European Journal of Molecular & Clinical Medicine
ISSN 2515-8260 Volume 07, Issue 02, 2020

INDEXING ON IR – RESULTS AND DISCUSSIONS

Jennifer .P[1], Dr. A. Muthukumaravel [2]

[1] Research Scholar & Assistant Professor, Department of CS, Faculty of Arts & Sci., BIHER, Chennai
jennifer.mca@bharathuniv.ac.in

[2] Dean-Faculty of Arts & Sci., BIHER, Chennai
muthukumaravel.mca@bharathuniv.ac.in

Abstract:

Indexing is an important process in Information Retrieval (IR) systems. It forms the core functionality of the IR process since it is the first step in IR and assists in efficient information retrieval. Indexing reduces the documents to the informative terms contained in them. It provides a mapping from the terms to the respective documents containing them. Once effective index has been built for the collection of documents, retrieval process is simplified. Indexing proceeds at four stages namely content specification, tokenization of documents, processing of document terms, and index building. The index can be stored in the form of different data structures namely direct index, document index, lexicon and inverted index. Index can be built by applying different algorithms or schemes such as single pass in memory indexing, blocked indexing, etc.

Keywords: Indexing, stemming, stopwords, stem word, Information recovery, parsing, duplication, etc.

Introduction:

Information Recovery (IR) is the process of extracting information system resources from a list of those resources that are important to a need for information. Searches can be based on full text or other content based indexing. Recovery of information is the science of searching for information in a document, searching for documents themselves, and also searching for the metadata which describes data, and for text, image or sound databases. Automated information retrieval systems are used to reduce what is called duplication of information. An IR system is a software system that provides access to books, journals, and other records; these records are stored, and maintained.

IMPLEMENTATION:

Implementation is the carrying out, execution, or practice of a plan, a method, or any design, idea, model, specification, standard or policy for doing something. As such, implementation is the action that must follow any preliminary thinking for something to actually happen.

Process of Implementation

The project is an endeavour to extract the stemmed word from the loaded document. Generally, all text documents are stored in a specific folder. First step of the process is to reload the page. Once the page is reloaded, the documents which have been saved in a particular folder will be extracted to the page. Second step is to select and upload a specific document in the system. Once the text document is selected the process begins. Stop words are removed and then duplicate words will also be removed. After the removal of redundant words, stemming is applied.

Type the specific word in the search tab and click on search button. The search engine will search for the word from the uploaded text document. The searched word will be extracted and converted to Stem (root) word. Then the extracted stem word will move to index. For example, consider the word ‘goodness’, if the word ‘goodness’ is supposed to be searched, the search engine will search for the word ‘goodness’ from the uploaded document. Then the searched word is converted into ‘good’ which is the stem word of the word ‘goodness’.

The project has the feature of arranging the words in ascending order using the index-based information retrieval system process. If searching word exists in the document the processor will report as ‘String has been found’. If not, then the processor will report as ‘String has not been found’. If the search word has been found then the processor will try to convert the word into stem word using stemming algorithm.

Advantages:

The advantage of the project is that

- The process runs at a speed of milliseconds.
- Cost is less as the project runs under offline mode.

RESULT AND DISCUSSION

This Session describes about the result and discu ssion with reference to aim of the study. The steps for index based information retrieval system are discussed below.

- **Reload the page:**

The first step is to reload the page. The JavaScript refresh page function can reload the current resource. In most cases, a page is selected to be refreshed. The method also has other perks, such as helping you get the URL address of the current page, redirect the browser to another page, and, of course, refresh page JavaScript.

- Click on the **Reload page** button.

Figure 1: Outline of system view to reload the page

- **Select the document:**

After reloading the page, click the text box it automatically lists out the documents that have been present in the folder. Then select any one of the document to retrieve the information. Once the document has been selected, the data will be extracted.

Figure 2: System view to select the document to retrieve the information

- **Remove the Stop words:**

Stop words are basically a set of commonly used words in any language, not just English.

Examples of minimal stop word lists that can be used:

- **Determiners** – Determiners tend to mark nouns where a determiner usually will be followed by a noun
  Examples: the, a, an, another
- **Coordinating conjunctions** – Coordinating conjunctions connect words, phrases, and clauses
  Examples: for, an, nor, but, or, yet, so
- **Prepositions** – Prepositions express temporal or spatial relations
  Examples: in, under, towards, before.

- **Remove the Redundant words:**

This part works to find and remove the duplicate words in selected document. The processor will check for redundant words in the selected document and remove it.

- **Convert the word to stem word:**

In linguistics, a stem is a part of a word used with slightly different meanings and would depends on the morphology of the language in question. In Athabaskan linguistics, for example, a verb stem is a root that cannot appear on its own, and that carries the tone of the word.

