Analysis of Web Content Filtering Factors and the Impact of Sieve Coupons

Neha Gupta [#1], Saba Hilal [*2]
[#] Department of Computer Applications, Manav Rachna International University
Sector – 43, Suraj kund Road, Faridabad, Haryana, India
[1] neha.fbc@mriu.edu.in
[*] Professor, School of Computer Sciences,
Lingayas University, Nachouli, Faridabad, Haryana, India
[2] saba21hilal@gmail.com

Abstract— **For Kids searching the relevant content and weeding out the distracted content is the most challenging task as there is no fix rule to classify which content is appropriate for a kid who is searching to complete his homework and which is not.**

The Main aim of the study is to design an IR in such a way that kids remain focused on what they are searching, instead of distracting and jumping over to sites which may cause mental stress and let them lost in this big world. The Paper presents the impact of sieve coupons to filter the search of a child, giving quality and age appropriate content to them.

Keyword- Content Filtering, Sieve Coupons, information retrieval, Kids, Students

I. INTRODUCTION

The Web creates new challenges for information retrieval. The magnitude of information on the Web is growing swiftly as well as the new entrants are inexpert in the art of Web search. It is estimated to contain approximately one billion pages of openly available information and it continued to grow at an exponential rate, tripling in size over the past two years. Now the amount of information available on the net is unimaginable

There is so much data on the web that websites devoted to searching the internet, like Google, have developed. Search engines take queries and return results. Textual queries are the way a user describes the information he or she needs. The results returned are web pages that the search engine gathers containing the information the user's query express. For instance, a user might type in the query "Manav Rachna College" hoping to find information about Manav Rachna College. In this case, Google or any other search engine would return a list of pages that it found included the words "Manav Rachna College." Still there is no guarantee that the list would contain useful results and the user may have to go through pages of results to find the desired information. A Search Engine helps to overcome these problems and navigate through the Internet to get the information that one requires quickly and accurately.

It points to the strong need of the tailor made services and content designed around the needs of the individual and that, which is available at a time and place and in a form, which suits the learner's needs [1]. The searching of educational information, content and material requires the development of better web content finding tools and techniques. Most of the times the required content is available on the Web but finding it is difficult. The Search Engines help to extract the information bundles from the vast ocean of the Web. However, finding of the correct collection still remains unsolved [1], moreover, most of the time the search engine is not designed for the purpose that matches the user searches perspective. And above all if the content is found, the possibilities of distracting from the main content to the useless content is more.

Thus, there is an increasing need for research aimed at understanding children's information needs and to provide IR systems that suit the characteristics of content for children. The Main aim of the study is to design an IR in such a way that kids remain focused on what they are searching, instead of distracting and jumping over to sites which may cause mental stress and let them lost in this big world.

II. CHALLENGES IN INFORMATION RETRIEVAL FOR KIDS

A. The need for query assistance

The use of longer queries to retrieve children-friendly content can be one of the causes of the lower retrieval performance found for theses queries since it has been shown that longer queries have poorer performance in current search engines. For this reason we consider that refining long queries is highly beneficial for these users. Additionally, query rewriting methods may also be explored to rewrite phrases to keywords or entities to boost the retrieval performance. Cue words are also a valuable resource to rewrite queries by using the cues terms associated to the contexts words or to the entities that occur in the query [2]. This method can ease the exploration of information by providing different content types and dimensions of

the topic being searched . Finally, query assistance functionality should also help and train children to improve their skills to search the web [2].

B. Difficulty of searching for kids

The low percentage of informational queries found in the children queries suggests that although these users are familiar with interactive applications, they are not fully harvesting the information content available in the Web and if they are able to found the content, they may land up on a website which may contain the content which may not be appropriate for them. This behavior may be due to the lack of expertise formulating information needs using keywords, to the difficulty of identifying relevant information from the web results or to the lack of more friendly methods to guide the search. This difficulty was also observed in the greater number of entries and longer duration of sessions.

This user search behavior suggests that more efficient ways to gather the information and more efficient ways to present it to the user are required for these types of users so that kid will get what he wants instead of any irrelevant, insecure, harmful and porn data.

The identification and clustering of sieve coupons shown in this paper has potential to assist the selection of relevant verticals for the information needs of children and above all provide the content which is secure and non distracting. Although current IR systems for children as Yahoo Kids! Or Ask Kids already offer categories associated to some of the cue words we identified in the clusters (e.g. games, coloring pages, poems, jokes, homework help, etc.), the dynamic selection of verticals for each query is still very limited. And if a kid is searching something for his school curriculum he or she may not only rely on these verticals as he/she may need the information from the current database. Current IR systems only provide very simple methods to aggregate results from the verticals. IR systems could parse the content of the web results and aggregate only the content type suggested by the sieve coupon of the query; this would highly reduce the cognitive load of users to find relevant information.

