INTRODUCTION

The capacity of storage device is increase and cost is decrease there is tremendous growth in database of all sorts. This explosive growth has led to huge, fragmented and become easy to collect and store information in document collection, it has become increasingly difficult to retrieve relevant information from this large document collection. Information Retrieval is a system which gets the information from the web related to a specific keyword in a order according to their rank. Information Retrieval System, that is a system used to store items of information that need to be processed, search and retrieved corresponding to the user’s query. The general objective of Information Retrieval System is to minimize the overhead can be express at the time a user spend in all of the steps leading to reading an item containing the needed information. The systems first extract keywords from documents and then assign weights to the keywords, by using different approaches. (Such a system has two major problems. One is how to extract keywords precisely and the other is how to decide the weight of each keyword.) The goal of an Information Retrieval System (IRS) is to help a user to locate the most similar documents that have the potential to satisfy the user information needs. The focus of information retrieval is the ability to search for information relevant to a user’s needs within a collection of data which is relevant to the users query. An Information Retrieval System Framework: Three main components of an information retrieval system is shown in fig-1. It is composed of Documentary Database, Query Subsystem and Matching Mechanism. Documentary database stores the documents and their representations. This component also contains an indexer module which automatically generates a representation for each document by extracting the document contents. Query Subsystem does query formulation. This component allows the user to formulate the queries. It contains a query language that collects the rules to select the relevant document. Matching Mechanism compares the set of documents in the document database with the query which is given by the user. The documents which match with the query given are termed as relevant documents. So this component helps to retrieve the relevant documents.

Several information retrieval models have been studied and developed in the IR area, some model is: - a. Boolean Model: - In a Boolean retrieval model, the indexer module performs a binary indexing in the sense that a term in a document representation is either significant (appears at least once in it) or not. Logical operator like AND, OR, and NOT is used. A document is represented as a set of keywords. i) Queries are Boolean expressions of keywords, connected by AND, OR, and NOT, including the use of brackets to indicate scope. ii) Output: Document is relevant or not. No partial matches or ranking. iii) Popular retrieval model because, iv) Easy to understand for simple queries, v) Clean formalism, vi) Boolean models can be extended to include ranking, vii) Reasonably efficient implementations possible for normal queries, viii) Very rigid: AND means all; OR means any, ix) Difficult to express complex user requests, x) Difficult to control the number of documents retrieved, xi) All matched documents will be returned. If a document is identified by the user as relevant or irrelevant, how should the query be modified? Vector Space Model:- VSM is a commonly used technique in the information retrieval and text excavation. In VSM, text’s each characteristic was considered independent characteristic which form a characteristic space. Cosine similarity is a commonly used measure in VSM. Based on VSM, the strategy uses topic keywords which extract from the user submitting query expression to evaluate topic relevancy. Probabilistic Model: - This model tries to use the probability theory to build the search function and its operation mode. The information used to compose the search function is obtained from the distribution of the index terms throughout
the collection of documents or a subset of it. This information is used to set the values of some parameters of the search function, which is composed of a set of weights associated to the index terms [1, 2, 3, 4, 8, 9, 11, 18, 19].

**BASIC INFORMATION RETRIEVAL SYSTEM**

A document based IR system typically consists of three main subsystems: document representation, representation of users’ requirements (queries), and the algorithms used to match user requirements (queries) with document representations. The basic architecture is as shown in figure.

![Figure 2: basic architecture of IR System](image)

The primary concern in representation is how to select proper index terms. Query formatting depends on the underlying model of retrieval used (Boolean models, vector space models, probabilistic models, fuzzy retrieval models [Borgodna & Pasi, 1993], models based on artificial intelligence techniques). A matching algorithm matches a user’s requests (in terms of queries) with the document representations and retrieves documents that are most likely to be relevant to the user. Typically the matching algorithms calculate a matching number for each document and retrieve the documents in the decreasing order of this number. Various system performance criteria like precision and recall have been used to gauge the effectiveness of the system in meeting users’ information requirements. Precision is the ratio of the number of relevant retrieved documents to the total number of relevant documents available in the document collection. Recall is defined as the ratio of the number of relevant retrieved documents to the total number of retrieved documents. Relevance feedback is typically used by the system (dotted arrows in figure 2) to improve document descriptions, or queries with the expectation that the overall performance of the system will improve after such a feedback.

