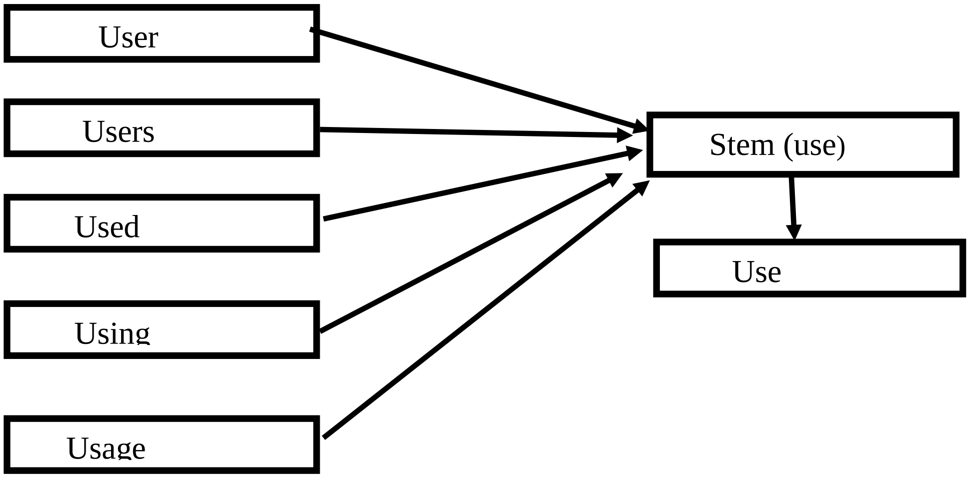


Fig 4. Stemming Process

7. Case Study

The information used were gathered from Kumasi Technical University libraries, SMART, and GOOGLE search for the experimentation which took place on the campus of Kumasi Technical University, Ghana.

8. Time Complexity

In order to ensure effective results, thirty one days period was set for the experiment. This started from 1st September to 31st 1st October 2016. This period was considered because that is the time students were still not active at the Kumasi Technical university campus where the experiment was carried out for verification. Secondly the participants who undertook the experiment had enough time for concentration and thoroughly work.

9. Experiment

The experiment was carried out to ascertain which tool is more efficient for information search. The participants are librarians, scientists, SMART, GOOGLE Technologist. Librarians. Scientists and Technologists are favorable participants because they interact with information retrieval systems on daily basis. SMART and GOOGLE Technologies on the other hand are the common implementers of information retrieval systems used worldwide.

9.1. Experimental Result

Algorithm Tool	Time set	Data volume	Accuracy
Raking Algorithm	30 days	280MB	High
Star Clustering Algorithm	30 days	280MB	Slow
Tokenization Algorithm	30 days	280MB	Very High

Table 1. Experimental Result

9.2. Experimental Results

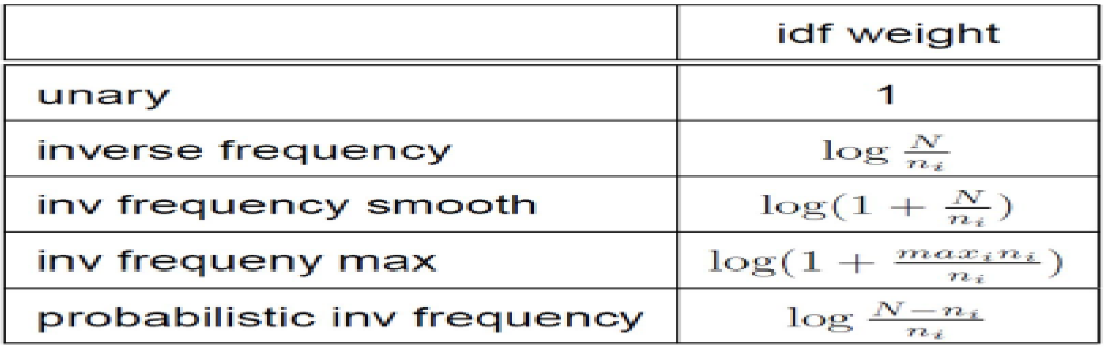
The Table 1 above shows the results of the experiment carried out at the computer Kumasi Technical University, Ghana applying the selected tools. Memory space utilization is a key feature in determining the adoptability of any technology. An Algorithm such as the tokenization algorithm fully agrees with utilization of memory space. Time and volume of data are as well key features in determining the functionality quotient of a technology. The cluster algorithm is best suited for this analogy.

