International Journal of Computer Applications (0975 – 8887)
Volume 126 – No.12, September 2015

Phrase Prioritization Algorithm and Supporting Data Structure for Retrieval

Sachin Kumar
Sr.Asst.Prof – Computer Science/IT
Fairfield Institute of Management & Technology
(Affiliated to GGSIP University, New Delhi)

Pratishtha Gupta, PhD
Associate Professor – Department of Computer Science
Banasthali Vidyapith,Jaipur,India

ABSTRACT
In the heart of this research work lies the proposed algorithm, which prioritizes the phrases of the search queries. This algorithm suggests the methodology of fetching phrases and then searching all possible phrases, so that recall value can be increased. The most important issue in this regard is the usage of such data structure, which facilitates the efficient search of phrases in documents. For this purpose, Linked Representation of Sparse Matrix has been suggested, which consists of linked lists not only rowwise but also columnwise. Columns correspond to the documents and hence make the search of every possible phrase efficient. Rows correspond to the dictionary of words. Linked Representation maintain the dynamic nature of documents as well as insertion and deletion of words from the documents.

Emphasis has also been given to the categorization of dictionary and query words into specific and general words, which will increase the precision of search results. Specific words will be given higher priority as compared to the general words.

Synonyms have also been considered for retrieval of documents, hence increasing the understanding between user requirement and search engine

Keywords
Linked Representation, Search Engine, Precision, Recall, F-measure, Algorithms, Database

1. INTRODUCTION
A novel idea is being presented here which gives highest priority to the phrases. Generally, when any user specifies the requirement, that specification is not limited to one word or words which are isolated or independent of each other. But specification generally comprises of two or more words, which can be considered as phrases. Hence, every possible phrase is retrieved from the user's query and is searched for the relevant documents. Longer the phrase available in a document, the more relevant it should be.

But to search for so many phrases in a document is again an issue, which demands an efficient solution. Hence, the data structure which makes this procedure the most efficient one has been suggested. Since, the relationship of documents with the words of dictionary can be best described in terms of a Sparse Matrix, linked representation of the Sparse Matrix has been suggested to maintain its dynamicity.

Moreover, a new idea of categorization of words of dictionary into specific and general words has also been proposed, which also increases the relevance of the documents to the greatest extent.

2. REVIEW OF LITERATURE
According to a study conducted by, [1] purposes of this research is to propose a new model for the integration of concurrent function deployment (CFD) and parameter (P) diagram in order to prioritize innovation factors. The matrixes of the two phases could become larger if the number of factors in the P diagram increases, and therefore, filling and analyzing the matrixes might become time consuming and difficult.[7] feel planning and scheduling are forms of decision making, which play a crucial role in manufacturing as well as in service industries. In the current competitive environment, effective sequencing and scheduling has become a necessity for survival in the marketplace. A great challenge for today's companies is not only how to adapt to this changing, competitive business environment but also how to draw a competitive advantage from the ways in which they choose to do so. Intelligent solutions, based on expert systems, to solve problems in the field of production planning and scheduling are becoming more and more widespread nowadays.

[4] feel that how can leaders adopt a mindset that maximizes learning, remains responsive to short-term emergent opportunities, and simultaneously strengthens longer-term dynamic capabilities of the organization? This will explores the organizational decisions and practices leaders can initiate to extend, strengthen, or transform "ordinary capabilities" into enhanced improvisational competence and dynamic capabilities. [8] suggest the purposed of this mixed methods study is to define the core components of a system wide, acute care program designed to meet the needs of older adults. The findings yielded eight clusters describing components of a geriatric acute care program: guiding principles, leadership, organizational structures, physical environment, patient- and family-centered approaches, aging-sensitive practices, geriatric staff competence, and interdisciplinary resources and processes. A total of 113 items that describe dimensions of quality were identified with these clusters.

[9] suggest the purpose of this study was to evaluate the effectiveness of teaching or learning in an e-learning system measures in linguistic preferences. The main contributions of this study are twofold. First, the evaluation can be considered as a complex-dependence, hierarchical decision-making problem. This study contains a review of the literature and identifies 21 criteria and five aspects to measure e-learning system effectiveness. Second, this study integrates fuzzy set theory and the ANP to develop an evaluation model that prioritizes the relative weights of the proposed measures. The proposed method can be used to handle dependence within a set of measures and to construct a hierarchical structure. [12] reveals that as they are currently conducted, missions by single ROVs consist of several sub-tasks. After a vehicle has been launched, a human operator or a small team is

responsible for controlling the flight, navigation, status monitoring, flight and mission alteration, problem diagnosis, communication and coordination with other operators, and often data analysis and interpretation. These tasks are similar in terms of their locus of control (e.g., keyboard and mouse input, joystick, trackball, visual display).

