COMPARATIVE ANALYSIS OF RELATIVE AND EXACT SEARCH FOR WEB INFORMATION RETRIEVAL

Yagnesh D. Dave[1], Bijendra S. Agrawal[2]

[1] Associate Professor, Shri Chimanbhai Patel Post Graduate Institute of Computer Applications,Gujarat,India
[2] Director, Kalol Institute of Management, Gujarat, India

Abstract

The volume of data on web repository is huge. To get specific and precise information for the web repository is a big challenge. Existing Information Retrieval (IR) techniques, given by contemporary researchers, are very useful in field of IR. Here, the authors have implemented and tested two of the techniques from the fields of IR. The authors dealt with Relative Search and Exact Search techniques one by one. Initially relative search tested on web repository data using web mining tool and then its results are analyzed. In the same manner, the exact search technique of IR tested on web repository data and the results are measured. The researchers have experienced the significant importance on exact search and relative search. The focused of the research paper is to retrieve relevant information from the web information repository. With the use of two searching criteria these can be done. With the use of the suggested methods the searchers may retrieve a relevant web data in a fewer time.

Key Words: *Web data Mining, Exact Search, Relative Search, PR, TM, CD, VSM and TASE.*

---***---

1. INTRODUCTION

The search engine is a common tool used for information retrieval from web repository. Information for multiple types are available on the public channels. The difficulties are to get exact or relative expected search responses depending on the searchers criteria. The web information retrieval involves three types of search systems: exact, relative and adaptive search. In this work relative and exact searches are analyzed for information retrieval.

The relative search will produce a resultant web data depending upon the likeliness from web repository. This method works on probabilistic model [1] which can summarize the ranks from either web page or web data. With the use of this, it can produce the nearest result.

The exact method compares the word with the already stored word in database. In the database, there works on two fields: web page name and keyword which consists of some certain kinds of keywords. The keyword belongs to classification dictionary which built from the domain sitemap [2]. This approach is also work on classification dictionary. Each entry of the classification dictionary contains a term-category pair, the contingency table for that pair and its calculated strength of association. The dictionary also consists of all possible term-category pairs with at least one searching content, category outcome.

2. RELATED WORK

Thangaraj,et. al. have indicated the ontology repository and thesaurus to get semantic web search for relative search in Information Retrieval[3]. Zhao et. al have given a framework emphasized on targeted data with the use of web crawled web pages and also for performing retrieved data depended on the location base rank. They focused on web locations and keywords from web information repository and compared keywords with its specific web page location in a pair[4]. Lin,et.al have targeted for retrieving web data depended on the retrieval time from specific web page by using temporal-textual Web queries. They proposed Time-Aware Search Engine [5]. Roy et. al. have proposed web IR technique content and intend for topic of query and explicit use of the word respectively[6]. Francès et. al. have suggested a technique for improving the document replication in a concern to web distribution techniques on the basis of cost and time effectiveness[7].

3. PROBLEM STATEMENT

From the understanding of the literature review, it is found that for obtaining the optimize and relevant retrieval searcher has to choose the right technique. The significance part of the research work has dealt with the comparison of two different techniques. The authors focused on retrieval results and also experimenting techniques to summarize with the retrieval results.

4. PROPOSED WORK

The purposed work has targeted to compare and analyzed the two different information retrieval methods for identifying the efficient retrieval from web information repository.

4.1 Methodology

The undertaken relative search methodology searches the word by opening the file before comparing all the word in the file. If the search word matches with the word in the file then it will be listed to the file. In database there be two field that are URL field which would be a filename and keyword which consists some certain kinds of keywords.

With the use of Term match, it produces a rank for retrieval depending upon the different available web source categories, which finds similar term in the web files available on web site or web domain. It also generates the relevance number of nodes for retrieving ranked based web retrieval in listed manner:

1. For identifying related categories,
    a. Determine the number of no duplicate term based on different category in a query.
    b. Determine the total frequency of no duplicate term from different title from the entire available web domain.
    c. Calculate the ratio of available web domain with the use of related term from the different categories.
2. Based on the above steps, assign rank from the step a, b and c in descending order for obtaining result from relative search.

It is to point that the retrieval rank retrieved by different web categories and these may contain similar types of relative data which the searcher is looking for. In this research paper, the authors have used categories available on web file are like document, paragraphs, sentences, files. Based on that ranking approach is quite producing a same result how the website is producing. Only difference is that this uses a combination of category and content match. The expansion from original retrieval query containing the identification depending on the basis of maximum retrieval occurrences, and also consisting titles and descriptions from the top three retrieval occurrences from most retrieving categories. The term weight of the expanded query vector, which is computed by multiplying term category.