10. Evaluation

The IR system can be evaluated under several criteria. Execution, Storage efficiency, retrieval effectiveness, and what it offers to the user are some of the criteria used for evaluation. The relevance of these factors are determined by the system developer and the equally most suitable algorithm and data structure for implementation are dependent on the decisions made by the developer. The efficiency in its Execution is measured is by the time it takes a module of the system or the system at large to perform a successful computation. It is measured using C supported systems. Execution efficiency is and always will be a major point of concern for IR systems.

Efficiency in terms of storage is measured in bytes required to store data. Conventionally, it is calculated as shown in table 3 below. For inverted files, the IDF depending on different operands is computed as follows

$$\frac{\text{Index file size + document file size}}{\text{Document file size}}$$



	idf weight
unary	1
inverse frequency	$\log \frac{N}{n_i}$
inv frequency smooth	$\log(1 + \frac{N}{n_i})$
inv frequeny max	$\log(1 + \frac{max_i n_i}{n_i})$
probabilistic inv frequency	$\log \frac{N - n_i}{n_i}$

Table 2. Efficiency in terms of storage

Several measures of retrieval effectiveness have been brought to light. The most common techniques used are recall and precision. The ratio of important documents retrieved from given query over the total number of important documents for that query in the database is the recall. The denominator should be estimated by sampling since its unknown. Precision, on the other hand, is the maximum total count of the relevant document retrieved over the total number of documents available in the database. To evaluate this, a graph recall-precision is drawn whereby the x-axis is precision and recall is the y-axis. When the graph is plotted, it is evident that recall and precision are inversely proportional to one another.

As stated above, the standard way of evaluating Information retrieval system is based on none relevance or relevance of the documents. With regards to the user information wants, the pooled document is assigned a binary classification as either relevant or irrelevant. This classification is referred to as ground truth or gold standard relevance judgment. The relevance of a document is not assessed in terms of the query but it is assessed relatively to the information need.

Most of these IR systems have several weights known as parameters that can be tuned to increase system performance. It is not right to report results obtained from the tuning of this parameters in an aim of increasing performance. This is because tuning of these parameters overshadows the expected system performance. These weights are assigned to a specific query rather than to a random query samples. With such an occurrence, the best way is to have more than one development test collections and tuning the parameters on the development test collections. The tester runs the IR system with those parameters and the final reported results from the collection are unbiased.

11. Conclusion

We can conclude that Information retrieval systems on the basis for information gain in the current age via the Internet. IR’s applications are in Google, SMART etc. which are the basic source of any kind of information in this technological convergence era. A study shows that by the end of 2016, over half the world population (3.9 billion) is online with 89 million from developing countries. This article finalizes on the algorithms, structures, and models used in coming up with a successful information Retrieval tool for efficient data retrieval.

The most powerful algorithm should be able to understand the expected results from the user query so as to satisfy him or her accordingly. The algorithm, however, should consider the necessary computer resources such as the Memory when retrieving information. The best algorithm should ensure that memory space needed for it to run efficiently and effectively be cheap and manageable. An algorithm that ensures such success is the tokenization algorithm. As explained in this article, it uses index model that stores document ID rather than the whole document.

It is ideally a huge dilemma when trying to choose the best algorithm for your IR tool. All these algorithms have weaknesses whose effect can be felt by either the user or the resource provider. An algorithm such as the clustering algorithm has its strength in handling huge information at the shortest time possible which is a win for the user. However, it is not all that economical when it comes to memory space usage. Equally the tokenization or the Ranking algorithms have their strongholds. They are accurate in a low degree of memory space usage but, they can be disastrous when used in retrieving information at bulk. They are not as fast as the star cluster algorithm. Choosing the best algorithm and model is a critical decision to be made. The system developers should always keep the user satisfaction with the available resources before anything else.