[13] feel that it stresses the importance of metadata and recognizes quality as a cyclical process which balances the necessity of national standards, the needs of the user, and the work realities of the metadata staff. This identifies decision points, outlines future action, and explains communication options. [10] suggest the novelty of this study lies in the application of a hybrid approach to a real industry case. This study has dealt with one of the most important subjects in supply chain management, providing a better decision for supplier selection using appropriate quantitative techniques. [11] feel that search engines exist to help sort through all the information available on the Internet, but have thus far failed to shoulder any responsibility for the content which appears on the pages they present in their indexes. Search engines lack any transparency to clarify how results were found, and how they are connected to the search terms. Thus, problems arise in connection with the protection of minors – namely, that minors have access, intentional or unwitting, to content which may be harmful to them. The findings of this study point to the need for a better framework for the protection of children. This framework should include codes of conduct for search engines, more accurate labeling of Web site data, and the outlawing of search engine manipulation. This study is intended as a first step in making the public aware of the problem of protecting children on the Internet.

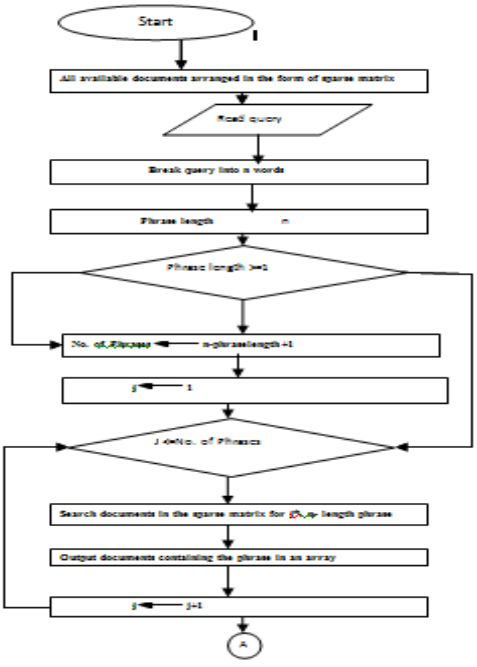
[2] feel that seeks to analyze the systemic effects of AML technologies and regulations, at both national and organizational levels. It calls for a reconsideration of the underlying assumptions within which AML- related technology is appropriated by financial institutions. It demonstrates how this technology creates multiple complex systemic phenomena that often act contrary to initial intentions. This complexity is generated not only by data mining and/or profiling technologies, but also by peripheral technologies as they interact with human activity systems in the AML domain.

[14] feel the competition in market penetration between the traditional and online channels has been intensified. The purpose is to assess predictors of business-to-consumer (B2C) channel preference and investigate the consumers' attitudes towards different shopping channels. This approach, which does not require restrictive assumptions, takes into account the difficulty of giving precise judgments by allowing respondents to be inconsistent to some extent. [3] suggest that about marketing accounting. It is about reading marketing writing and writing marketing reading and what calls them into being. It is about our "abouting" practices; those signifying practices by means of which we week to capture a piece of the world and show it off, wrapped in a suitable tale of discovery, in a cabinet in the museum of marketing knowledge.

3. SPARSE MATRIX REPRESENTATION

Here as the research goes into new phase i.e. the research would introduce the concept of sparse matrix representation. A sparse matrix is which have many of the entries as zero. Here much space and computing time could be saved if only the nonzero terms were retained explicitly. In the case where these nonzero terms did not form any nice pattern such as a triangle or a band, there derived a sequential scheme in which each nonzero element was represented by a structure with three data members row, column and value. The sequential representation permits easy access of matrix terms by row.However, accessing all the terms in a specific column of a matrix is difficult. To provide easy access both by row and by column, there must devise a linked representation for a sparse matrix. In the data representation, the researcher use each nonzero element is in two circular lists, one is a row list and other as a column list. Here it have a circular list for each row and each column of the matrix. Each circular list has a header node. In sparse matrix representation, it uses nodes of the type MatrixNode. This class has a Boolean field head, which is used to distinguish between header nodes and nodes that represent nonzero elements. Each header node has three additional fields : down, right and next. The total number of header nodes is max(number of rows,number of columns). The header nodes for row i is also the header node for column i.The down field of a header node is used to link into a column list , the right field is used to link into a row list. The next field links the header nodes together. All other nodes have six fields : head, row, col, value, down and right. The down field is used to link to the next nonzero term in the same column and the right field links to the next nonzero term in the same row.Thus if a_{ij} !=0 , then there is a node with head= false, value =a_{ij}, row <=I, and col=j. This node is linked into the circular linked lists for row i and column j . Hence, the node is simultaneously in two different lists. Here reading in a sparse matrix and obtaining its linked representation have assumed that the first input line gives the number of rows, the number of columns and the number of nonzero terms in the matrix. Each subsequent input line is a triple of the form(i,j,a_{ij}). These triples consist of the row, column and value data members of the nonzero terms of the matrix. Here ,assume that these triples are ordered by rows and within rows by columns, ELLIS H. et al.

