FOSTERING THE RETRIEVAL OF SUITABLE WEB RESOURCES IN RESPONSE TO CHILDREN'S EDUCATIONAL SEARCH TASKS

by

Oghenemaro Deborah Anuyah

A thesis

submitted in partial fulfillment

of the requirements for the degree of

Master of Science in Computer Science

Boise State University

August 2018

© 2018
Oghenemaro Deborah Anuyah
ALL RIGHTS RESERVED

BOISE STATE UNIVERSITY GRADUATE COLLEGE

DEFENSE COMMITTEE AND FINAL READING APPROVALS

of the thesis submitted by

Oghenemaro Deborah Anuyah

Thesis Title: Fostering the Retrieval of Suitable Web Resources In Response to Children's Educational Search Tasks

Date of Final Oral Examination: 9th July 2018

The following individuals read and discussed the thesis submitted by student Oghenemaro Deborah Anuyah, and they evaluated the presentation and response to questions during the final oral examination. They found that the student passed the final oral examination.

Maria Soledad Pera, Ph.D.	Chair, Supervisory Committee
Casey Kennington, Ph.D.	Member, Supervisory Committee
Jerry Alan Fails, Ph.D.	Member, Supervisory Committee
Michael Ekstrand, Ph.D.	Member, Supervisory Committee

The final reading approval of the thesis was granted by Maria Soledad Pera, Ph.D., Chair of the Supervisory Committee. The thesis was approved by the Graduate College.

To everyone that have said to me “Maro, I believe you can make it through”, I dedicate this thesis to you.

ACKNOWLEDGMENTS

I would like to express my utmost gratitude to my advisor, Dr. Maria Soledad Pera, for granting me the opportunity to work in her lab and benefit from her rich knowledge in Information Retrieval. Thank you for your mentorship, encouragement, patience, and guidance throughout the research process. This work would also not have been possible without the guidance and feedback from my committee members: Dr. Casey Kennington, Dr. Jerry Alan Fails, and Dr. Michael Ekstrand.

To everyone that supported me throughout my Masters degree program, I extend my sincere gratitude. First, to my parents, Mr. and Mrs. Anuyah, and my sister, Fejiro Anuyah. I would not have made it this far if not for your love and prayers. God bless you! I would like to thank Akpofure for all of his support and encouragement right from the first day I started my degree.

I would also like to appreciate my co-lab member and friend, Ion Madrazo Azpiazu, for his constant encouragement and guidance as this thesis work expanded in new directions. To Emily Suchan, words are not enough to express how grateful I am for all your love, prayers, and emotional support during the course of my thesis. I would also not forget to appreciate my friends Rezvan Joshaghani, Aprajita Shukla, and Sanaz Alamian, you ladies are the best! Thank you for making sure I was always in high spirits and for making my Master's program fun.

Lastly, I would like to appreciate the Department of Computer Science for giving me the opportunity to benefit from the great teaching, assistantship, and support throughout my program. To all professors that extended their rich knowledge to me

through their teaching, I appreciate you.

This work has been funded by NSF grant with Award number: 1565937.

ABSTRACT

Children regularly turn to search engines (SEs) to locate school-related materials. Unfortunately, research has shown that when utilizing SEs, children do not always access resources that specifically target them. To support children, popular and child-oriented SEs make available a safe search filter, which is meant to eliminate inappropriate resources. Safe search is, however, not always the perfect deterrent as pornographic and hate-based resources may slip through the filter, while resources relevant to an educational search context may be misconstrued and filtered out. Moreover, filtering inappropriate resources in response to children's searches is just one perspective to consider in offering them the right resources, as aspects that are key for this audience are overlooked, including reading level, resource subjectivity, or the context of the search (i.e., educational setting). To verify impediments of existing SEs in response to children's searches conducted at school, we conduct several empirical studies on well known SEs: Google, Bing, their safe search counterparts, Kidrex and Kidzsearch. Based on our findings, we present **K**i**S**u**RF**, a novel filtering and ranking strategy that not only eliminates inappropriate resources while retaining education-relevant ones, but also simultaneously examines multiple qualitative aspects of online resources in order to offer suitable ones. Empirical studies conducted using diverse datasets, including one comprised of children's search sessions in the school setting, showcase (i) the usefulness of simultaneously integrating evidences from multiple perspectives in order to inform resource suitability detection, and (ii) the correctness and effectiveness of **K**i**S**u**RF** in prioritizing child-suitable resources.

