Integrated Searching Technique for IR from Web Repository

Yagnesh Dave
Shri Chimanbhai Patel Post Graduate Institute of Computer Applications, Gujarat Technological University

Bijendra S. Agrawal, PhD
Kalol Institute of Management, Gujarat Technological University

ABSTRACT

The volume of unstructured text and hypertext data is increasing exponentially over that of well organized structured data in web data repositories. These spectacular data assets with increasing volume of repositories is generating challenges in efficient access of data to produce information by way of processing as well as extracting patterns by way of web mining. The effective and efficient retrieval as well as mining hidden patterns in a large volume of unstructured data as well as hypertext data has opened a window of research on web data mining. The undertaken research work is motivated on the centralized thought of exploratory research along with experimental justification to achieve research targets in planned research track. The research work initially carried out by the literature review of web data mining and information retrieval techniques through its models. The proposed work dealt with an integrated approach in searching technique to retrieve information from web data repository. The proposed model **Amalgamate Web Search Methodology (AWSM)** increases the level of Information Retrieval performance by integrating Exact, Relative and Adaptive search.

Keywords

Web data Mining, Exact Search, Relative Search, Adaptive search, HITS, PR, TM, CD and VSM.

1. INTRODUCTION

The research work framed and implemented by focusing the searching approach like exact, relative and adaptive search. The data are experimented under data mining environment.

1.1 Exact Search

The Boolean model in Information Retrieval system became more widely distributed, however, the general public untrained in logic found it difficult to formulate effective queries. In addition, the Boolean system retrieves depending upon the contenting document that logic may retrieve less or high volume of web data depending upon the form of the Boolean logic in the absentee of the retrieving keyword [1]. Here it is framed as a documents, paragraphs, sentences and files as source of search category of web data to be tested.

1.2 Relative Search

Rooted in probabilistic notions of the PR model focuses on the basic of information retrieval depending upon the web document or web data rank with due respect of the order of summarizing minimum possible information with relation of User's retrieval [2].

1.3 Adaptive Search

Modified HITS embraces the link analysis maxim that says a hyperlink is an annotation of human judgment conferring authority to pointed pages; it differs from other link-based approaches in several regards. Instead of simply counting the number of links, Modified HITS calculates the value of page p based on the aggregate values of pages that point to p or are pointed to by p, much in the same fashion as Page Rank [3].

2. RELATED WORK

Bucklan et al. [4] have implemented query terms expressed with Boolean operators and were compared against the inverted index terms. With the use of this the Boolean operator combining abstract concepts (AND) and synonyms (OR) as well as the fast response time from the inverted index search easily. They also observed that the general public untrained in logic found it difficult to formulate effective queries. Salton et al. [5] identified a hybrid system, called the extended Boolean system and were devised to incorporate the strengths of Boolean operators into the VS model. In the extended Boolean system, a strictness indicator called the p-value is used in conjunction with Boolean operators in the query to indicate the degree of strictness. N. Lalmas et al. [6] found that the automatic categorization of web documents is crucial for managing large amount of data available on the Internet. They found a fact that web documents are rich in structure and varied different structure. They have suggested a solution for managing with the help of web data presentation available on web page by its classification types. With the selection of a new similarity measure it affects the retrieval performance. Salton et al. [7] used the extended Boolean approach is used to decrease the computational cost. Dik Lun Lee [8] focused on the VSM for retrieval of more number of resultant data based on text retrieving approaches with the use of electronic form. They were intended to find the minimal processing with highly efficient and effective retrieval ranking from the relevant documents. Robertson et al. [9] used the probabilistic (PR) model which is similar to VS but the different between is that it's retrieves data depending upon the probable common retrieval from the search query. The PR model views the main objective of information retrieval depending on the page ranking of web documents for achieving minimal resultant data with respect to user's requirement. The basic idea of this is to calculate the term weights, which is depending on the probable relevant web data or web documents depending on the distributed query term assessed from query term basis of finding relevant retrieval. The review findings related work has indicated to develop a model that is capable to integrate all three specified challenges found thereat. The Boolean search methodology requires improvement in the significance level of search methodology. This modification will produce search output comparatively nearer to the target under search. The next modification is to use term match mechanism instead of probabilistic technique in the case when the exact search is not giving fruitful targeted results of targeted search. The next modification is carried out on adaptive search wherein the authorship attributes in addition to the attributes of page and node to obtain the increase in the preciseness of search.

