International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 09 Issue: 04 | Apr 2022 www.irjet.net p-ISSN: 2395-0072

CANDIDATE SET KEY DOCUMENT RETRIEVAL SYSTEM

Kunal Gawade[1], Akash Parthe[2], Sameer Deshwal[3], Nirjhar Jaiswal[4], Prof. Sagar Kulkarni[5]

[1,2,3,4]UG Student, Dept. of Computer Engineering, Pillai College of Engineering, New Panvel, India
[5]Assistant Professor, Dept. of Computer Engineering, Pillai College of Engineering, New Panvel, India

---

Abstract **-** *Our main aim is to build a document retrieval system for this project. Suppose we have a lot of documents and the user has to retrieve specific information from that bunch of records, then they have to go through all the documents to retrieve that information which will take a lot of time. To solve this problem, we need a system to help users recover information quickly. This is where our project can be used. Following will be the main features of our project: Processing English queries of users, Interacting with users to correct the incorrect syntax queries, and Giving results of the queries. Our project will be limited to queries in the English language. One of the striking points of this system model is introducing a semantic relationship between query and corpus documents. We can consider the System as an application of a candidate set document retrieval system. The System can be implemented using a heuristic retrieval method. The input query will be processed using NLP techniques*

Keywords : Document retrieval, queries, semantic, NLP

1. Introduction

Recently there has been a growing interest in developing natural interaction between humans and computers. Information retrieval system is one of the ways in which interaction between humans and computers is achieved. Information is the knowledge that has been communicated or received about a specific event or circumstance.. Searching through stored information to retrieve information relevant to the task at hand is referred to as retrieval. As a result, information retrieval (IR) is concerned with representing, storing, organizing, and retrieving data. Here, types of information items include documents that are stored in a directory. The chief goals of the IR are indexing text and searching for useful documents in a collection. A good information retrieval system would rate most of the relevant documents ahead of less relevant documents in response to a user query, thereby allowing the user to use relevant documents.

2. Literature Survey

A. **Evaluation of Information Retrieval Performance Metrics using Real Estate Ontology - Namrata Rastogi , Parul Verma, Pankaj Kumar (2020)[1]:**

The paper focuses on the analysis of various information retrieval performance evaluation metrics for the real estate information retrieval model as proposed by us. The analysis covers all major IR metrics being used by researchers and will help in providing an insight into the set retrieval and rank retrieval metrics. The set retrieval metrics focus on basic precision-recall that uses an unordered result set of web documents.

B. **Cross-lingual Information Retrieval: application and Challenges for Indian Languages - Jay Patel , Kamlesh Makvana , Dr.Parth Shah (2019)[2]:**

In this study, we came to know that the Information Retrieval in the native language is more difficult due to the difference in the rule of sentence formation. But people's thinking and writing of the sentence varies in a broader sense. The relevant information can be dug out by accurate transformation of query words with the incorporation of semantic context. This approach can bridge the gap of words or the language barrier that a naïve user feels.

C. **Correction of Spaces in Persian Sentences for Tokenization - Mahnaz Panahandeh , Shirin Ghanbari (2019)[3]:**

In this paper, a method for correcting the problems in not inserting full-space and half-space in texts typed by users is proposed. In the Persian language, the only preprocessing tool in which the correction of not inserting full-space among words is Step1. In comparing the performance of the proposed method for space correction with STeP1, as well as the Hazm tokenizer, which does not correct full-space mistakes, the results show the superiority of the proposed method.

***D.* Proposed Language Independent Stemmer for Information Retrieval Systems Using Dynamic Programming - Mrs.M.Kasthuri, Dr.S.Britto Ramesh Kumar, Dr. Souheil Khaddaj (2017) [4]:**

In this paper, they have discussed the Proposed Language-Independent Stemmer is useful to find out stem words for four morphologically different languages such as English, French, Tamil, and Hindi. Various research projects to implement Stemmer and Lematizer for multiple languages have recently been completed. The framework and algorithm proposed to support multi-linguistic Information Retrieval is present in the paper.

3. Proposed Work

Document Retrieval process where potentially relevant documents are identified. The identification process is often conducted as a set intersection – from the set of all documents, the potentially relevant documents are those that contain all or some of the search items.

3.1 System Architecture

The system architecture is given in Figure 1. Each block is described in this Section.