IR Challenges:-

i) User and context sensitive retrieval
ii) Multi-lingual and multi-media issues
iii) Better target tasks
iv) Improved objective evaluations
v) Substantially more labeled data
vi) vi. Greater variety of data sources
vii) vii. Improved formal models

There are two types of challenges: i) Near term challenges, 2) Long term challenges.

**THE PROBLEM OF WEB-BASED IR SYSTEM**

The characteristics of the information content of the Web and its dynamics, determine a different kind of IR problem with different requirements for IR systems, different tasks to be carried out, and different approaches to perform those tasks. In this section, we briefly visit these issues in order to describe the Web-based IR problem. The information base of the Web shows the following general properties. Large size: The information base on the Web is huge with billions of Web pages.

- Unlabeled: The content and the keywords summarizing the content are not sufficient to categorize the information base.
- Dynamic: The information base changes in time and it changes fast.
- Duplicated: There may be more than one copy of the same information distributed on the Web.
- Interconnected: The content is interconnected through hyperlinks.

Precise delivery of content is subject to user preferences and interpretation.

- Insufficient query: The user queries are limited to a few keywords and the users often do not know the best query to retrieve the information they need.
- Heterogeneous users: There are various users, and their perspectives and interpretation, even for exactly the same content, may be different.
- Dynamic requirements: The users’ needs for information change in time. Even a single user may have different needs on different occasions [12].

Keyword optimization:- It would not be a very wise decision to rush into search engine marketing without having a set of relevant keywords ready beforehand. Search engines are all about keywords. The web surfer type in their search in the form of keywords and the search engines show results analyzing the relevance of those keywords in the web world. The systems first extract key-words from documents and then assign weights to the keywords by using different approaches. Such a system has two major problems. One is how to extract keywords precisely and the other is how to decide the weight of each keyword [2, 5].

Focussed crawler- A focussed crawler or topical crawler is a web crawler that attempts to download only contents that are relevant to a pre-defined topic or set of topics and avoid downloading all others. Topical crawling generally assumes that only the topic is given, while focussed crawling also assumes that some labeled examples of relevant and not relevant pages are available. Therefore a focussed crawler may predict the probability that a link to a particular page is relevant before actually downloading the page. The performance of a focused crawler depends mostly on the richness of links in specific topic being searched, and focused crawling usually relies on a general web search engine for providing starting points. Issues of Consider: According to the characters of focused crawler, there are four issues as the following.

A. Where to start crawling?
B. Which link do you crawl next?
C. What pages should crawler download?
D. How to minimize the load on visited pages?
According to the four issues proposed above, we can conclude that the Key topic of focused crawler is Similarity Computation of Web Pages. Starting with an initial set of keywords, our system expand the set by adding the most suitable term that intelligently selected during the crawling process by genetics algorithm. The definitive guide to gathering, sorting and organizing your keyword into a high performance SEO. As keyword based relevance ranking does not guarantee relevance in meanings, different semantic models have been introduced to improve relevance ranking. To improve the efficiency of retrieval, it has been proposed that the document which are generally retrieved together in response to some query, should be kept close together within the system in the form of clusters. A cluster based search proceeds to satisfy a query efficiently by identifying and retrieving only those clusters which exhibit a sufficiently high degree of match with the query improve the effectiveness of retrieval as it results in the retrieval of a higher number of relevant documents for a given amount of effort [6, 7, 8, 13, 14, 15, 16, 17].

CONCLUSION AND FUTURE WORK

The main advantage of keyword optimization is effective information retrieve using genetics algorithm and similarity function. Initially download a set of document from seed URLs. Based on weight scheme marks the entire document in the above set and select n document at highest marks. Take m words from each document basis of maximum word frequency in a document. Then combine all words of n document and become a single keyword. And convert this word into Boolean model 0 and 1 form. Use the genetics algorithm and use jaccard similarity function to find the fitness value of each document. Jaccard coefficient:

\[
\text{sim}(x,y) = \frac{|x \cap y|}{|x \cup y|}
\]

After find the fitness value apply genetics operator like selection, mutation, crossover and find the optimize result.

REFERENCES