4. FLOW-CHART OF PROPOSED ALGORITHM



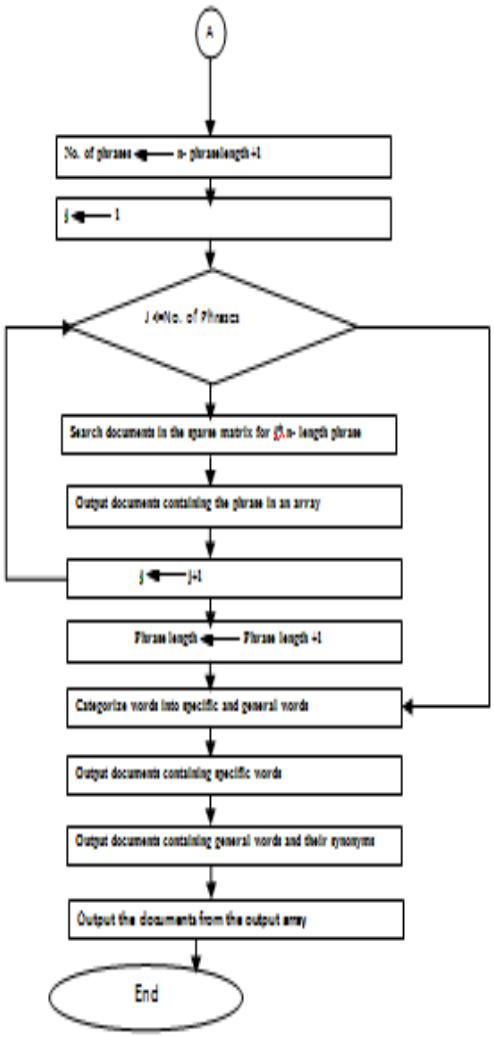


Fig 1: Flow Chart for the proposed algorithm with its steps

5. EXAMPLE

If there are three documents.

Doc 1:- This is a unique representation of complete range of literature.

Doc 2:- A cakes and bakers manufacturing company has many distributors.

Doc 3:- Every distributor has one or more counters.

Here, analysis would be based on as described earlier in introduction part of this research paper i.e. there must be every word analysed and then make a posting lists of every words in the dictionary as a nodes. There is a sorted form of dictionary alphabetically as shown in figure 1 would give dictionary in vertical part of the page and horizontal part must contain different documents and their ID's . As shown in figure 1, if the first node of the dictionary i.e, A has existence in Doc 1 and Doc 2 ,there must put pointers or arrows in their postings list to the main linked list nodes of Doc 1 and Doc 2 and if A would not be present in Doc 3 then leave it blank . For second node in the dictionary have another word , same as if this is present in Doc 2 and Doc 3. There put arrows or pointers to join these two linked list in different horizontal part of the page and make a circular linked list to join it reverse to the dictionary node by putting its frequency in the horizontal part of the nodes and join them as a circular linked list. And similarly we must apply the same pattern on every node that may be of horizontal part or vertical part of the page.One more thing is applied here that there must joined the linked list sentence wise also as shown in figure 1.

In this the research has investigated the proposed algorithm in information retrieval. There are many algorithms on information retrieval having multiple, conflicting objectives to be met. Here sparse matrix representation has given this algorithm for doing information retrieval simple. The meaning of the term information retrieval can be very broad. Just getting a credit card out of the wallet so that a person can type in the card number is a form of information retrieval. Information retrieval is finding material of an unstructured nature that satisfies an information need from within large collections.