3. PROPOSED WORK

The purposed work has targeted to amalgamate three separate approaches into a single method in order to attain the efficient level of search output which can be used for subsequent classification of search space responses into relevant cluster subspace responses. The Amalgamate Web Search Methodology (AWSM) has three modules namely exact search, relative search and adaptive search.

3.1 Purposed Prototype Model Methodology

The researchers have developed a model which is based on three different criteria: Exact Search, e.g. Robinson is famous personality in sport. So if user searches on the search engine, he may get multiple links (MSE) regarding Robinson. So on basis of exact search works with the help of CD. While Relative Search will use the TM. Researcher has also included VSM and WD for using this model.

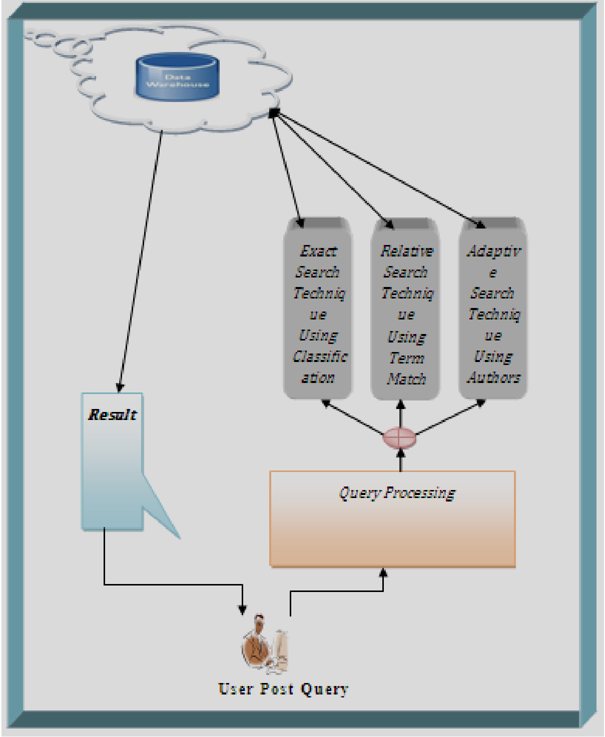
Data Warehouse
Exact Search Techniq ue Using Classific ation
Relative Search Techniq ue Using Term Match
Adaptiv e Search Techniq ue Using Authors
Result
Query Processing
User Post Query

Fig 1: AWSM Proposed Model.

And at last the Adaptive Search will work based on modified Hits. The researcher has tried to develop user friendly model.

3.1.1 Exact Search

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

The exact search 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 database there be two field : one is URL field which is be a filename and second one is keyword which consists some certain kinds of keyboards.

To retrieve resultant data from exact search, the classification dictionary is built from the domain sitemap where the files are residing. Using the association based methodology, it’s perform quite moderate in both the criteria. For each term searched by searcher from a content as a category, a possible table is demonstrated. The number of the sum for each of the probable resultant from searching content and category. Where “¬’ denotes the absence either of searching content or the category from the searcher is likely to be found. The feasible possibilities required are both i.e. what searcher is trying to search as well as the searcher must be sure about the category from the result may obtain. Searching content category, where searching content can retrieve result when the searcher is not able find the category. In second row, the first possibility is that the researcher has suggested the result from the different categories in the absences of the identified search locations. and in last case, where none of the case has mentioned so in that the searcher will not be able to find anything. As from the above table considered, the table is built and also identified that the all processed web data retrieved from the different web documents available in web directory are modified and also the results are updated. With use of the above suggested search, it gives result based on statistics provided by the association measure. Here the category defined as the types of web pages like web document (structured or unstructured).

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.

3.1.2 Related search

The Relative search is searching 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.

TM method produce a rank from available different categories to retrieve result, which matches query terms to terms in the Website sitemap files for finding a set of related nodes generates a ranked category 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 also to mention that the rank generated based in the categories may contain different kinds of related data or variables mentioned as category label. Here the researchers have used categories 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 expanded query vector consists of terms in the original query, the label of the best matching category, and the titles and descriptions of the top three sites in the best matching category. The term weight of the expanded query vector, which is computed by multiplying term category

3.1.3 Adaptive search

The concept of using adaptive search is a research which can monitor what are the pages; user visited or clicked during the current visit of the program.