C. Questions that needs to be answered while searching for kids

Q1. Do the available search engines display or extract the relevant content as per the need of the child aged 8 to 12 years?

Q2. What is the impact of adding simple keywords like kids with search term? Do they improve performance?

Q3. To improve the search experience of the child, What are the most appropriate information types and sources that an IR system should select given a child's information request?

Q4. Can we improve the search experience of a child with effective presentation method?

Q5. What is the role of quality in terms of content extracted by the search engine for kid?

III. Related Researches

The paper [3] discusses various characteristics of kids that one should keep in mind while developing any information retrieval system for kids. The experiments and results are shown using adult based search interfaces. The author of [4] has focused on information need of kids by examining their search behaviour. The author has tried to find the answer of few questions that needs to be answered by any IR system of kids. Another paper [5] analyses the session and queries for kids information need and compare them with general queries and sessions. The author has enriched the AOL query log by implementing the result of kid's queries. The paper [6] presents the use of query assistance and search moderation techniques for kids so that kids have a better experience searching the web. The author has also focused on interface design for kids. The author [7] has analysed a large query log from a commercial search engine and identify the problems related to child search behaviour. The target audiences of his work was child from age 6 to adult of 18. They have also worked on search difficulties based on query matrices.

The paper [8] presents a study of 64 fifth grade students who were using science library catalogue for searching the content on the web. The study highlights the problems of kids while searching and the possible solution. The paper [9] presents an automatic way of identifying the web page suitable for kids. The focus is on child psychology and cognitive science. The author has investigated the potential of combining topical and non-topical aspect of identifying age appropriate content for kids. The paper [10] discuss cognitive specifies of children and the way they can be encoded for classification. The author has worked on two dimensions: child friendliness and focus toward child audiences. The author [11] discuss project Gutenberg to make available classic literature to children in a secure way. The paper [12] presents an interaction based information filtering system for kids. This system focuses on user interaction modelling, user evaluation , automatic detection of child friendly information etc. The paper [13] presents a system named Tad Polemic which will assist children in searching the web for difficult topics and also provide filtering of content based upon child interest and age. The author of [14] has conducted a study to gather the quantitative and qualitative data about children interaction with web search engines. They identified that kids perform poorer on metaphorical interfaces and good on Google. The paper [15] presents a paradigm to identify the suitable videos for kids on youtube on the basis of various features like people reviews, comments, author information, community information etc. The

paper [16] tries to uncover methods and techniques that can be used to automatically improve search results on queries formulated by children. The author presents a prototype of a query expander that implements several of these techniques.

IV. IMPORTANT FACTORS FOR KIDS' CONTENT FILTERING

Children's information retrieval is not just about a child searching for information. First of all, 'a child' is a very broad term. What 'groups' of children are we talking about and do the components in the information retrieval paradigm change by different characteristics such as age, gender, reading skills, computer experience and cognitive developmental stages? To compare different 'groups' of children, it is important to group children with the same characteristics, so that found effects can be associated with the differences in that particular characteristic. While conducting the research, following factors has been kept in mind for retrieving the content from the web. These factors are:-

A. Relevance of data (Age appropriate content)
B. Quality of data
C. Secure data
D. Presentation of data(Ease of use)

A. Relevance of data (Age appropriate content)

What information are children looking for? In other words, what is a child's information need? While searching the content on the web, child gets million of pages as search result. It is a well known fact that all the pages are not relevant. The main challenge is to display the content that is age appropriate for the child. Most of the current web search engines are designed for adults and the content they display is also for adults. So the main challenge is to extract the content suitable for kids according to their age. Lets take a term "structure of leaf" as an example. Let's group children in two age groups 8 to 10 and 11 to 13. If a child of age group 8 to 10 will search structure of leaf he will need only the basic structure of the leaf describing only the main parts while the child between the age group of 11 to 13 will be looking for a detailed description about the leaf structure. So while filtering the content for the child it is necessary to group the content according to the age requirement of the child. Important relevance criteria can be topicality, novelty, interest, clarity and completeness.

B. *Quality of data*

Child can retrieve data by formulating a query but to determine the quality of the data child need assistance and training from their elders. Another way to determine the quality may be repeated searching. Web consists of millions of pages, and to determine the quality and authenticity of data is next to possible. The quality of data can only be determined if the child has previous knowledge about the subject, or he has been given instruction in how to search and navigate electronic resources, and how to judge the relevance of results to meet their information needs. There are certain sources on web which are authenticated by certain experts like website of Britannica, encyclopedia, Wikipedia, certain journals, e-books etc. Apart from all this the age of the child also plays an important role for determining the quality of the data. According to analysis done in my research, the child between age 8 to 10 needs supervised browsing to determine the quality of data and the child between ages 11 to 13 can determine the quality of data at his own, but it is not true in every case.