Information retrieval used to be an activity that only a few people engaged in i.e. Librarians, paralegals and similar professional searchers. Now the world has changed, and hundreds of millions of people engage in information retrieval every day when they use a web search engine or search their email. Information retrieval is fast becoming the dominant form of information access, overtaking traditional database style searching. IR can also cover other kinds of data and information problems beyond that specified in the core definition above. The term "unstructured data" refers to data which does not have clear, semantically overt, easy-for-a-computer structure. It is the opposite of structured data, the canonical example of which is a relational database, of the sort companies usually use to maintain product inventories and personnel records. In reality, almost no data are truly "unstructured". This is definitely true of all text data if you count the latent linguistic structure of human languages. But even accepting that the intended notion of structure is overt structure, most text has structure, such as headings and paragraphs and footnotes, which is commonly represented in documents by explicit markup (such as the coding underlying web pages). IR is also used to facilitate "semi structured" search such as finding a document where the title contains Java and the body contains threading. Christopher D. M. et al[2005].

The field of information retrieval also covers supporting users in browsing or filtering document collections or further processing a set of retrieved documents. Given a set of documents, clustering is the task of coming up with a good grouping of the documents based on their contents. It is similar to arranging books on a bookshelf according to their topic. Given a set of topics, standing information needs, or other categories (such as suitability of texts for different age groups), classification is the task of deciding which class(es), Information retrieval systems can also be distinguished by the scale at which they operate, and it is useful to distinguish three prominent scales. In *web search*, the system has to provide search over billions of documents stored on millions of computers. Distinctive issues are needing to gather documents for indexing, being able to build systems that work efficiently at this enormous scale, and handling particular aspects of the web, such as the exploitation of hypertext and not being fooled by site providers manipulating page content in an attempt to boost their search engine rankings, given the commercial importance of the web. Email programs usually not only provide search but

also text classification: they at least provide a spam (junk mail) filter, and commonly also provide either manual or automatic means for classifying mail so that it can be placed directly into particular folders. Distinctive issues here include handling the broad range of document types on a typical personal computer, and making the search system maintenance free and sufficiently lightweight in terms of startup, processing, and disk space usage that it can run on one machine without annoying its owner. In between is the space of *enterprise, institutional, and domain-specific search*, where retrieval might be provided for collections such as a corporation's internal documents, a database of patents, or research articles on biochemistry. In this case, the documents will typically be stored on centralized file systems and one or a handful of dedicated machines will provide search over the collection, Christopher D. M. et al[2005].