After using the method for finding relative search, if still user is unable to find data from the web so hence researcher know there are only main two techniques which every search engine is working, i.e. PageRank and Hits.

HITS initially uses for text based retrieval where the S is a set of domain which responses to a query from the web data or documents who have same S to produce a graph G, which work until the G achieved and retrieve the resultant set of retrieved document with their hub based on weights and also produces a weights of authorities.

Starting with all weights initialized to 1, computes web data or web documents of every page p in T, normalizes resultant data that the calculation of the squares adds up to the weight i.e. 1, and repeats until the weights stabilize.

4. RESULTS AND ANALYSIS

4.1 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 "User". All the results shown under the head of exact search are retrieved based on the keyword retrieved as "User". Researcher retrieved results from all the phases i.e. documents, paragraphs, sentences and files.

Table 1 Exact SEARCH with document specific

Case #	***Case***	***Nb hits***
1	*action controlling event*	*1*
2	*Multimedia Project Delivering*	*1*
3	*Multimedia Project Report*	*1*
4	*Multimedia Systems*	*1*
5	*flash*	*1*
6	*HW*	*1*
7	*HW-SW*	*1*
8	*Important Questions for Advanced Database Management Systems*	*1*
9	*Introduction to flash CS4 authoring environment*	*1*
10	*motion guide*	*1*
11	*Multimedia Authoring Tool*	*1*

Table 2 Exact SEARCHES with document Classification table

Case	***Predicted class***	***Ratio***	***Economy***	***Education***	***Politic***	***Psychology***	***Sociology***
Flash	*Economy*	*20.9*	***0.9361***	*0.0006*	*0.0448*	*0.0001*	*0.0185*
Multimedia Systems	*Psychology*	*4.3*	*0.0889*	*0.1472*	*0.0724*	***0.6318***	*0.0596*
Multimedia Authoring Tools	*Psychology*	*3.8*	*0.0828*	*0.1659*	*0.0709*	***0.6244***	*0.056*
Multimedia Project Delivering	*Education*	*3*	*0.0933*	***0.4727***	*0.1348*	*0.1565*	*0.1426*
Important Questions for Advanced Database Management Systems	*Politic*	*2.4*	*0.1272*	*0.0891*	***0.4839***	*0.0975*	*0.2022*
HW	*Psychology*	*2.3*	*0.2209*	*0.0725*	*0.104*	***0.5161***	*0.0865*

4.1.1 Exact Search From specific Document by Paragraph

This section shows data identified based on exact search from document by paragraph. The data shown below retrieved in the paragraphs contained in documents from the keyword "User".

Table 3 Exact SEARCHES from Specific Document by Paragraph Classification

Case	***Predicted class***	***Ratio***	***Economy***	***Education***	***Politic***	***Psychology***	***Sociology***
Flash	***Economy***	*20.9*	***0.9361***	*0.0006*	*0.0448*	*0.0001*	*0.0185*
Flash	*Economy*	*20.9*	***0.9361***	*0.0006*	*0.0448*	*0.0001*	*0.0185*
Flash	*Economy*	*20.9*	***0.9361***	*0.0006*	*0.0448*	*0.0001*	*0.0185*
Flash	*Economy*	*20.9*	***0.9361***	*0.0006*	*0.0448*	*0.0001*	*0.0185*
Flash	*Econ*	*20*	***0.93***	*0.000*	*0.0*	*0.000*	*0.01*