Another methodology is based on the exact search with the objective of comparative analysis searches the word by compare with the already stored word in database. If the search word matches with database word then it will be listed to the file. In implemented database there are two fields: one is URL field which is being a filename and second one is keyword which consist some certain kinds of keyboards.

To use this method for retrieving web data, the classification method (CD) should be defined from the web domain on which the web pages are resides. This methods works on the association model which can be better for handling performance issue in all criteria. To find the exact retrieval, the searcher has to move from the web content to web page and its categories. The following conditions are defined below:

[Searching Content, Category]	[Searching Content] ¬ [Category]
¬[Searching Content], [Category]	¬[Searching Content]¬ [Category]

Each entry of the classification dictionary contains a term-category pair, the contingency table for that pair, and its calculated strength of association. The dictionary entries consist of all possible term-category pairs with at least one Searching Content, Category outcome.

5. RESULTS AND ANALYSIS

5.1 Exact Search Result

To retrieve data from the web, the first approach that the researcher has used is exact search, where the researcher has identified data by the keyword. All the results shown under the head of exact search are retrieved based on the keyword retrieved. Researcher retrieved results from all the phases i.e. documents, paragraphs, sentences and files.

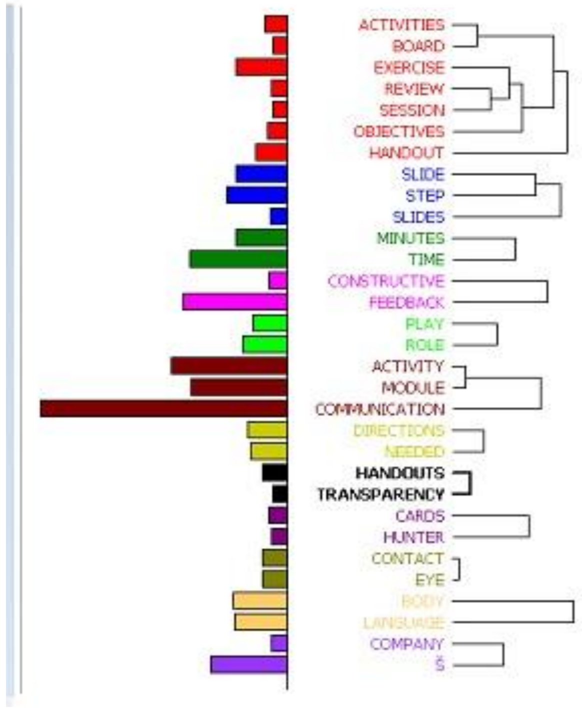
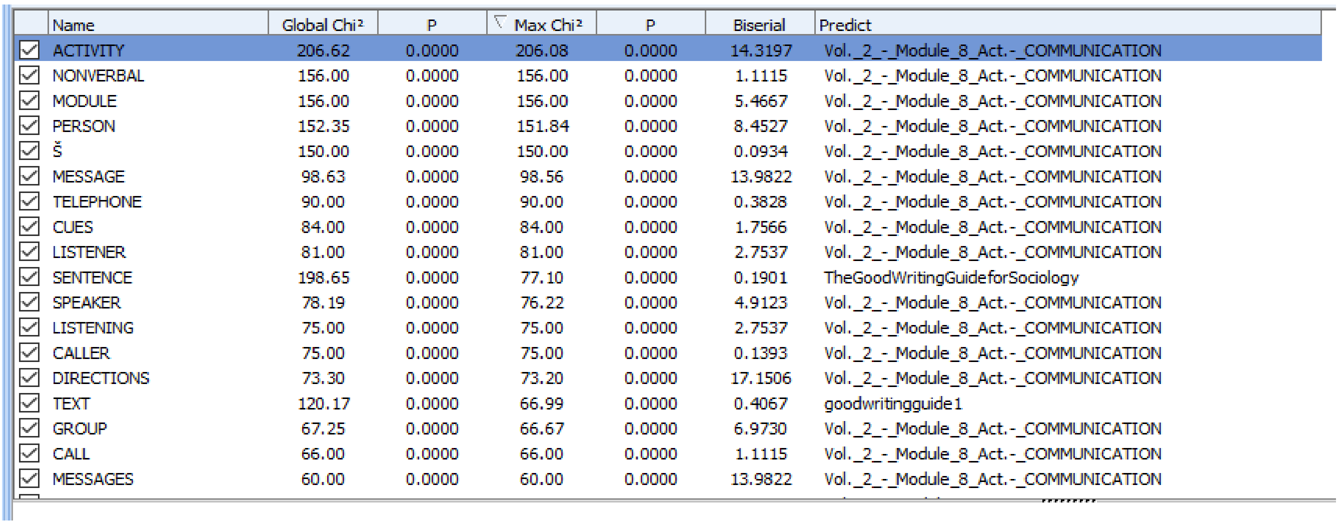