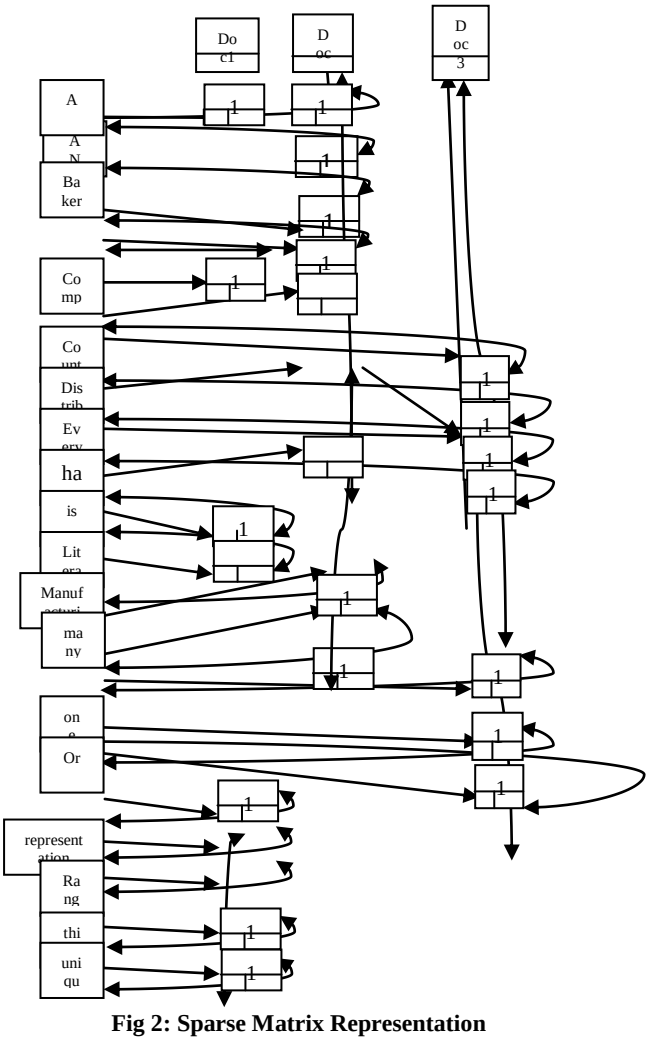


Fig 2: Sparse Matrix Representation

6. CRANTOP DATABASE

Crantop database is an xml based database in which this research is carried out. It contains 3 parts that is documents with their Doc IDs in their dictionaries. This database would contains Doc.ID 1 to Doc.ID 1400 in which there are several nodes of linked list structure which can be found by the help of software's like MS-Word . In the second database we contains different queries i.e. the research would be carried out on query 1 to 50 its analysis would be carried out in MS-Excel by sorting the every words into ascending order and make it into a new sparse matrix representation. In the third database its every documents ranking would be given that is how many times a single link list can be carried out in a single query from 1 to 50 for the Doc ID's 1 to 1400 and its rankings would be given here on the basis of that its precision and recall can be calculated here with its f-measure with different Intersection, Intersection with skip pointers and positional intersection algorithms and based on that new algorithm would be proposed and based on that second table would be analysed. This is the need of this crantop database by this research can be carried out.

7. PROPOSED ALGORITHM STEPS

7.1 Break Query

Given a character sequence in a paragraph and a defined document unit, break-row is the task of breaking it up into pieces, called *tokens*, perhaps at the same time throwing away certain characters, such as punctuation. Here is an example of break-row query:

Input: This, is, a, unique, representation, of, complete, range, of, literature;

Output:
These break-row queries are often loosely referred to as terms or words, but it is sometimes important to make a type distinction. A *paragraph* is an instance of a sequence of characters in some particular document that are grouped together as a useful semantic unit for processing. A *term* is a type that is included in the IR system's dictionary. The set of index terms could be entirely distinct from the words, for instance, they could be semantic identifiers in a taxonomy, but in practice in modern IR systems they are strongly related to the words in the document. However, rather than being exactly the words that appear in the document, they are usually derived from them by various normalization processes. Here if it is taking like a document as a query it may contain a root word, stop words, two-words phrase, three or more words phrases and when it compares with other documents in the postings lists and it retrieved some better results then it is a break row queries, Christopher D. M. et al[2005].

7.2 Phrase Queries

Many complex or technical concepts and many organization and product names are multiword compounds or phrases. The research would like to be able to pose a query such as manufacturing company by treating it as a phrase so that a sentence in a document like "A cakes and bakers manufacturing company has many distributors" is not a match. Most recent search engines support a double quotes syntax ("manufacturing company") for *phrase queries*, which has proven to be very PHRASE QUERIES easily understood and successfully used by users. As many as 10% of web queries are phrase queries, and many more are implicit phrase queries (such as person names), entered without use of double quotes. To be able to support such queries, it is no longer sufficient for postings lists to be simply lists of documents that contain individual terms. In this section we consider two approaches to supporting phrase queries and their combination. A search engine should not only support phrase queries, but implement them efficiently. A related but distinct concept is term proximity weighting, where a document is preferred to the extent that the query terms appear close to each other in the text, Christopher D. M. et al[2005].

7.3 Two –Words Phrases

One approach to handling phrases is to consider every pair of consecutive terms in a document as a phrase. For example, the text "A cakes and bakers manufacturing company has many distributors" would generate the two-words Phrase .Two-words index cakes,bakers,manufacturing,company,many,distributors In this model, we treat each of these two-words as a vocabulary term. Being able to process two-word phrase queries is immediate. Longer phrases can be processed by breaking them down. The query cakes, bakers can be broken into the Boolean query on two-words: "cakes bakers" AND "manufacturing company" AND "many distributors" This query could be expected to work fairly well in practice, but there can and will be occasional false positives. Without examining the documents, it cannot verify that the documents matching the above Boolean query do actually contain the original 4 word phrase. Among possible queries, nouns and noun phrases have a special status in describing the concepts people are interested in searching for. But related nouns can often be divided from each other by various function words, in phrases such as the abolition of slavery or renegotiation of the constitution. These needs can be incorporated into the two-word indexing model in the following way. First break the text and perform part of speech tagging. It can then group terms into nouns, including proper nouns and function words including articles and prepositions among other classes. Here if it is taking like "A cakes and bakers "this is a four words phrase and if it remove "and" stop word then it have increased its priority as a retrieval of documents. Similarly if we remove "A" then it will again of only two word phrase and here its priority of word retrieval from the database may be increased i.e. why it is divided into two-word, three-words or more words phrases because it have increased the priority of retrieving the documents from the database as less is the length of the words more will be the accurate retrieval of the words or phrases, Christopher D. M. et al[2005].