	omy	.9	**61**	6	448	1	85
Multimedia Systems	Psychology	4.3	0.0889	0.1472	0.0724	**0.6318**	0.0596
Multimedia Systems	Psychology	4.3	0.0889	0.1472	0.0724	**0.6318**	0.0596
Multimedia Systems	Psychology	4.3	0.0889	0.1472	0.0724	**0.6318**	0.0596
Multimedia Systems	Psychology	4.3	0.0889	0.1472	0.0724	**0.6318**	0.0596
Multimedia Systems	Psychology	4.3	0.0889	0.1472	0.0724	**0.6318**	0.0596
Multimedia Systems	Psychology	4.3	0.0889	0.1472	0.0724	**0.6318**	0.0596
Multimedia Authoring Tools	Psychology	3.8	0.0828	0.1659	0.0709	**0.6244**	0.056
Multimedia Authoring Tools	Psychology	3.8	0.0828	0.1659	0.0709	**0.6244**	0.056
Multimedia Authoring Tools	Psychology	3.8	0.0828	0.1659	0.0709	**0.6244**	0.056
Multimedia Authoring Tools	Psychology	3.8	0.0828	0.1659	0.0709	**0.6244**	0.056
Multimedia Authoring Tools	Psychology	3.8	0.0828	0.1659	0.0709	**0.6244**	0.056
Multimedia Authoring Tools	Psychology	3.8	0.0828	0.1659	0.0709	**0.6244**	0.056
Multimedia Project Delivering	Education	3	0.0933	**0.4727**	0.1348	0.1565	0.1426
Multimedia Project Delivering	Education	3	0.0933	**0.4727**	0.1348	0.1565	0.1426
Multimedia Project Delivering	Education	3	0.0933	**0.4727**	0.1348	0.1565	0.1426
Multimedia Project Delivering	Education	3	0.0933	**0.4727**	0.1348	0.1565	0.1426
Multimedia Project Delivering	Education	3	0.0933	**0.4727**	0.1348	0.1565	0.1426
Important Questions for Advanced Database Management Systems	Politic	2.4	0.1272	0.0891	**0.4839**	0.0975	0.2022
Important Questions for Advanced Database Management Systems	Politic	2.4	0.1272	0.0891	**0.4839**	0.0975	0.2022
HW	Psychology	2.3	0.2209	0.0725	0.104	**0.5161**	0.0865
HW	Psychology	2.3	0.2209	0.0725	0.104	**0.5161**	0.0865
HW	Psychology	2.3	0.2209	0.0725	0.104	**0.5161**	0.0865

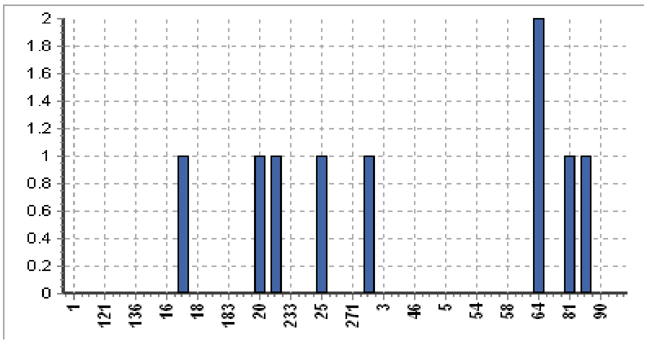

Figure 2 Exact Searches with Case Occurrence Document by Paragraph

The above figure shows that case occurrences found by the paragraph. It is clearly indicating that these paragraphs are classified by predicted classes and based on that the case occurrences are identified by the paragraphs. As this figure shows the paragraph 64 retrieves the most case occurrences by the paragraph. It calculates from the length of paragraphs and retrieval result.

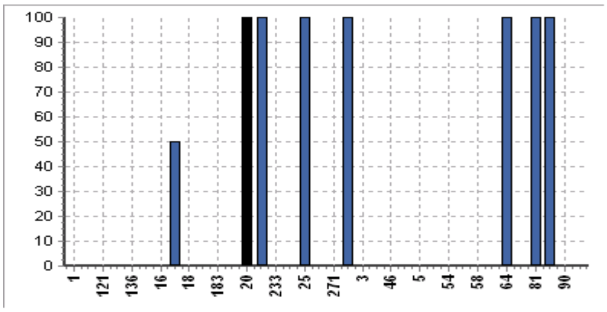

Figure 3 Exact Searches with Category Percent from Document by Paragraph

The above figure represents the percent by category. All these paragraphs which are defined on x -axis represent the occurrences of percent and it shows that paragraphs 233,25,271,64,81 and 90 are achieving 100%. While as paragraph no.16 is obtaining only 50% due to the nature of paragraph. It identified based on the length of paragraph and obtained the keyword from the paragraph.