Fig-1 Agglomeration order after relatively

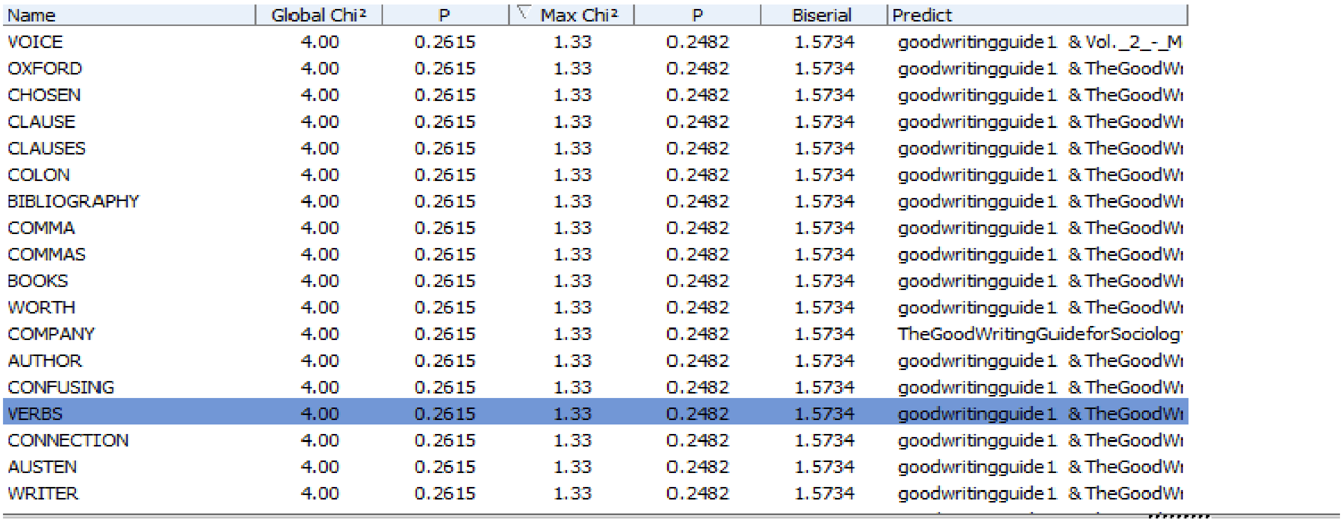
The Fig- 1 summarizes the relative frequencies retrieved keywords from the selected data and it summarizes to the maximum to minimum relative keyword found from the resultant web data. As seen in classification, the retrieval results are vast but the authors have removed clusters from the table which retrieved in figure 1.



	Name	Global Chi^2	P	Max Chi^2	P	Biserial	Predict
☑	ACTIVITY	206.62	0.0000	206.08	0.0000	14.3197	Vol._2_-_Module_8_Act.-_COMMUNICATION
☑	NONVERBAL	156.00	0.0000	156.00	0.0000	1.1115	Vol._2_-_Module_8_Act.-_COMMUNICATION
☑	MODULE	156.00	0.0000	156.00	0.0000	5.4667	Vol._2_-_Module_8_Act.-_COMMUNICATION
☑	PERSON	152.35	0.0000	151.84	0.0000	8.4527	Vol._2_-_Module_8_Act.-_COMMUNICATION
☑	Š	150.00	0.0000	150.00	0.0000	0.0934	Vol._2_-_Module_8_Act.-_COMMUNICATION
☑	MESSAGE	98.63	0.0000	98.56	0.0000	13.9822	Vol._2_-_Module_8_Act.-_COMMUNICATION
☑	TELEPHONE	90.00	0.0000	90.00	0.0000	0.3828	Vol._2_-_Module_8_Act.-_COMMUNICATION
☑	CUES	84.00	0.0000	84.00	0.0000	1.7566	Vol._2_-_Module_8_Act.-_COMMUNICATION
☑	LISTENER	81.00	0.0000	81.00	0.0000	2.7537	Vol._2_-_Module_8_Act.-_COMMUNICATION
☑	SENTENCE	198.65	0.0000	77.10	0.0000	0.1901	TheGoodWritingGuideforSociology
☑	SPEAKER	78.19	0.0000	76.22	0.0000	4.9123	Vol._2_-_Module_8_Act.-_COMMUNICATION
☑	LISTENING	75.00	0.0000	75.00	0.0000	2.7537	Vol._2_-_Module_8_Act.-_COMMUNICATION
☑	CALLER	75.00	0.0000	75.00	0.0000	0.1393	Vol._2_-_Module_8_Act.-_COMMUNICATION
☑	DIRECTIONS	73.30	0.0000	73.20	0.0000	17.1506	Vol._2_-_Module_8_Act.-_COMMUNICATION
☑	TEXT	120.17	0.0000	66.99	0.0000	0.4067	goodwritingguide1
☑	GROUP	67.25	0.0000	66.67	0.0000	6.9730	Vol._2_-_Module_8_Act.-_COMMUNICATION
☑	CALL	66.00	0.0000	66.00	0.0000	1.1115	Vol._2_-_Module_8_Act.-_COMMUNICATION
☑	MESSAGES	60.00	0.0000	60.00	0.0000	13.9822	Vol._2_-_Module_8_Act.-_COMMUNICATION