C. Secure data

Security is the main concern of every parent while giving his child the platform to search the content on web. As the content on the web is very large and there is no authority to check the invalid and inappropriate content so to provide secure data to the child is the most challenging issue. Secure data means data which is valid, have quality and does not contain any link, image or text which is harmful or inappropriate for the child. Web is full of porn content as well; to protect the child from such content which can cause mental stress is the most important task.

There are a lot of educational software's , filters, monitoring soft wares etc are available which do the necessary filtering but are restricted and don't give access to the current knowledge base. The aim of the research is to filter the content for kids in such a way that only secure content is displayed to the kids. For this purpose we have introduced the concept of sieve coupons in our research.

D. Presentation of data (Ease of use)

After an IR-system has run a query, it finds relevant results. It is important to examine how these results can be presented best for children: on the same page on which the child is searching (simultaneous) or on a new page (sequential). It is also important to examine which results must be presented first: the most

relevant results by scoring and ranking the matching documents, or the documents that are most referred to by others. Further, how the individual results can best be presented for children is important: with or without a short summary of the found document. Another question is how the link labels of the results must be formulated to help children make the optimum choice. Finally, the differences and similarities between adult and child preferences for all these aspects have to be examined. Search performance on these different variants of result presentation must be examined to help decide what works best for children. In my research, the focus is on mobile based presentation of results. Now days most of the parents browse net on their mobile and even child find it convenient to browse the content on mobile, as it is always connected, easily available and handy. So we have focused on mobile based presentation of results in a convenient manner.

IV. OBJECTIVE OF THE RESEARCH

For Kids searching the relevant content and weeding out the distracted content is the most challenging task as there is no fix rule to classify which content is appropriate for a kid who is searching to complete his homework and which is not.

The Main aim of the study is to design an IR in such a way that kids remain focused on what they are searching, instead of distracting and jumping over to sites which may cause mental stress and let them lost in this big world.

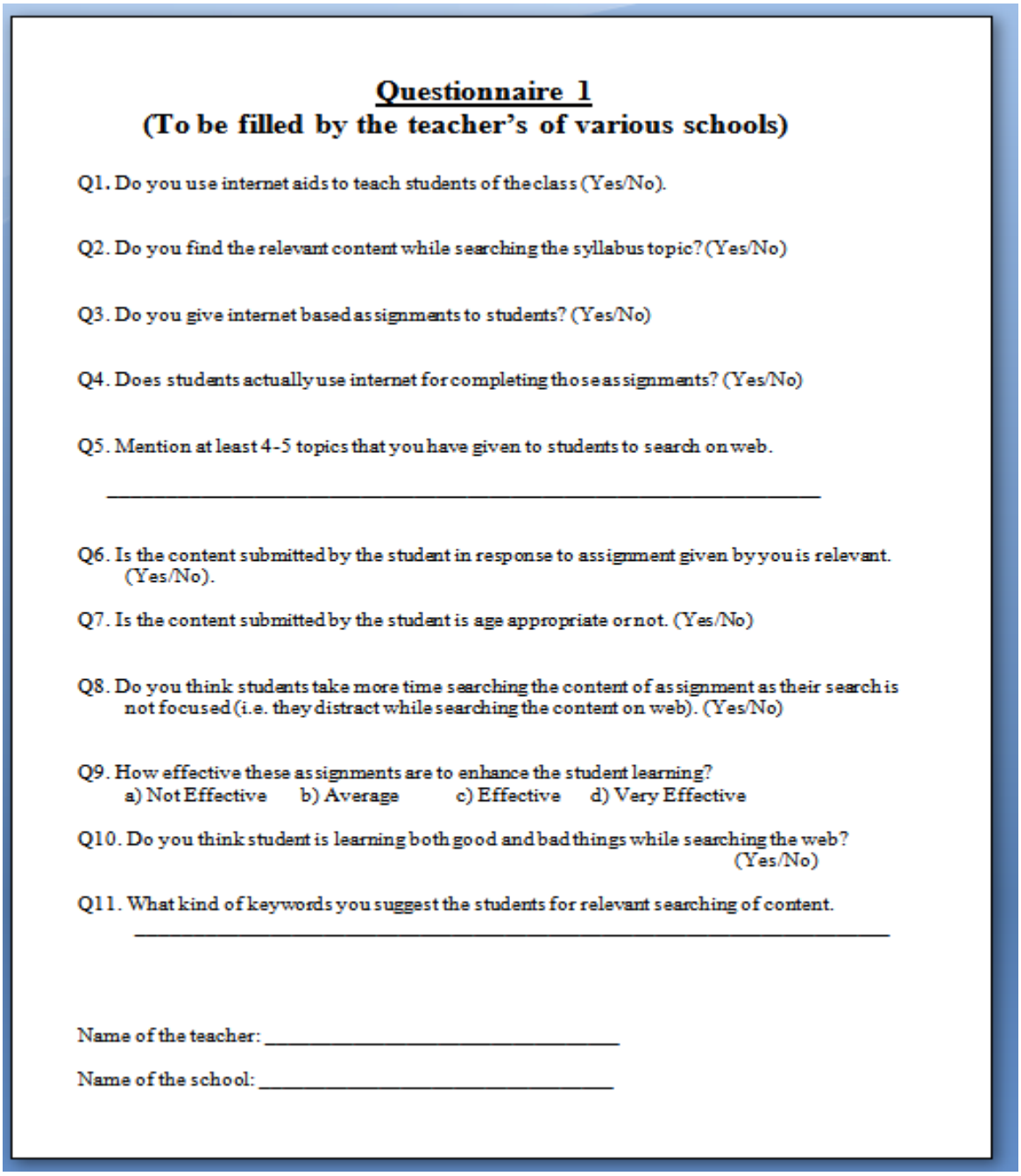
As we are working on designing an IR that suits kids' requirement, we classify our approach into top down layer system, where each layer performs a task to enhance the query entered by the kid in such a way that the kid will get the relevant content, hiding most of the irrelevant, distracted, porn or useless content. To achieve the above mention procedure we have introduced the concept of sieve coupons which will do the filtering and mine the web in such a way that only the useful information will be displayed to the kids' hiding others.