The above figure identified the maximum occurrences from document by paragraph. That identified paragraph numbers and frequencies of that occurrence in a paragraph contained in documents

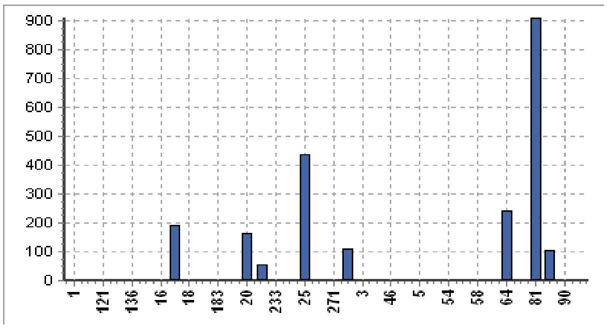

Figure 4 Exact Search with Rate per from Document 10000 words by Paragraph

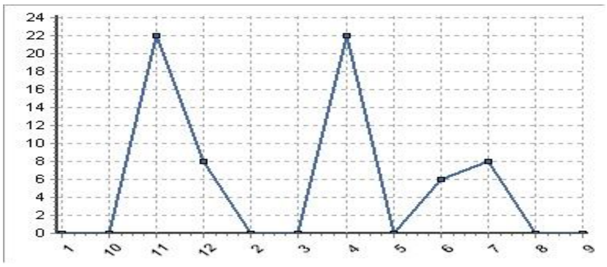

Figure 5 Exact Searches with Coded Statements with Case Occurrence

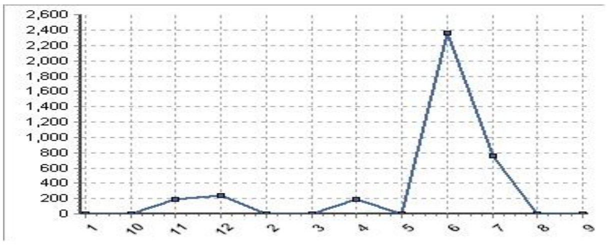

Figure 6 Exact Searches with Coded Statements with Rate per 10000 Words

4.2 Relative Search

This section shows retrieved results based on relative search. This section retrieved results from the files, documents, sentences and paragraphs. to retrieve results based on relative search , it identified any keyword containing "Us" anywhere in either files or documents or paragraphs or sentences.

Table 4 RELATIVE SEARCHES by document

Case #	*Case*	*Nb hits*
1	*action controlling event*	*1*
2	*Multimedia Project Delivering*	*1*
3	*Multimedia Project Report*	*1*
4	*Multimedia Systems*	*1*
5	*flash*	*1*
6	*HW*	*1*
7	*HW-SW*	*1*
8	*Important Questions for Advanced Database Management Systems*	*1*
9	*Introduction to flash CS4 authoring environment*	*1*
10	*motion guide*	*1*
11	*Multimedia Authoring Tools*	*1*
12	*Multimedia Project Delivering*	*1*
13	*Multimedia Project Report*	*1*

Table 5 Relative SEARCHES by document classification

Case	***Predicted class***	***Ratio***	***Economy***	***Education***	***Politic***	***Psychology***	***Sociology***
Multimedia Systems	*Psychology*	*129.*	*0*	*0.0076*	*0.004*	***0.9878***	*0*

		5			6		
action controlling event	Psychology	129.5	0	0.0076	0.0046	**0.9878**	0
flash	Economy	20.9	**0.9361**	0.0006	0.0448	0.0001	0.0185
Multimedia Project Report	Education	5.1	0.0001	**0.8266**	0.0098	0.1633	0.0001
Multimedia Systems	Psychology	4.3	0.0889	0.1472	0.0724	**0.6318**	0.0596
Multimedia Authoring Tools	Psychology	3.8	0.0828	0.1659	0.0709	**0.6244**	0.056
Multimedia Project Delivering	Education	3	0.0933	**0.4727**	0.1348	0.1565	0.1426
Important Questions for Advanced Database Management Systems	Politic	2.4	0.1272	0.0891	**0.4839**	0.0975	0.2022
HW	Psychology	2.3	0.2209	0.0725	0.104	**0.5161**	0.0865