Fig-2 Classifications of Most Frequent Data

The Fig- 2 retrieved a classification of relative data from the keywords and it shows the directions and action name have more weight age then any one have.



Name	Global Chi^2	P	Max Chi^2	P	Biserial	Predict
VOICE	4.00	0.2615	1.33	0.2482	1.5734	goodwritingguide1 & Vol._2_-_M
OXFORD	4.00	0.2615	1.33	0.2482	1.5734	goodwritingguide1 & TheGoodWi
CHOSEN	4.00	0.2615	1.33	0.2482	1.5734	goodwritingguide1 & TheGoodWi
CLAUSE	4.00	0.2615	1.33	0.2482	1.5734	goodwritingguide1 & TheGoodWi
CLAUSES	4.00	0.2615	1.33	0.2482	1.5734	goodwritingguide1 & TheGoodWi
COLON	4.00	0.2615	1.33	0.2482	1.5734	goodwritingguide1 & TheGoodWi
BIBLIOGRAPHY	4.00	0.2615	1.33	0.2482	1.5734	goodwritingguide1 & TheGoodWi
COMMA	4.00	0.2615	1.33	0.2482	1.5734	goodwritingguide1 & TheGoodWi
COMMAS	4.00	0.2615	1.33	0.2482	1.5734	goodwritingguide1 & TheGoodWi
BOOKS	4.00	0.2615	1.33	0.2482	1.5734	goodwritingguide1 & TheGoodWi
WORTH	4.00	0.2615	1.33	0.2482	1.5734	goodwritingguide1 & TheGoodWi
COMPANY	4.00	0.2615	1.33	0.2482	1.5734	TheGoodWritingGuideforSociolog
AUTHOR	4.00	0.2615	1.33	0.2482	1.5734	goodwritingguide1 & TheGoodWi
CONFUSING	4.00	0.2615	1.33	0.2482	1.5734	goodwritingguide1 & TheGoodWi
VERBS	4.00	0.2615	1.33	0.2482	1.5734	goodwritingguide1 & TheGoodWi
CONNECTION	4.00	0.2615	1.33	0.2482	1.5734	goodwritingguide1 & TheGoodWi
AUSTEN	4.00	0.2615	1.33	0.2482	1.5734	goodwritingguide1 & TheGoodWi
WRITER	4.00	0.2615	1.33	0.2482	1.5734	goodwritingguide1 & TheGoodWi

Fig- 3 Classifications of Occurrences from Relevance Data

Fig- 3 shows the classification of occurrences of retrieval data depending upon the relevance. It identified the likelihood of retrieval based on the relevance.