7.4 Three –Words Phrases

As it is explained in two-word phrases that if it has increased the priority of a root words in the database to be retrieved break them a phrase in single root words so that they can retrieve accurate amount of information from the database. And if it is a three-words phrase its priority must be less as compared to single root words to be searched and retrieved from the database as in the given query as a three words phrases .One approach to handling phrases is to consider every pair of consecutive terms in a document as a phrase. For example, the document "A cakes and bakers manufacturing company has many distributors" would generate the three-words Phrase as "cakes, bakers, manufacturing". These three-words index "cakes, bakers, manufacturing" In this model, we treat each of these three-words as a vocabulary term. Being able to process three-word phrase queries is immediate. Longer phrases can be processed by breaking them down. Like two-words phrases , three –words phrases are also important in information retrieval like to gain the speed benefits of indexing at retrieval time. It have to build the index in advance, the steps are as

1) Collect the documents to be indexed:

A cakes and bakers manufacturing	Company has many distributors.

2) Break words into text, turning each document into a list of tokens:

Cakes	Bakers	manufacturing	Company	…..

3) Do linguistic preprocessing, producing a list of normalized words, which are the indexing terms.

4) Index the documents that each term occurs in by creating an inverted index, consisting of a dictionary and postings.

When the dictionaries are created among the various document collections of different words then only it can take two-words phrases, three- words phrases or more words phrases. Here when applied sparse linked representation then it can be retrieved any number of words phrases. Basically, intersection algorithm is used to select common index in more than two queries and here it will fill its effect. Similarly, positional intersect algorithm is also retrieved the common positions at a query's different words. So two, three or more words phrases can be easily retrieved by using sparse linked representation effectively except other algorithms which are not too much effective as compared to their precision and recall findings.As it is proved to be that different information retrieval algorithms have slow precision and recall as compared to the sparse linked representation which is discussed above.And it is proved here, Christopher D. M. et al[2005].

7.5 Remove Stopwords

Sometimes, some extremely common words which would appear to be of little value in helping select documents matching a user need are excluded from the vocabulary entirely. These words are called stop words. The general STOP WORDS strategy for determining a stop list is to sort the terms by collection frequency (the total number of times each term appears in the document collection), and then to take the most frequent terms, often hand-filtered for their semantic content relative to the domain of the documents being indexed, as a stop list, the members of which are then discarded during indexing. Using a stop list significantly reduces the number of postings that a system has to store. And a lot of the time not indexing stop words does little harm: keyword searches with terms like the and by don't seem very useful. However, this is not true for phrase searches. The phrase query "A cakes and bakers manufacturing company has many distributors", which contains two stop words, is more precise than cakes AND bakers the meaning of it to be lost if the word to is stopped out. Some song titles and well known pieces of verse consist entirely of words that are commonly on stop lists

The general trend in IR systems over time has been from standard use of quite large stop lists (200–300 terms) to very small stop lists (7–12 terms) to no stop list whatsoever. Web search engines generally do not use stop lists. Some of the design of modern IR systems has focused precisely on how they can exploit the statistics of language so as to be able to cope with common words in better ways.

Finally, it shows how an IR system with impact-sorted indexes can terminate scanning a postings list early when weights get small, and hence common words do not cause a large additional processing cost for the average query, even though postings lists for stop words are very long. So for most modern IR systems, the additional cost of including stop words is not that big – neither in terms of index size nor in terms of query processing time.When we remove

stopwords then its query's processing speed must be increased, Christopher D. M. et al[2005].

7.6 Convert Remaining Into Root Words

Here may be a root words like representation, represented or representing. It will make the accurate data to be retrieved from the set of database as a root word. A root word for English, for example, should identify the string "cats" and possibly "catlike", "catty" etc. as based on the root "cat", and "stemmer", "stemming", "stemmed" as based on "stem". This algorithm reduces the words "fishing", "fished", and "fisher" to the root word, "fish". On the other hand, "argue", "argued", "argues", "arguing", and "argus" reduce to the stem "argu" illustrating the case where the stem is not itself a word or root but "argument" and "arguments" reduce to the stem "argument".