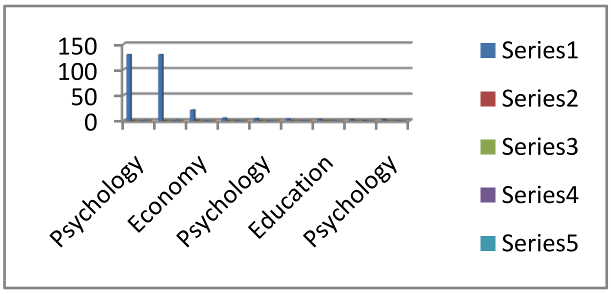

Figure 7 Relative Searches by Ratio Classification

This Figure shows the result based on ratio found in table 5 and it shows that the multimedia systems and action controlling file have a major impact on the relative search rather than the other. The x-axis is the retrieval documents from the resultant keyword are achieved by the researcher.

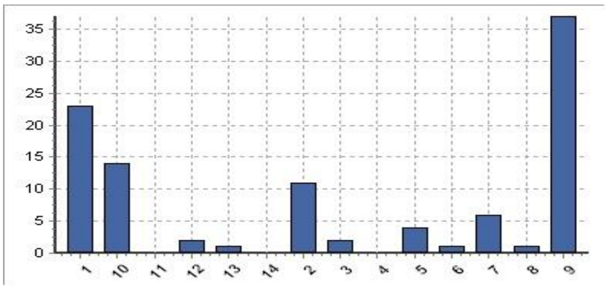

Figure 8 Relative Searches with Paragraph Frequency By Paragraph

The number of frequency identified based on retrieval is shown in figure 8. It identifies that Case number 9 have retrieved most of the cases as compare to the all. It retrieves more than 35 times.

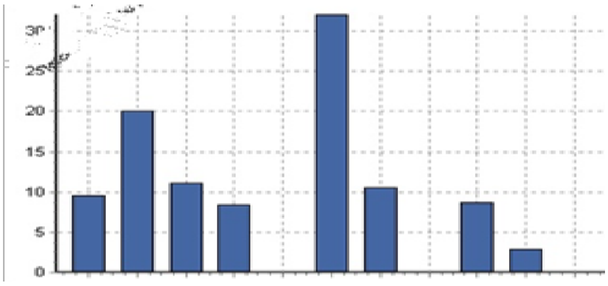

Figure 9 Relative Searches with Same File with Category Percent

The figure 9 retrieves the category percentage based on the different file on the basis of category.

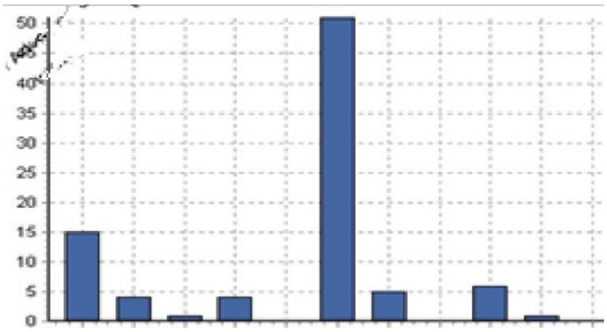

Figure 10 Relative Searches with Same file with Word Frequency

The above figure retrieve that resultant keyword is available in the shown file which have highest level of finding relative keyword and i.e. action controlling event file. This file found more number of resultant data as compare to other resultant files.

4.3 Adaptive Search

This section identifies the retrieval results based on the either on number of hits or based on the key content. This identifies the resultant data based on the nos of hits generated while retrieving result. It identifies that if the user has clicked or hits on the targeted host, it definitely have the resultant data into that document.

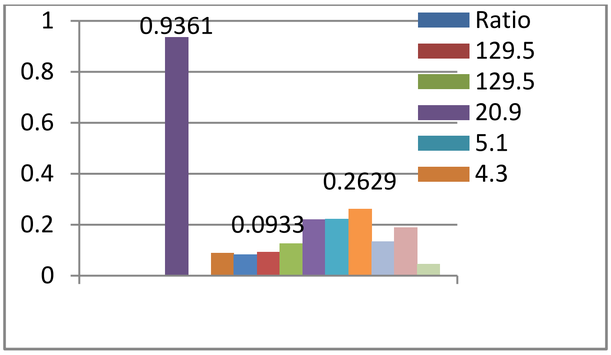

Figure 11 Adaptive charts for Document Classification

This figure identifies the classification ratio retrieved with the help of adaptive search and it clearly shows that purple bar indicating the highest ratio i.e. .9361 of occurrence with the adaptive search .