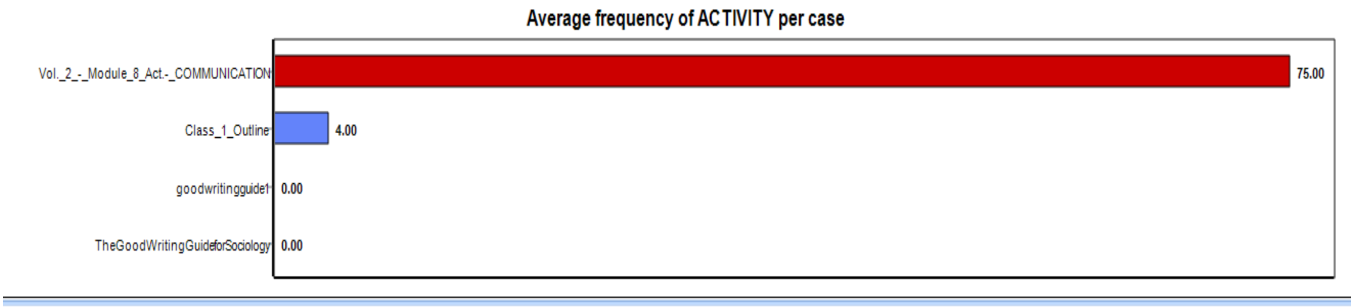


Fig- 4 Classifications of Frequencies of Relevance Data

Fig- 4 summarizes the classification chart of a frequency obtained by each case. It shows that the activity name occurred more than 75% in vol_2_module which colored as red and the blue colored activity obtained 4% in class_1_outline.

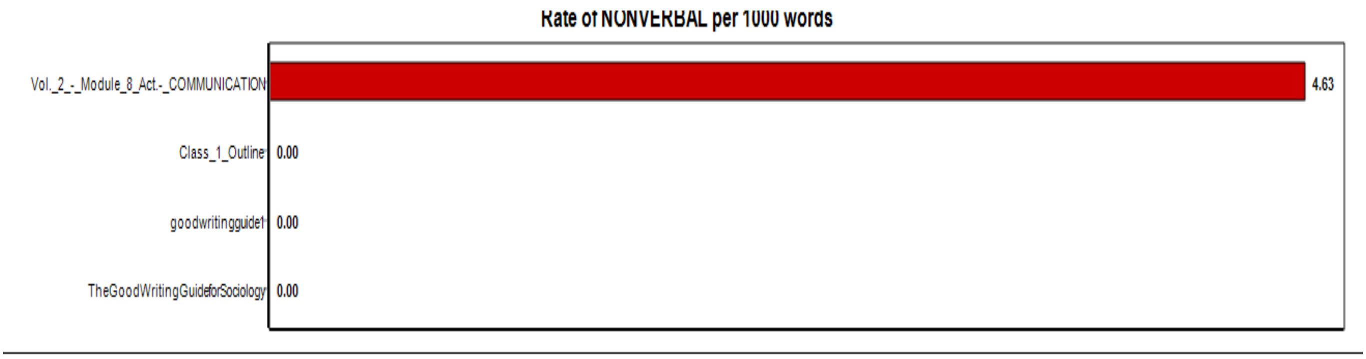


Fig- 5 Rate of Nonverbal per 1000 words

The Fig- 5 summarized that the vol_1_modules retrieved more non verbal resultant data i.e. 4.63%. it summarized that with the relative search it also retrieved non verbal keyword on the basis of relevant search.

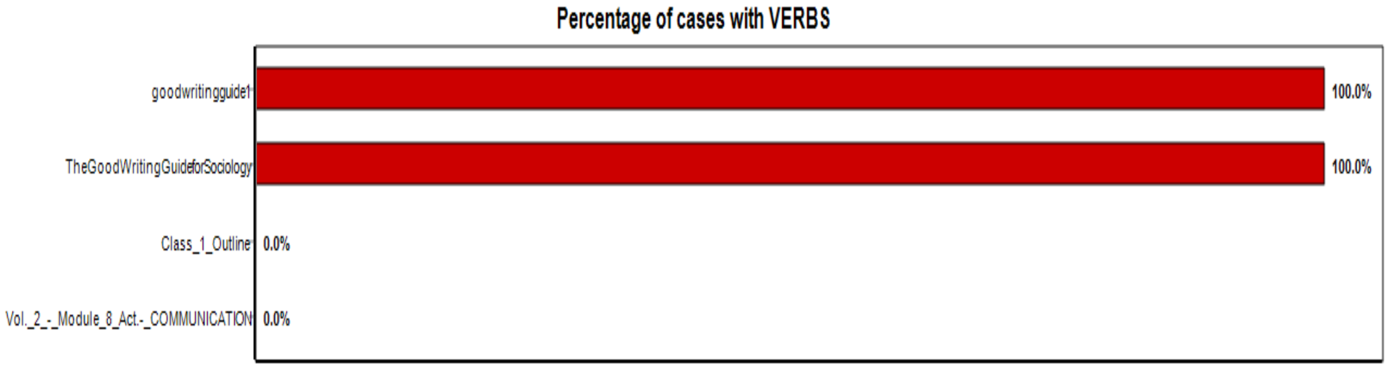


Fig- 6 Rate of verbal per 1000 words

The Fig- 6 summarized that the resultant data retrieved in the chart found the verb from the test data and found that good writing guide and the good writing guide sociology have more verbs than any case has.