While searching, the human point of view, decisions and tricks are recorded and sieve coupons are specified. A kid becomes a search expert if same types of searches are performed continuously, for more than time T. The promising result (sieve coupons), are then given for review to parents and teachers involved in searching different areas for their kids and their opinion is taken. A questionnaire is designed to decide the priority and selection of the sieve coupons. After this prioritize the sieve coupons to be used in searches.

V. RESEARCH PROCESS

As the aim of the research is to provide secure and relevant data to the kid for their assignments, the research process started with a questionnaire that was filled by teachers of various schools to get an idea , whether they are using internet for educational work and are giving internet based assignments to the students or not. Based upon that Questionnaire an analysis is done on various assignments, holiday homework's and projects of various schools using Google. Various improvement terms are embedded with the search terms to improve the impact of relevant pages on search page. Various content filtering, quality factors have been taken into account while finalizing the sieve coupons that can be used to improve the result of relevant pages on search engines.

The questionnaire discussed above is given below. The questionnaire was filled by 35 teachers. According to this questionnaire 84% of the teachers use internet and give internet based assignments to their students. The main aim of teachers for giving internet based assignments is to make students aware of latest technologies and getting pace with the new trends in the market. The teachers who have filled this questionnaire teach students of age group 8 to 13 (i.e. class 3rd to class 8th).



Questionnaire 1
(To be filled by the teacher's of various schools)

Q1. Do you use internet aids to teach students of the class (Yes/No).

Q2. Do you find the relevant content while searching the syllabus topic? (Yes/No)

Q3. Do you give internet based assignments to students? (Yes/No)

Q4. Does students actually use internet for completing those assignments? (Yes/No)

Q5. Mention at least 4-5 topics that you have given to students to search on web.

_____

Q6. Is the content submitted by the student in response to assignment given by you is relevant. (Yes/No).

Q7. Is the content submitted by the student is age appropriate or not. (Yes/No)

Q8. Do you think students take more time searching the content of assignment as their search is not focused (i.e. they distract while searching the content on web). (Yes/No)

Q9. How effective these assignments are to enhance the student learning?
a) Not Effective b) Average c) Effective d) Very Effective

Q10. Do you think student is learning both good and bad things while searching the web? (Yes/No)

Q11. What kind of keywords you suggest the students for relevant searching of content.

_____

Name of the teacher: _____

Name of the school: _____

Fig 1: Questionnaire

After getting this information we gathered various school assignments, holiday homework's and tried to search the terms on Google using various keywords given by the teachers of various schools. Based upon the search term selected from various assignments and homework sheets of various schools, an analysis sheet is made which shows a comparison of results fetched with Google without keywords and with keywords used. The analysis sheet has shown a remarkable difference in the output when the search terms are combined with keywords. The analysis can be illustrated with the help of following chart.