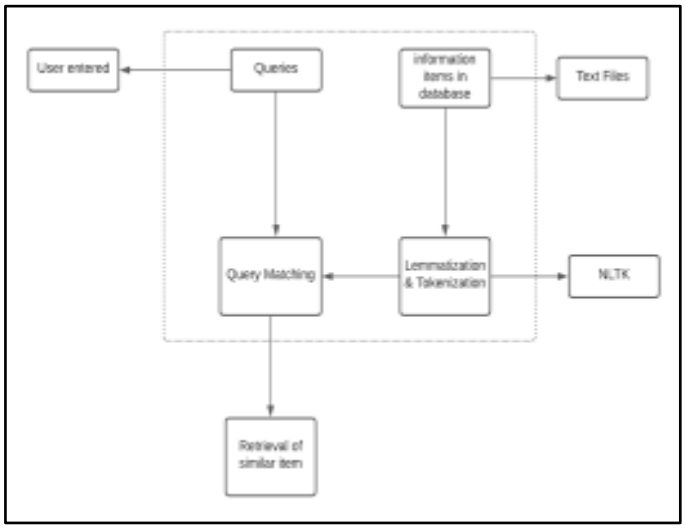

Fig. 1 System architecture

The various components of Information Retrieval System are as follows:

3.1.1 Indexing :

A pre-process called indexing is commonly used in document retrieval systems to effectively determine whether documents from a corpus fit a particular query. It refers to how papers are stored and handled in the collection. A retrieval system saves documents in an abstract representation to make searching more efficient. A list of keywords is kept, as well as links to the documents in which they appear. An inverted file is a structure for storing indexing information. Although there are other possibilities, the inverted file is the most common data format used by IR systems (IF). An IF is a traversed, posting-list-organized version of the original document collection. Each entry in the inverted file refers to a single term in the dictionary. The indexing process includes several steps, which are described as follows:

3.2.1 Tokenization :

The primary organizing of the ordering handle is regularly known as tokenization. In this stage, record content is parsed, and file words called Tokens are produced. In expansion, all characters contained within the tokens are frequently lower-cased, and all accentuations are expelled at this stage. Each dialect has a diverse inner double encoding for the characters within the dialect. We expect all the reports (English) to be encoded in Unicode based on UTF-8, utilizing different 8-bit bytes.There are a variety of tokenization strategies that can be used depending on the language and modeling aim.

3.2.3 Stop Words Removal :

Luhn mentioned that the frequency of a time period within a document might be a very good discriminator of its significance inside the file. Similarly, there are numerous extremely common terms (e.g., “the”) that appear in almost all files of a corpus. These terms are called prevent phrases, which convey little value for the cause of representing the content material of files and are generally filtered out from the list of ability indexing phrases for the duration of the indexing system. Getting rid of the prevent words also lets in the reduction of the scale of the generated report index. But, removing stopwords from one report at a time is time eating. A fee-effective technique is composed of doing away with all phrases which usually seem inside the report series and for you to no longer improve retrieval of applicable files. Those stopwords have one-of-a-kind impacts on the information retrieval process. Relational stopwords indicate semantic relevance that is important for green records retrieval. Doing away with relational stopwords from the file would bring about a lack of such applicable semantic facts resulting in a decrease in the relevant performance of the system. At the same time, casting off non-relational stopwords would lessen the record duration resulting in quicker seek. We remove the best non-relational stopwords to perform relation inclusive looking. White space tokenization technique is the most commonly used tokenization technique .

3.2.4 Stemming :

Stemming is the process of converting an inflection (or derivative) word into a stem, stem, or root form (usually the form of the written word) in inflection and information retrieval. The stem does not have to be the same as the morphological root of the word. It is usually sufficient to assign the same stem to related words, even if the stem itself is not a valid root. Computer scientists have been using stemming algorithms since the 1960s. As a form of query extension, many search engines treat words that have the same root as synonyms. This is a technique called conflation.

3.2.8 Lemmatization :

Lemmatization is a term that describes how to perform things correctly by using a vocabulary and morphological examination of words. Typically, the goal is to remove just inflectional endings from a word and return it to its base or dictionary form.

In stemming, a portion of the word is fairly chopped off at the tail conclusion to reach the stem of the word. There are certainly distinctive calculations utilized to discover how numerous characters got to be chopped off, but the calculations don't really know the meaning of the word within the language it has a place. In lemmatization, on the other hand, the calculations have this information. In truth, you'll indeed say that these calculations allude to a lexicon to get the meaning of the word, sometime recently decreasing it to its root word, or lemma. So, a lemmatization calculation would know that the word way better is derived from the word great, and thus, the lemme is sweet. But a stemming calculation wouldn't be able to do the same. There can be over-stemming or under-stemming, and the word way better may well be diminished to either wagered, bet, or fair held as way better. But there's no way in stemming that it can be diminished to its root word great. This, essentially, is the contrast between stemming and lemmatization.

3.2.5 Doc2Vec :

Doc2vec converts a document to a vector using an unsupervised machine learning approach. The main aim of Doc2vec is to represent documents numerically.Doc2vec is similar to word2vec, but unlike words, it does not maintain a logical structure. So while developing doc2vec, another vector named Paragraph ID is added to it.