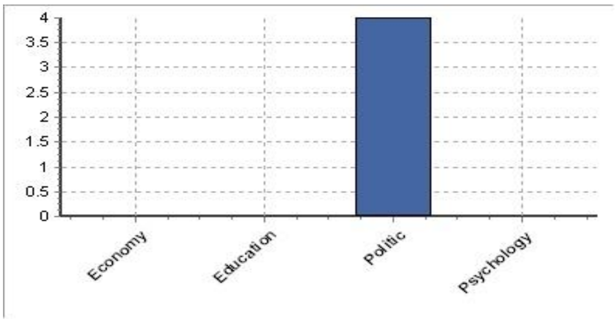

Figure 12 Adaptive Word Frequencies

The figure 12 indicates that the resultant data retrieved in politics category is higher in that case from the above figure 12. Both the figures are retrieving result from the specific category.

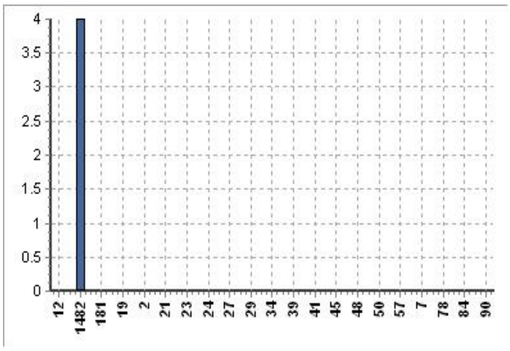

Figure 13 Adaptive Word Frequencies in Words

This figure 13 is of the classification of politics where the frequency is on highest level for occurrence as compare to the entire document retrieved with the help of adaptive search.

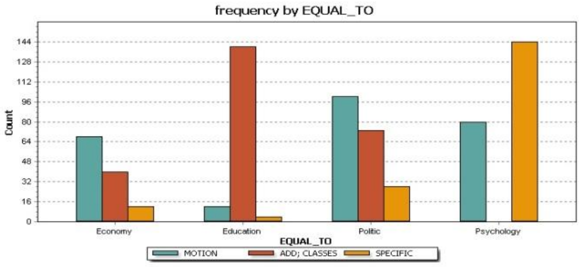

Figure 14 Adaptive Word Frequencies by Equal to

The figure 14 focused on the classification shown in three different colors like orange, carrot and sky blue. This identifies that the ec onomy, politics categories containing the motion while as the education category focused on add; classes and the psychology category contain specific retrieved.

5. CONCLUSION

The research paper undertaken has explored the selected approaches confirmed to text and link based information retrieval. From the results it has been found that if the searching keyword is of short string, exact search will produce better outcomes in comparison to rest of the approaches. The attempt to retrieve relevant information without using the keyword, the relative search produces comparatively faster retrieval but the resulting outcome has larger volume. In this approach additional time is needed to filter relevant information. The integrated approach in combination delivers higher search performance level compared to the outcomes in case of individual approaches. The obtained result has Indicated and justified the performance level attained with integration/fusion of approaches rather than their individual uses. The Framework enables enhance, multi-parametric information retrieval over web data based on distinguishing properties and attributes of files, documents and sentences and allows to enhance the conceptual consolidation during the retrieval time. As future extension, the authors commit for significance enhancement by increasing accuracy level.

6. REFERENCES

[1] 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.

[2] 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.

[3] Marchiori "The Quest for Correct Information on the web", Italy,2005.

[4] Buckland, M. Chen Mapping Entry Vocabulary to Unfamiliar metadata Vocabularies D-Lib Magazine,5(1),1999.

[5] Salton. G.&Buckley Weighting Approaches in Automatic Text Retrieval. Information Processing and Management,24,513-523,1998.

[6] N. Lalmas,M.& Fuhr,N.(1999) A Probabilistic description oriented approach for categorizing Web documents. Proceeding of the 8th ACM International Conference on Information and Knowledge Management, 475-482.

[7] Salton, G. Buckley and Allan J.(1999). Automatic restructing and retrieval of text files. Communications of the ACM.

[8] Dik Lun Lee, Kent E. Seamons. Documents Ranking and Vector space model, IEEE Software, March-April, 1997.

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