There are several types of root words algorithms which differ in respect to performance and accuracy.

a) Lookup algorithms

A simple root word looks up the inflected form in a lookup table. The advantages of this approach is that it is simple, fast, and easily handles exceptions. The disadvantages are that all inflected forms must be explicitly listed in the table: new or unfamiliar words are not handled, even if they are perfectly regular and the table may be large, [5].

b) The production technique

The lookup table used by a root word is generally produced semi-automatically. For example, if the word is "run", then the inverted algorithm might automatically generate the forms "running", "runs", "runned", and "runly", [6]

c) Suffix-stripping algorithms

Suffix stripping algorithms do not rely on a lookup table that consists of inflected forms and root form relations. Instead, a typically smaller list of "rules" is stored which provides a path for the algorithm, given an input word form, to find its root form [6].

d) Proposed algorithm on root words

But here in the authors present algorithms a dictionary are created vertically of the page and similarly horizontally different documents are placed in a MS-word page. To analyze all these if taken an example like "cat is staring" or "Every distributor has one or more counters", it takes the word" staring", "stare" etc in the dictionary as a form of sparse matrix representation. Or it can take "distributor", "distributed","distribution" etc. Similarly, it will make the accurate data to be retrieved from the set of database as a root word. Its main goal is to reduce inflectional forms and sometimes derivationally related forms of a root word to a common base form. There are two processes which give it flavor. Stemming usually refers to a crude heuristic process that chops off the ends of words in the hope of achieving this goal correctly most of the time, and often includes the removal of derivational affixes whereas lemmatization usually refers to doing things properly with the use of a vocabulary and morphological analysis of words, normally aiming to remove inflectional endings only and to return the base or dictionary form of a word.

7.7 Classify Into Specific And General Words

General words and specific words are not opposites instead, they are the different ends of a range of words. General words refer to groups whereas specific words refer to individuals.

Furniture is a general word it includes within it many different items. If anybody ask to form an image of furniture, it won't be easy to do. Do anybody see a department store display room? a dining room? an office? Even if anybody can produce a distinct image in their mind, how likely is it that another reader will form a very similar image? Furniture is a concrete term, but its meaning is still hard to pin down, because the group is so large.It is hard to develop much of a response, because the group represented by this general words is just too large.

If any user have taken "This is a unique representation of complete range of literature". Here if representation is a specific words and literature is a general words then if none of the document is searched out like other document if there is other document "Every distributor has one or more counters". Again same if taken as distributor and counters as specific and general words and if these roots words are there then these documents would be searched out. Here for another example if taken "Jaipur Tourism" then in such kind of documents Jaipur is a specific word and Tourism is a general word. But when the user search the Jaipur Tourism and none of the document is searched out then there is another document is taken i.e. Delhi Tourism again the user search the Delhi as specific word and Tourism as general word and something is searched out from the documents as a noun then it can take intersection or union algorithm as the matter or text to search out the document.

7.8 Synonyms

Here similar meaning root words can be searched by the user like in the example given above in the sparse matrix representation in three different documents. There synonyms can be searched out by the given algorithm like if two root words which is searched have same meaning then it can be search and this document can be retrieved. Like there are bakers, cakes are the root words in document 2 but its synonyms are in document 3 i.e. distributors as a root word. It has a similar word liked pastries in another documents so it can be search from any other query And distributors have cakes, bakers with the help of this most other root words can be search with the help of the proposed algorithm. And there are so many synonyms are available here in the dictionary as well as in the posting list and in different documents. So it is a very good feature in this algorithm as have synonyms of the similar meaning root words.

7.9 Intersection With Specific Words

Here again intersection have taken as the common postings together as if it have two Boolean queries located in the dictionary. Then intersection operation need to efficiently intersect postings lists so as to be able to quickly find documents that contain both terms. As there may be queries like "Mahatma Gandhi and Lal Bahadur Shastri are the great son of India".So here Mahatma Gandhi or Lal Bahadur Shastri are the specific words but if some people does not know about Lal Bahadur Shastri and know about Mahatma Gandhi then its document can be search with the specific word Lal Bahadur Shastri. Because it is present in that document where Mahatma Gandhi was present so this kind of documents can be retrieved by using this algorithm. Similarly if we have taken all three documents of figure 1 and use intersection with specific words then common meaning words can be searched out like representation, manufacturing and distributors. And other documents would

be retrieved here which have common postings list in the documents. Suppose there are these three documents here

Doc 1:- This is a unique representation of complete range of literature.

Doc 2:- A cakes and bakers manufacturing company has many distributors.

Doc 3:- Every distributor has one or more counters.

If users are taking all the three above documents and compare them as intersection algorithm with some specific words let us there may be another document and in that Chandimal is a businessman who sell pastries and cakes so by comparing it with other three documents Chandimal is a specific word by which other documents would be retrieved here which have common postings list in these documents.