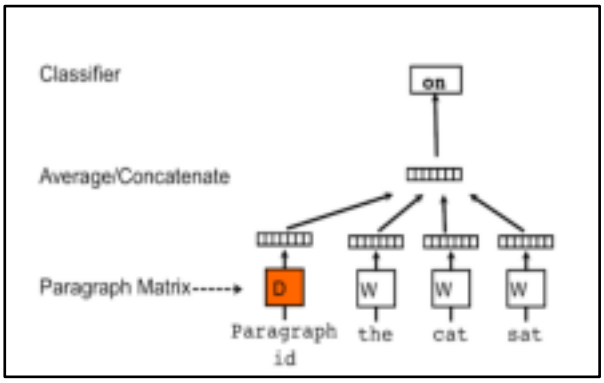

Fig 3.2.5 : Distributed Memory version of Paragraph Vector (PV-DM)

In the above figure, there is a feature vector added through which the uniqueness of the document can be identified. While training such a model, the vectors named 'W' are the word vectors that hold the numeric representation and represent the concept of a word. Similarly, the document vector, designated as 'D,' has the numeric representation and conveys the concept of a document.

3.2.6 Query Matching :

Query matching is done using similarity matching. In similarity matching, documents that are similar to user queries are returned. The typical method to compute text similarity between documents is to convert the input documents into real-valued vectors. The purpose is to create a vector space in which similar papers are "near" based on a predetermined similarity measure.

"Cosine similarity" is the approach used to calculate similarity.Cosine similarity can be calculated using the below formula:-

$$cos(a,b) = \frac{\sum_{i=1}^{n} a_i b_i}{\sqrt{\sum_{i=1}^{n} a_i^2}\sqrt{\sum_{i=1}^{n} b_i^2}}$$

The document having highest cosine similarity will be the most similar document with the user query.

3.2.9 Dataset :

Dataset is defined as the data on which processing and information retrieval needs to take place. In our case documents are the dataset. The documents included in the dataset are of various domains. These domains include cloud computing, web development, Human machine interaction etc. Choosing a proper dataset is also an important task in information retrieval as the documents included should be related to the user queries.

3.2.8 Retrieval of similar items:

After the query matching process, system returns the documents which are related to the user queries. System returns either a list of documents related to the user query or it can also return no document if the user query is not found.

4. CONCLUSION:

A system where document retrieval based on user query has been done. The preprocessing tasks like tokenization, stop word removal and lemmatization have been implemented on documents and user queries.Doc2vec has been used to retrieve documents and the results have been accurate.

REFERENCES:

[1] Evaluation of Information Retrieval Performance Metrics using Real Estate Ontology - Namrata Rastogi , Parul Verma, Pankaj Kumar (2020)

[2] J. Patel, K. Makvana and P. Shah, "Cross-lingual Information Retrieval: application and Challenges for Indian Languages," 2019 IEEE 5th International Conference for Convergence in Technology (I2CT), 2019, pp. 1-4, doi: 10.1109/I2CT45611.2019.9033563.

[3] Correction of Spaces in Persian Sentences for Tokenization - Mahnaz Panahandeh , Shirin Ghanbari (2019)

[4] Proposed Language Independent Stemmer for Information Retrieval Systems Using Dynamic Programming - Mrs.M.Kasthuri, Dr.S.Britto Ramesh Kumar, Dr. Souheil Khaddaj (2017)

[5] W. Zhang, W. Wang, L. Zhu, R. Zheng and X. Liu, "Python-Based Unstructured Data Retrieval System," 2020 International Conference on Smart Grid and Electrical Automation (ICSGEA), Xiangtan, China, **2020**, pp.

[6] Dahab, Mohamed & Alnefaie, Sarah & Kamel, Mahmod. (2018). A Tutorial on Information Retrieval Using Query Expansion. 10.1007/978-3-319-67056-0_35.

[7]A. Gadag and B. M. Sagar, "A review on different methods of paraphrasing," 2016 International Conference on Electrical, Electronics, Communication, Computer and Optimization Techniques (ICEECCOT), 2016, pp. 188-191, doi: 10.1109/ICEECCOT.2016.7955212.

BIOGRAPHIES:

Kunal Gawade is an undergraduate student of Mumbai university. His area of interests are NLP and Data science.

Akash Parthe is an undergraduate student of Mumbai University. His area of interests are Web Development and NLP.

Sameer Deshwal is an undergraduate student of Mumbai university. His area of interests is Data Science and NLP.

Nijhar Jaiswal is an undergraduate student of Mumbai university. His area of interests are data mining and NLP.