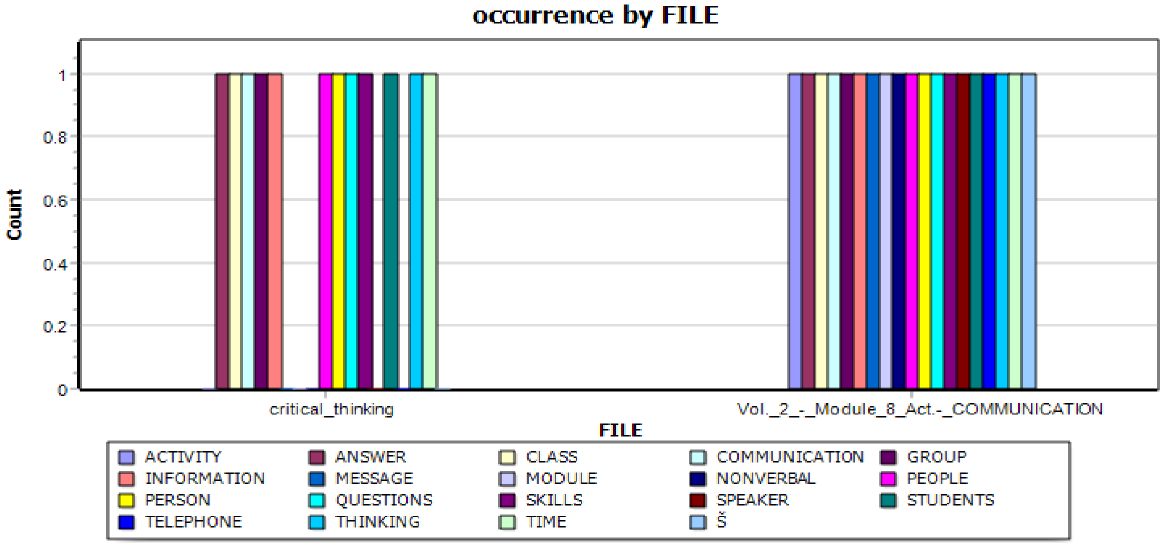


Fig- 7 Occurrences By Files

The Fig- 7 retrieved most occurrences keyword by each file and found that the different colored defined as keyword which mapped in each different file. It identified that the vol_2_module.and Act_communication have more occurrences then critical_writing test case.

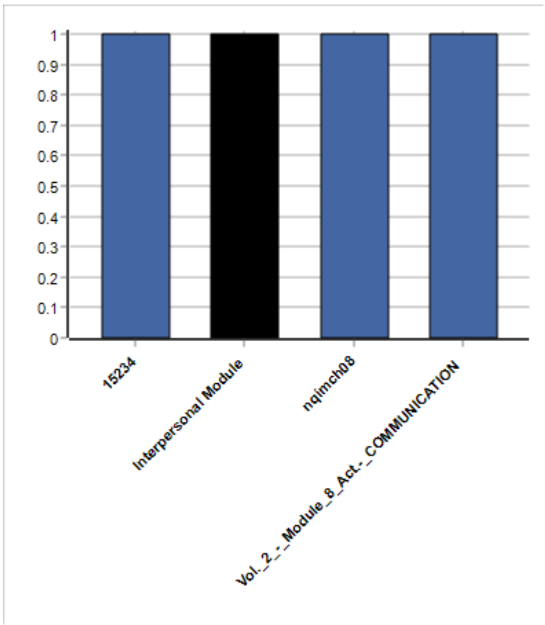


Fig- 8 Exact Searches With Case Occurrences By File

The Fig- 8 shows occurrences of retrieval data depending upon the exact search retrieval and found that the four web retrieval have common value i.e.1.

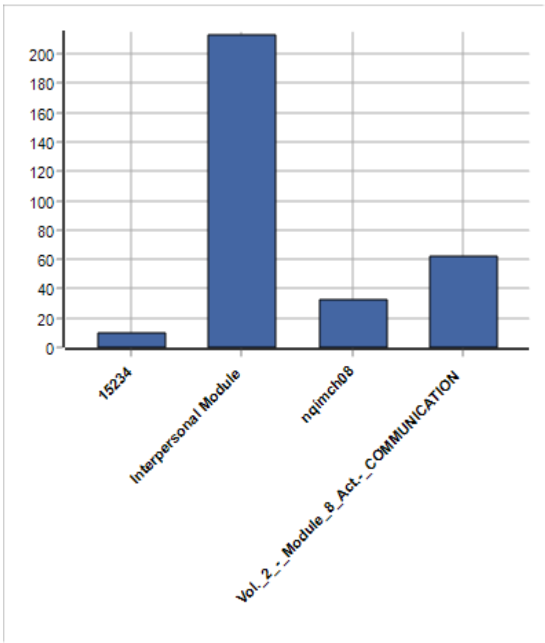


Fig- 9 Exact Searches With Word Frequency By File

Fig- 9 retrieved web information from the sentences containing in a file and focused that the interpersonal module have more frequencies than the rest. It achieves more than 200%.