8. CONCLUSION

A novel algorithm based on Phrase Prioritization has been proposed. Emphasis and priority has been given to the phrases, so that the user typing more than one word may be considered together and the documents containing those phrases must be given the higher priority, which should generally be the case. A data structure supporting the phrase search has been proposed, which is the linked implementation of a sparse matrix, which maps the nature of the relationship between documents and the words of dictionary. Moreover, emphasizing the need of categorizing the words into specific and general words is also a very new and beneficial idea, which can increase the relevance of those documents containing specific words as compare to general words. The proposed algorithm is a best possible solution to the phrase prioritization and its supporting data structure that is link list representation.

9. FUTURE WORK

Proposed Algorithm will be practically implemented in Python and comparison with the other existing algorithms will be presented.

10. REFERENCES

[1] Arash S et al., 2013 "Prioritization of innovation factors by the integration of concurrent function deployment and P diagram with a case study in Sepahan Industry Group", Journal of Manufacturing Technology Management, Vol. 24 Iss: 6, pp.952 – 971

[2] Dionysios S. et al., 2006 "AML-related technologies: a systemic risk", Journal of Money Laundering Control, Vol. 9 Iss: 2, pp.157 – 172

[3] Douglas B. et al., 1997 "Beyond ethnography: Towards writerly accounts of organizing in marketing", European Journal of Marketing, Vol. 31 Iss: 3/4, pp.264 – 284

[4] Ethan S. et al., 2011, Strategic Change and the Jazz Mindset: Exploring Practices that Enhance Dynamic Capabilities for Organizational Improvisation, in Abraham B. (Rami) Shani, Richard W. Woodman, William A. Pasmore (ed.) Research in Organizational Change and Development (Research in Organizational Change and Development, Volume 19) Emerald Group Publishing Limited, pp.55 – 90

[5] Hull, D. A. et al., 1996" A Detailed Analysis of English Stemming Algorithms, Xerox Technical Report"

[6] Hull, D. A. et al.,1996 " Stemming Algorithms – A Case Study for Detailed Evaluation", JASIS, 47(1): 70–84

[7] Kostas S. M. et al., 2002 "GENESYS: an expert system for production scheduling", Industrial Management & Data Systems, Vol. 102 Iss: 6, pp.309 – 317

[8] Marie Boltz et al., 2010 "Building a framework for a geriatric acute care model", Leadership in Health Services, Vol. 23 Iss: 4, pp.334 – 360

[9] Ming-Lang T. et al., 2011 "Evaluating the effectiveness of e-learning system in uncertainty", Industrial Management & Data Systems, Vol. 111 Iss: 6, pp.869 – 889

[10] Mehmet S. et al., 2008 "Hybrid analytical hierarchy process model for supplier selection", Industrial Management & Data Systems, Vol. 108 Iss: 1, pp.122 - 142

[11] Marcel M. et al., 2003 "Transparency on the Net: functions and deficiencies of Internet search engines", info, Vol. 5 Iss: 1, pp.52 – 74

[12] Shawn A. W. et al., 2006, 16. Design of a Multi-Vehicle Control System: System Design and User Interaction, in Nancy J. Cooke, Heather L. Pringle, Harry K. Pedersen, Olena Connor (ed.) Human Factors of Remotely Operated Vehicles (Advances in Human Performance and Cognitive Engineering Research, Volume 7) Emerald Group Publishing Limited, pp.223 – 236

[13] Sarah H. T. et al., 2013 , All Metadata Politics Is Local: Developing Meaningful Quality Standards, in Jung-Ran Park , Lynne C. Howarth (ed.) New Directions in Information Organization (Library and Information Science, Volume 7) Emerald Group Publishing Limited, pp.229 – 250

[14] Wei-yu K. C. et al., 2010 "An analytic hierarchy process approach to assessing consumers' distribution channel preference", International Journal of Retail & Distribution Management, Vol. 38 Iss: 2, pp.78 – 96

[15] ConvertedtoXMLfrom:ftp://ftp.cs.cornell.edu/pub/smart /cran/ (for 1400 documents)

[16] ConvertedtoXMLfrom:ftp://ftp.cs.cornell.edu/pub/smart /cran/(for 50 queries)

[17] Ellis Horowitz, Sartaj Sahni, Dinesh Mehta, Fundamentals of data structures in C++,2nd Edition ,2008, Universities Press(India) Private Limited.

[18] Christopher D. Manning(Stanford University) ,Prabhakar Raghvan(Yahoo! Research), Hinrich Schutze(University of Stuttgart),IntroductiontoInformationRetrieval,2009,Cambridge University Press.