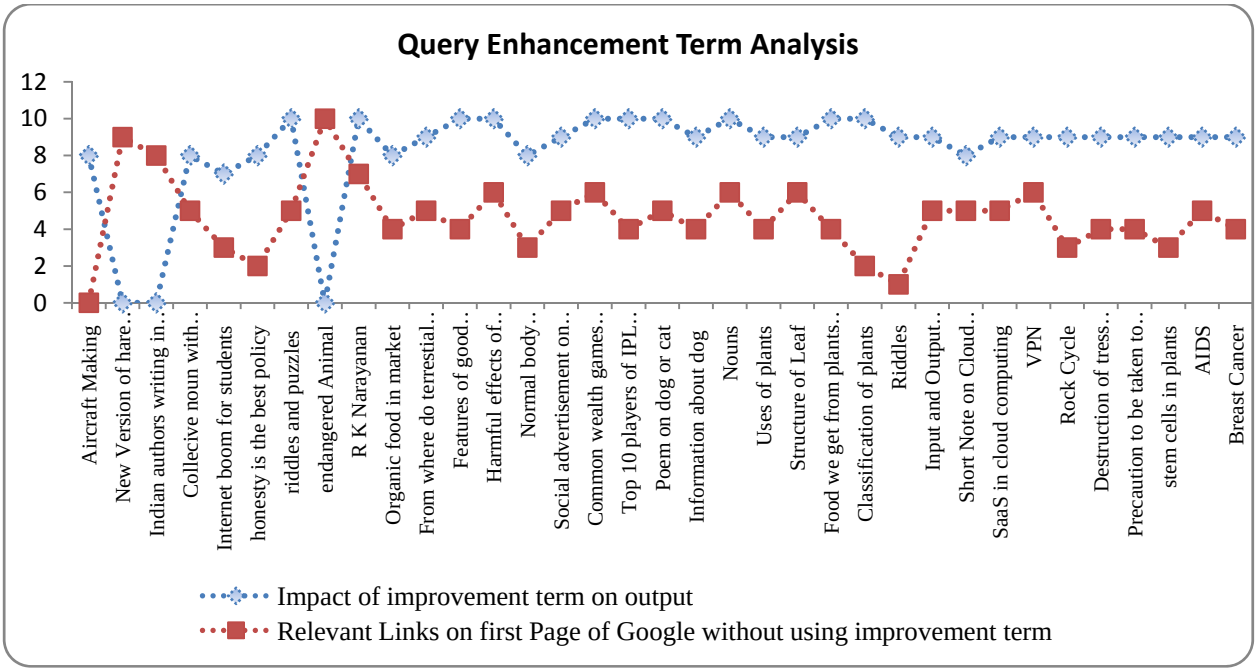


Fig 2: Chart showing impact of improvement term on the Search Term

The Fig 2 shows a remarkable improvement in relevant links on first page of web result after using improvement term. The blue dotted line shows the impact of improvement term appended with query search term. The red dotted line shows relevant links without using improvement terms. Though variation has been found on three terms like endangered animals, Indian authors writing in English and new version of hare and tortoise story, where improvement term is not required.

A. Various improvement terms used and their impact on search terms

The following are list of the keywords given by various teachers for searching the content on web.

[Kids, Children, Wikipedia, Encyclopedia, Britannica, Educational, School, Students, Notes]

The above keywords have been gathered from the questionnaire discussed above. Using these keywords we tried to search various terms and shown their result in the form of a excel sheet given below.

Impact of improvement terms on search terms(For Class 1 to 5)

	Suggessted Improvement term and their Impact out of 10								
Search Term	Kids	Children	Wikipedia	Encyclopedia	Britanica	Educational	School	Students	Notes
Aircraft Making	8	8	0	1	1	0	1	4	0
New Version of hare and tortoise story	10	10	10	10	6	10	8	7	4
Indian authors writing in English	10	10	9	9	6	9	6	6	3
Collecive noun with pictures	10	10	9	3	2	1	4	5	2
Internet boom for students	5	5	0	0	0	1	6	7	2
honesty is the best policy	8	5	0	2	2	1	1	2	0
riddles and puzzles	10	7	2	3	3	8	4	5	0
endangered Animal	9	9	8	9	7	8	4	7	7
Poem on dog or cat	10	8	3	3	2	5	2	1	1
Information about dog	9	7	5	4	4	4	3	2	2
Nouns	10	7	4	5	4	5	5	5	3
Uses of plants	9	7	7	6	5	3	3	1	1
Structure of Leaf	9	7	5	4	4	4	3	2	2
Food we get from plants and animals	10	6	6	5	4	5	4	3	3
Classification of plants	10	6	4	4	3	4	2	4	2
Riddles	9	8	2	2	2	5	5	3	2
Input and Output devices of computers	9	7	7	6	6	5	5	5	5

Fig 3: Excel sheet to show the search results for class 1 to 5

Figure 3 shows the search terms for class 1 to 5 and the impact of various improvement terms on search term. The analysis of this sheet can be illustrated with the help of following chart.