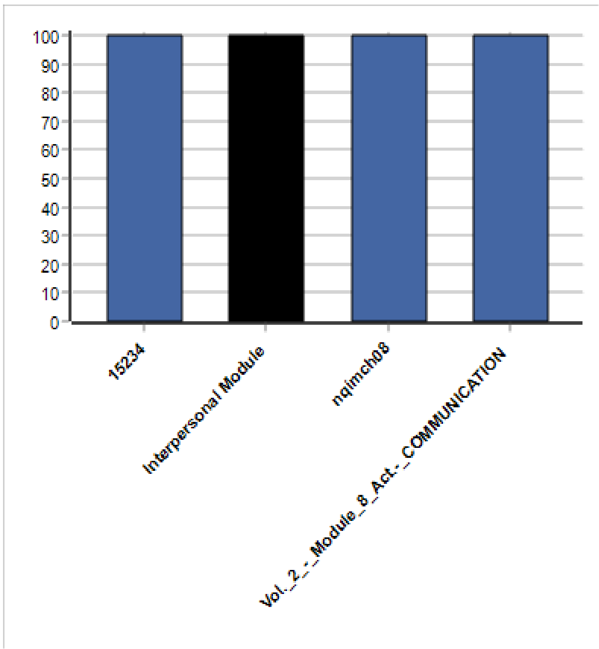


Fig- 10 Exact Searches By Case Occurrences From Sentences

Fig- 10 containing the exact searches from the case occurrence from the sentences containing in a test data and found that the black colored are the file which is in active and summarized that the all the mentioned web data files are obtaining 100% result for a specified keyword.

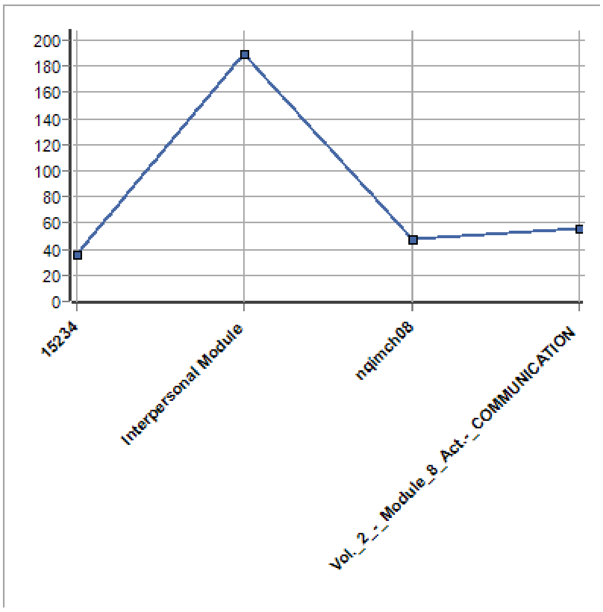


Fig- 11 Exact Searches With Rate Per 10000 By File

The Fig- 11 shown a result based on rate per 10000 words by file. It found that more than 180% achieves in the file from interpersonal module and then it immediately move down to 45% in nqimch08 and then increases in vol_2_module in 60%.

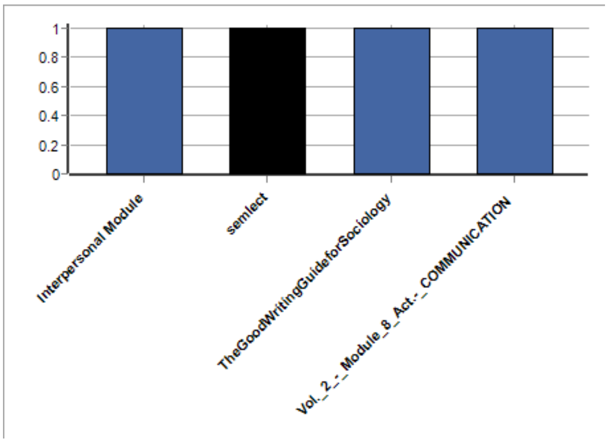


Fig- 12 Exact Searches By Case Occurrences From Sentences

The Fig- 12 retrieved web resultant data and found number of case occurrences and found that the test files are obtaining 100% exact result.