Figure 3: System for converting a word into stem word

For example, **Act**: to move or do (actor, acting, react, enact)

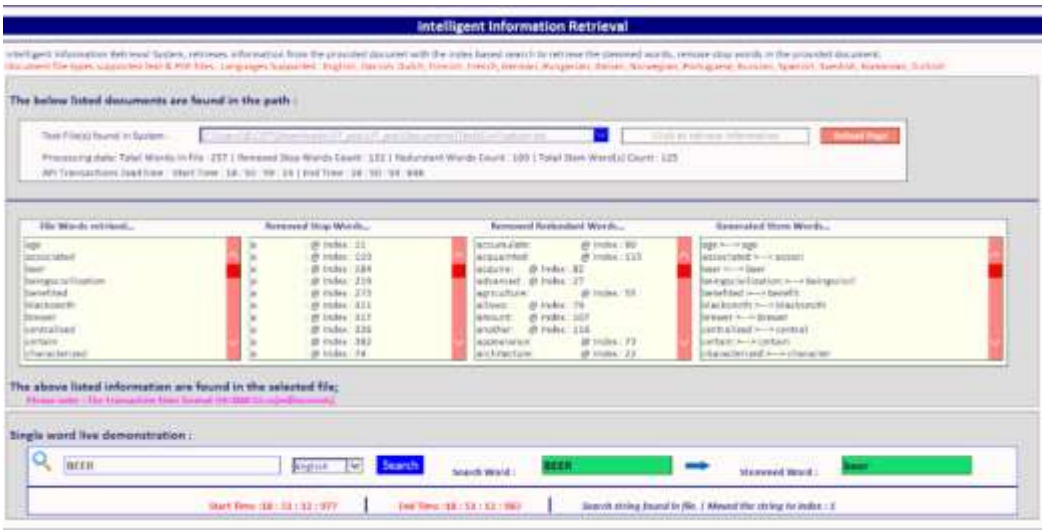
- Search the particular word in the index.

Type any word in a single word live demonstration and select the word in which language to search for the word in the particular document.

Figure 4: System view to search a particular word in the index

• If a word given in a single word live demonstration is found in a document, it searches for a word and finds a root for that word.

Figure 5: System view to search a particular word and find the root word



• If a word given in a single word live demonstration is not found in a document, it searches for a word and says that the search string is not found in a file. These are the process of Index based information retrieval system.

Figure 6: System view for search string is not found

CONCLUSION

Information retrieval manages the capacity and portrayal of learning and the retrieval of information pertinent to a particular client issue. Information retrieval systems react to questions which are normally made out of a couple of words taken from a characteristic dialect. The question is contrasted with record portrayals which were removed amid the indexing stage. This article discussed about Information Retrieval covered with application areas and domain application, Evaluation of information retrieval includes measures and experimental, Ontology covered domain and sub-ontology, Index methods includes technique of indexing and searching and OCR Technique, Stop word removal

and finally discuss about query processing. The research work present about the implementation of searching stem words in a particular word document. In this method, the process is fast, i.e. the transaction time is in milliseconds.

RECOMMENDATIONS

The fact that the article has achieved the objectives of this research work, the maintenance and enhancement of this research work, is still needed. The article is merely an exercise in information building; more functionality can still be applied to the proposed program. As a future enhancement, Synonyms for the extracted stem word can be given. Synonyms of extracted stem word will be useful in analyzing and correcting the sentence meaning in a particular document.

REFERENCES

1. M. Mitra, B.B. Chaudhuri, "Information Retrieval from Documents: A Survey", Information Retreival 2, 141-163, 2000.
2. Ricardo Baeza-Yates and Berthier Ribeiro-Neto. Modern Information Retrieval. ACM Press, 1999.
3. D. D. Lewis and K. S. Jones, "Natural language processing for information retreival", Communications of the ACM, 39(1):92–101, 1996.
4. T. Brants and Google Inc, "Natural language processing in information retrieval," in Proceedings of the 14th Meeting of Computational Linguistics in the Netherlands, pp. 1–13, 2004.
5. H. Joho and M. Sanderson. Document frequency and term specificity. In the Recherche d'Information Assiste par Ordinateur Conference (RIAO), 2007.
6. F Song and W. B. Croft, "A general language model for information retrieval" In Proceedings of the eighth international conference on Information and knowledge management, CIKM '99, pages 316–321, 1999.
7. Tasmin Maxwell, PhD thesis: Term Selection in Information Retrieval, University of Edinburg, 2014.
8. B. Nanus, 'The Use of Electronic Computers for Information Retrieval', Bull Med Libr Assoc, vol. 48, no.3, pp. 278 - 291, Jul. 1960.

9. G Salton and M McGill, Introduction to modern information retrieval. McGraw-Hill, 1983.

10. C Zhai, Statistical Language Models for Information Retrieval. Morgan and Claypool Publishers, 2008