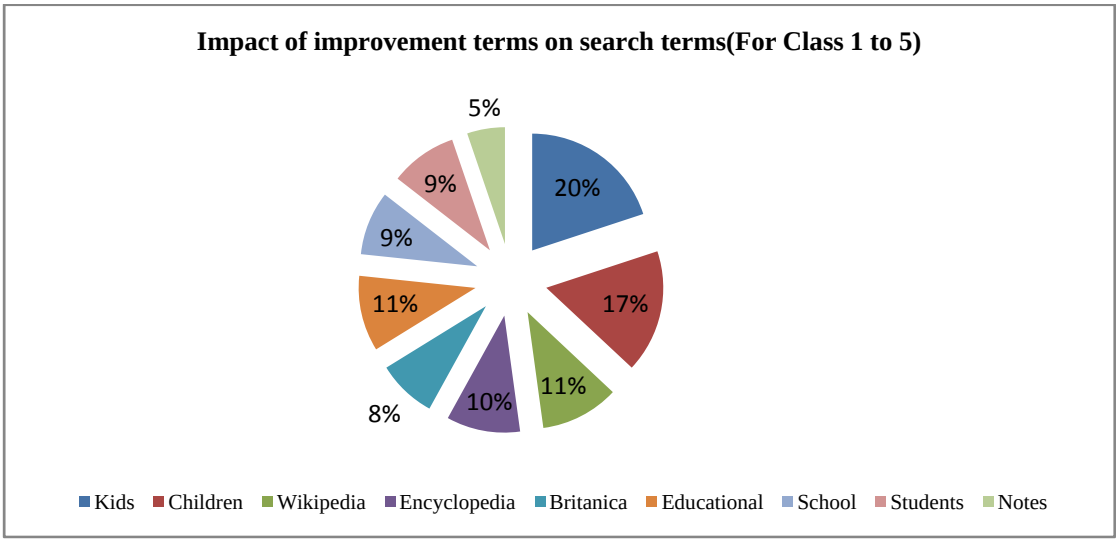


Figure 4: Illustration of Fig 3 with the help of a chart

The chart illustrated above used various improvement terms, but the best term that can be used as improvement term for students of class 1 to 5 is either kids or children. The above data collected focuses on various factors of content filtering discussed above.

While searching the content for students of class 1 to 5 various factors need to be considered like is the content age appropriate, is the data quality data , is the website secure, is data presentable and easy to understand and soon. Based upon all these factors it has been found that kid's keyword is the most appropriate term for searching the content for students of age 6 to 10. Another alternative can be the term children. All the above

mentioned terms like encyclopedia, Wikipedia etc are analyzed for all the above mentioned factors and a conclusion has been drawn about using the keyword kids for searching the content for kids'.
Further we tried to search the content for students of class 6th to 8th. Again various content filtering factors were kept in mind while searching the content and the analysis has been drawn with the help of analysis sheet and the chart given below. Again a detailed analysis has been done on search term by appending various improvement terms and the most appropriate term for students of class 6 to 8 is Wikipedia. Rest of terms can give results but one or the other content filtering factor violates. The near by improvement term can be encyclopedia with 14% results.

Impact of improvement terms on search terms(Class 6 to 8)

	Suggessted Improvement term and their Impact out of 10								
Search Term	Kids	Children	Wikipedia	Encyclopedia	Britanica	Educational	School	Students	Notes
R K Narayanan	9	9	10	6	5	4	3	6	3
Organic food in market	5	5	8	8	7	6	4	4	5
From where do terrestial animals obtain oxygen	6	5	9	6	6	6	4	5	4
Features of good respiratory surface	5	5	10	7	7	7	6	6	4
Harmful effects of smoking	4	3	10	4	5	5	4	3	3
Normal body temperature of domestic birds and animals	4	4	8	6	5	3	3	2	2
Social advertisement on equality	3	3	9	5	2	2	1	1	1
Common wealth games medals won by india 2010	3	3	10	6	5	4	3	2	2
Top 10 players of IPL match 2011	3	3	10	5	4	3	3	2	2
Short Note on Cloud Computing	3	2	8	5	4	4	2	4	3
SaaS in cloud computing	2	2	9	5	4	4	2	4	3
VPN	3	2	9	4	3	2	2	3	3
Rock Cycle	3	3	9	3	3	2	4	4	2
Destruction of tress effect the soil cover	2	3	9	5	4	4	2	3	1
Precaution to be taken to live in earthquake prone area	3	3	9	6	5	2	3	3	2
Stem cells in plants	2	2	9	5	4	3	3	2	2
AIDS	3	2	9	4	3	2	2	3	3
Breast Cancer	2	1	9	3	3	2	4	4	2

Fig 5: Excel sheet to show the search results for class 6 to 8

The result of the above data can be illustrated with the help of following charts:

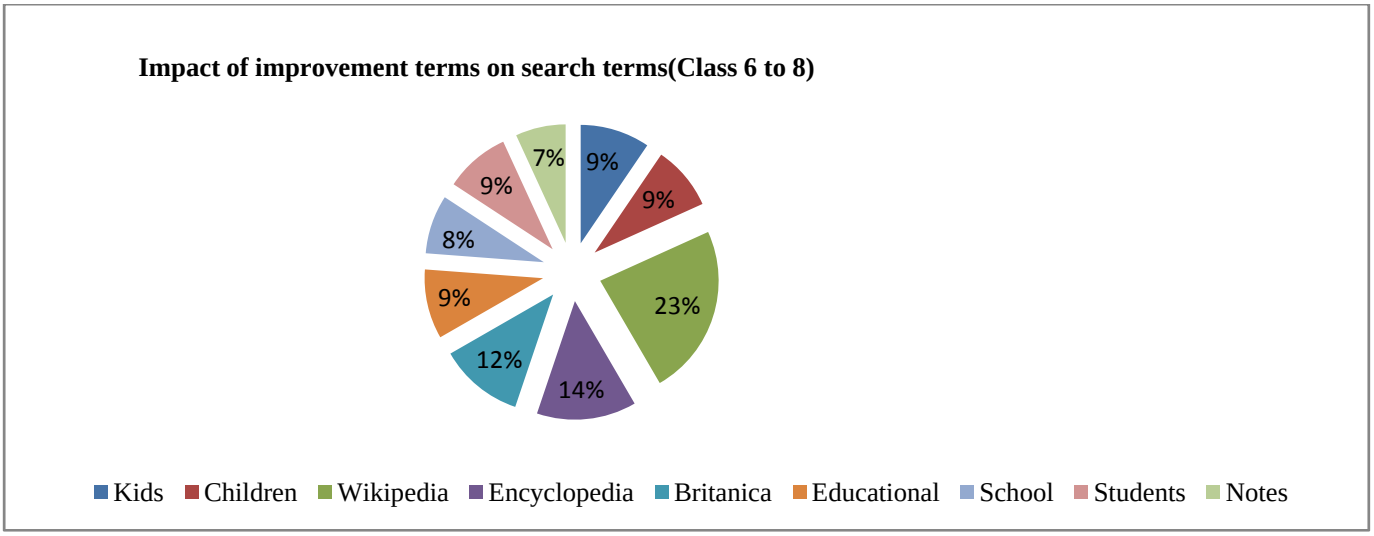


Figure 6: Illustration of Fig 4 with the help of a chart

B. Sieve coupons used and their impact analysis on the basis of various content filtering factors

From the above presented analysis sheet and the charts it has been found that the selection of the appropriate improvement term enhance 80-90% probability of relevant pages at a go. The selection of relevant term according to the age of the kid and the type of the content he/she is searching has introduced the concept of sieve coupons. Sieve coupons are the coupons that are appended with the search term which help in finding the relevant pages on the web, keeping content filtering factors as key to search the pages on web.

Various content filtering factors and their impact using sieve coupons can be illustrated with the help of following chart:-