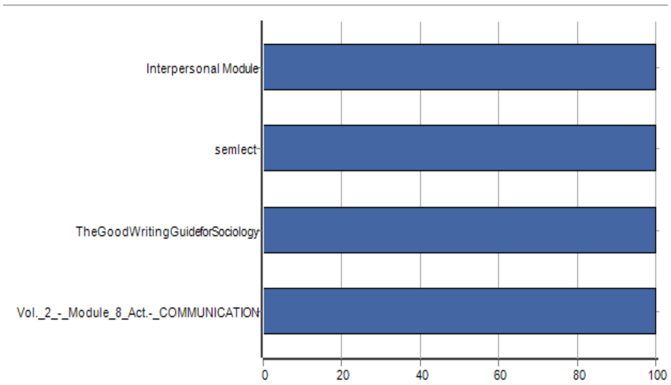


Fig- 13 Exact Searches By Category Percentage From Sentences

The Fig- 13 shows that retrieval based on the category percentage and from the selected test data, the vertical bar shown the files (test file) which obtained 100% result based on the category.

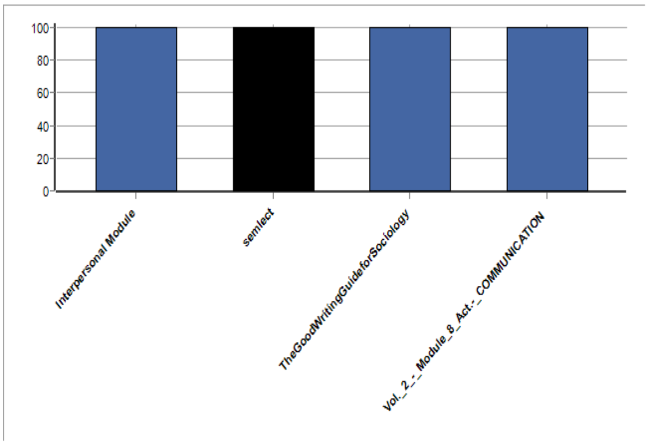


Fig- 14 Exact Searches By Case Occurrences From Sentences

The above Fig- 14 shows that case occurrences found by the sentences. It is clearly indicating that these sentences are classified by predicted classes and based on that the case occurrences are identified by the paragraphs.

6. CONCLUSION

The result analysis led to conclude that to exact search, searcher is expected to find the minimal retrieval in fewer times, where as in the relative search the searcher get vast retrieval and will also take bit more time than exact search. In the exact search will not work on uncertain or unclear keyword. In contrast to this, relative search gives comparatively improved response option and is expected to more possible values that the searcher is looking for from the web information repository. While using the relative search it also require more time to filter out the relevant resultant data. Both the approaches are useful in text and link based retrieval process. The extensibility of relative search may give more options to retrieve from web domain. The exact search produces better outcomes from different sources like web documents, paragraphs containing in the web documents, sentences containing the paragraphs. Thus with the use these option searcher retrieves the better outcome as compare to relative search. The obtained result has justified and validated that the targeted outcomes tht has been attained with integration/fusion of approaches rather than their individual uses.

REFERENCES

[1] Robertson, S.E. Maron & Cooper (1982), Probability of Relevance: A unification of two competing models for document retrieval. Information Technology: Research and Development,1,1-21.

[2] McCallum, A. Rosenfeld, R.,Mitchell & A.Y. Improving text classification by shrinkage in a hierarchy of classes. Proceeding of the 15th international conference on Machine Learning,359-367.

[3] Thangaraj, M., and G. Sujatha. "An architectural design for effective information retrieval in semantic web." Expert Systems with Applications 41.18 (2014): 8225-8233.

[4] Zhao, Jie, et al. "Exploiting location information for web search." Computers in Human Behavior 30 (2014): 378-388.

[5] Lin, Sheng, et al. "Exploiting temporal information in Web search." Expert Systems with Applications 41.2 (2014): 331-341.

[6] Roy, Rishiraj Saha, et al. "Discovering and understanding word level user intent in Web search queries." Web Semantics: Science, Services and Agents on the World Wide Web 30 (2015): 22-38.

[7] Francès, Guillem, et al. "Improving the efficiency of multi-site web search engines." Proceedings of the 7th ACM international conference on Web search and data mining. ACM, 2014.