12. Acknowledgment

The work was supported by the National Natural Science Foundation (NSF) under grant (No.61472294, No.61672397), Key Laboratory of Spatial Data Mining & Information Sharing of Ministry of Education, Fuzhou university (No. 2016LSDMISO5) program for the High-end Talents of Hubei Province Any opinions, findings, and conclusions are those of the authors and do not necessarily reflect the views of the above agencies.

References

[1] Beebe, N. H. A Complete Bibliography of ACM Transactions on Information Systems (2017).

[2] Beel, J, Gipp. B, Stiller. J, “Information Retrieval On Mind Maps - What Could It Be Good For? Proceedings of the 5th International Conference on Collaborative Computing: Networking, Applications and Work sharing (CollaborateCom'09). Washington, DC: IEEE. (2009).

[3]Broderse`n,K.H.,Ong,C.S., Stephan,K.E., Buhmann J.M “The binormal assumption on precision-recall curves”. Proceedings of the 20th International Conference on Pattern Recognition, (2010). 4263-4266.

[4] Center for Intelligent Information Retrieval | UMass Amherst". (2016) ciir.cs.umass.edu. Retrieved -07-29.

[5] Christopher D. M., Prabhakar, R., Hinrich, S., "Chapter 8: Evaluation in information retrieval" (PDF) (2009). Retrieved 2015-06-14. Part of Introduction to Information Retrieval

[6] Della Rocca, P, Senatore, S, & Loia, V. A semantic-grained perspective of latent knowledge modeling. Information (2017) Fusion (2015) n, *36*, 52-67. https://doi.org/10.1016/j.inffus.2016.11.003

[7] Doslu, M., & Bingol, H. O. Context sensitive article ranking with citation context analysis. *Scientometrics*, *108*(2) (2016) 653-671. https://link.springer.com/article/10.1007/s11192-016-1982-6.

[8] Frakes. A, William. B, ‘Information Retrieval Data Structures & Algorithms’ Prentice-Hall, Inc. (1992) ISBN 0-13-463837-9

[9] Foote, J "An overview of audio information retrieval". Multimedia Systems. Springer. (1999)

[10] Goodrum, A. A. (2000) "Image Information Retrieval: An Overview of Current Research". Informing Science. **3** (2).

[11] Jansen, B. J. and Rieh, S. “The Seventeen Theoretical Constructs of Information Searching and Information Retrieval”. Journal of the American Society for Information Sciences and Technology. (2010) 61(8), 1517-1534.

[12] Lavrenko, V. “Introduction to Probabilistic Models for Information Retrieval”. (2010) http://homepages.inf.ed.ac.uk/vlavrenk/doc/ pmir-1x2.pdf.[13] Manning, C. D., Raghavan, P., Schütze, H. “Introduction to Information Retrieval. Cambridge University Press”. (2008)

[14] McSweeney, D, A Data Driven Guide To Anchor Text (And Its Impact On SEO) (2016). https://ahrefs.com/blog/anchor-text/.

[15] Mark S., Bruce, W.C., "The History of Information Retrieval Research". Proceedings of the IEEE. **100**: (2012) 1444–1451. doi:10.1109/jproc.2012.2189916.

[16] Powers, D. M., W "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation" (PDF). Journal of Machine Learning Technologies. (2011) **2** (1): 37–63.

[17] Sreedhar, G & Chari, A. A. First Look on Web Mining Techniques to Improve Business Intelligence of E-Commerce Applications. In Handbook of Research on Advanced Data Mining Techniques and Applications for Business Intelligence. (2017) (pp. 298-314). IGI Global.

[18] Ting, K. M. “Encyclopedia of machine learning” (2011). Springer. ISBN 978-0-387-30164-8.

[19] "University of Glasgow, School of Computing Science, Research overview - Information Retrieval" (2016). www.gla.ac.uk. Retrieved 2016-07-29

[20] Wang, L., Zhang, Y., Qian, D., & Yao, M. “Extraction of Information from Public Health Emergency Web Documents (2016)