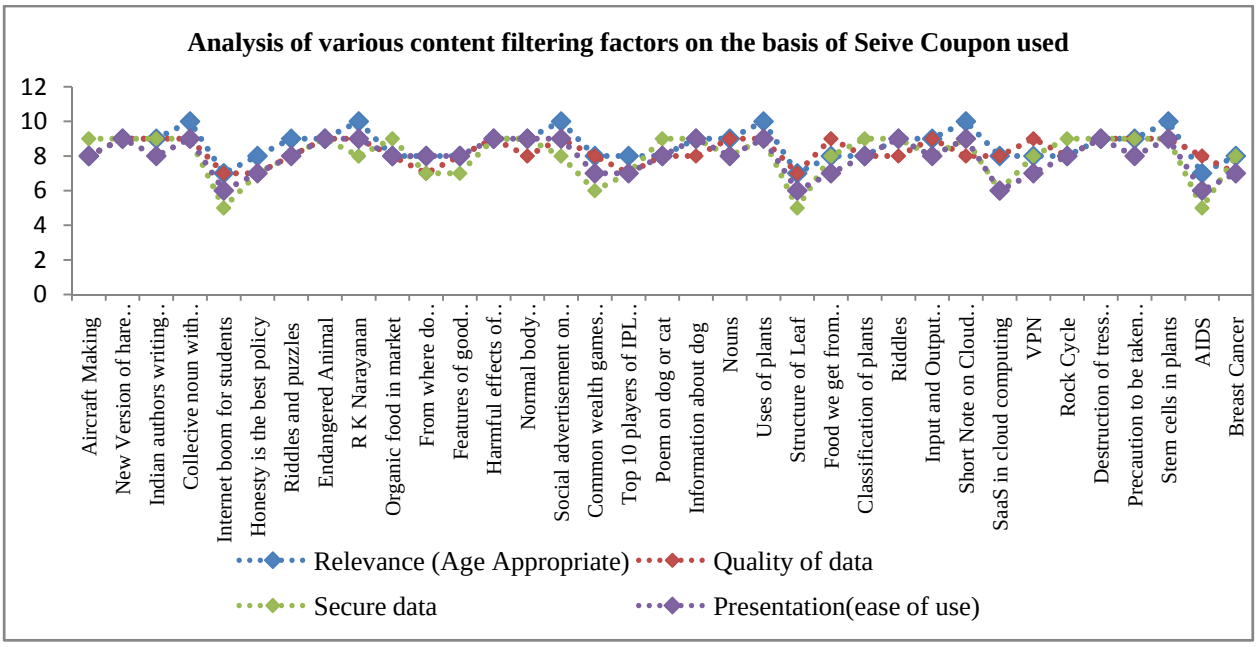


Figure 7: Illustration of various content filtering factors using sieve coupons discussed in Section III

VI. Conclusion

The paper presented above clearly depicts the role of Sieve coupons that can be used to enhance the quality and reliability of educational material searched by the child on the web. The experiment conducted evaluates Google results with the sieve coupons results and shows a remarkable improvement in the fetched results. Not only links are improved but the content filtering factors are also satisfied. The work presented above is only the part of the research conducted and have many other aspects that will be explored in due course of the research.

REFERENCES

[1] Saba Hilal, S. A. M. Rizvi, "The Syllabus Based Web Content Extractor (SBWCE)", Communications of the IIMA, Vol 8 issue 1, 2008

[2] Duarte Torres, Sergio and Hiemstra, Djoerd and Serdyukov, Pavel (2010) Query log analysis in the context of information retrieval for children. In: 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, July 19-23, 2010, Geneva, Switzerland.

[3] Jochmann-Mannak, H.E. and Huibers, T.W.C. and Sanders, T.J.M. (2008) Children's information retrieval: beyond examining search strategies and interfaces. In: The 2nd BCS-IRSG Symposium: Future Directions in Information Access, 22 Sept 2008, London.

[4] S.Duarte,"Information retrieval for children based on aggregate search paradigm", University of Twente (Internal Report), 2011.

[5] Duarte Torres, Sergio and Hiemstra, Djoerd and Serdyukov, Pavel (2010)an An analysis of query intented to retrieve information for children. In: 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, June 13, 2011, Geneva, Switzerland.

[6] Eickhoff, Carsten and Polajnar, Tamara and Gyllstrom, Karl and Duarte Torres, Sergio and Glassey, Richard (2011) Web Search Query Assistance Functionality for Young Audiences. In: 33rd European Conference on Information Retrieval, ECIR 2011, 18-21 April 2011, Dublin, Ireland.

[7] Duarte Torres, Sergio and Weber, Ingmar (2011) What and how children search on the web. In: 20th ACM International Conference on Information and Knowledge Management, CIKM 2011, 24-28 October 2011, Glasgow, UK.

[8] Sandra G. Hirsh, "How Do Children Find Information on Different Types of Tasks", Children's Use of the Science Library Catalog. Library Trends 45(4): (1997)

[9] Carsten Eickhoff et.al, "A combined topical/non-topical approach to identifying web sites for children", WSDM '11 Proceedings of the fourth ACM international conference on Web search and data mining Pages 505-514, 2011

[10] Carsten Eickhoff Et.al, " Web page classification child suitability", CIKM '10 Proceedings of the 19th ACM international conference on Information and knowledge management Pages 1425-1428,2010

[11] Hauff, C. and Trieschnigg, R.B., "Enhancing Access to Classic Children's Literature." BooksOnline'10 Workshop at CIKM 2010, 26 Oct 2010

[12] R. Glassey, L. Azzopardi, D. Elliott and T. Polajnar., "Interaction-based Information Filtering for Children", In Proceedings of the 3rd Information Interaction in Context Symposium (IIiX '10) New Brunswick, NJ, USA, 2010.

[13] Karl Gyllstrom, Marie-Francine Moens, "Clash of the Typings: Finding controversies and children's topics within queries", ECIR'11 Proceedings of the 33rd European conference on Advances in information retrieval, pg 80-91, 2011

[14] Jochmann-Mannak, H., Huibers, T., Lentz, L., Sanders, T. (2010) Children searching information on the Internet: Performance on children's interfaces compared to Google. In Proceedings of the Workshop on Accessible Search Systems, SIGIR '10, July 23, 2010

[15] Carsten Eickhoff, Arjen P. de Vries, Identifying Suitable YouTube Videos for Children, Proceedings of the 3rd Networked & Electronic Media Summit (NEM),2010

[16] M.G. van Kalsbeek , J.J. de Wit, "Automatic Reformulation of Children's Search Queries", University of Twente (Internal Report), 2010.