TABLE OF CONTENTS

ABSTRACT vii

LIST OF TABLES xii

LIST OF FIGURES xiv

LIST OF ABBREVIATIONS xvi

1 Introduction 1

2 Background and Related Work 9

2.1 Children's use of Search Engines in the Classroom 9

2.2 Safe Search 10

2.3 Personalization of Web Resources 11

2.3.1 Readability 12

2.3.2 Educational Relevance 13

2.3.3 Objectivity 14

2.3.4 Target Audience 15

2.4 Ranking Web Resources 15

3 Examining the Limitations of Filtering and Ranking Strategies . . . 17

3.1 Search Engines 17

3.2 Data Sources for Empirical Analysis 18

3.3 Empirical Analysis and Discussion 20
3.3.1 Are SEs too restrictive when it comes to locating resources in response to educational information discovery tasks? 20
3.3.2 Does safe search always disregard sexually explicit resources? . . 21
3.3.3 Are safe search filters effective in disregarding resources that potentially promote violence? 23
3.3.4 Do SEs retrieve resources that match the reading abilities of school-aged users? 24
3.3.5 Are educational resources ranked higher for education-related searches? 25
3.3.6 Do SEs prioritize specific web domains? 27
3.4 Lessons Learned 28

4 KiSuRF: A Multi-Perspective Strategy for Offering Suitable Web Resources to Children 30
4.1 Programming Language and Tools 30
4.2 Text Extraction and Processing 31
4.3 Design Overview of **KiSuRF** 33
4.4 Data Sources used in the Design and Development of **KiSuRF** 34
4.5 Eliminate Inappropriate Resources 35
4.5.1 Quantify the Degree of Sexual Explicitness in Resources 36
4.5.2 Determine the Degree to which Resources Promote Violence . . . 38
4.5.3 Retain Education-relevant Resources 39
4.5.4 Aggregate Data Points for Filtering 41
4.6 Prioritize Suitable Resources 43

4.6.1 Estimate the Reading Level of Resources 43
4.6.2 Quantify the Degree of Objectivity in Resources 45
4.6.3 Quantify the Degree to which Resources are Educational 46
4.7 Integrate Qualitative Aspects 56

5 Evaluation 59
5.1 Experimental Setup 59
5.1.1 Evaluation Strategy 59
5.1.2 Metrics 60
5.1.3 Test of Statistical Significance 63
5.2 Examine Correctness of **KiSuRF** 64
5.2.1 Which readability formula performs best for estimating the reading levels of web resources? 64
5.2.2 Can **KiSuRF** identify objective resources? 66
5.2.3 Do topics and concepts in $\mathbf{Edu_{KB}}$ abide by pre-defined contextual similarities? 69
5.2.4 Is **KiSuRF** able to identify relevant educational topics in resources? 70
5.2.5 Are educational resources assigned higher educational value scores when compared to non-educational resources? 73
5.2.6 Which model is best suited to aggregate data points for **KiSuRF**'s filtering purposes? 75
5.2.7 Which model is best suited for simultaneously integrating aspects considered for ranking? 76
5.3 Overall Performance Analysis of **KiSuRF** 79

5.3.1 How does **KiSuRF**'s filtering strategy perform when compared to safe search filters on SEs under study? 80
5.3.2 Is considering multiple perspectives essential for prioritizing suitable resources? 82
5.3.3 How is the overall performance of **KiSuRF** in handling children education-related searches? 83

6 Conclusions and Future Work 88
6.1 Applicability 90
6.2 Future Directions 90

REFERENCES 93

A Framework for Archiving Children's Search Sessions 108

B Summary of Datasets 112

LIST OF TABLES

3.1 Percentage of searches initiated with queries in $\mathbf{Ed_{kw}}$ that led to no results. 21

3.2 Assessment of SEs in responding to searches formulated for accessing sexually explicit content. 22

3.3 Assessment of SEs in handling searches conducted with the aim of accessing violence-related content. 24

3.4 Average grade levels computed for top–10 results retrieved for queries in $\mathbf{K_{qry}}$. 25

3.5 Assessment of the effectiveness of SEs in handling educational searches initiated using queries in $\mathbf{E_{qry}}$. 27

3.6 Average ranking assigned to resources from different domains for searches initiated using queries in $\mathbf{K_{qry}}$. 28

4.1 Example of text processing operations performed on a sample sentence extracted from the children's book "Alice's Adventures in Wonderland" written by Lewis Carroll, where NN–noun, VBD–verb in the past participle, VBG–verb in the present participle, JJ-adjective, and PRP–pronoun. 32

4.2 Example of different variations of the term "stay" and their corresponding stem. 33

4.3 Overview of **KiSuRF**'s filtering data points. 43

4.4 Overview of **KiSuRF**'s ranking data points. 57

5.1 Summary of the datasets used in validating the effectiveness and correctness of **KiSuRF**. 60

5.2 Comparison of the correctness of traditional readability formulas in estimating the reading level of resources, using the RMSE. 66

5.3 **KiSuRF**'s performance in extracting relevant educational topics from resources. 72

5.4 Performance of different models considered in the design of **KiSuRF**'s filtering strategy. 76

5.5 Resource distribution on $\textbf{Edu\_Search}_{\textbf{DS\_A}}$ and $\textbf{Edu\_Search}_{\textbf{DS\_B}}$. 77

5.6 Assessment of **KiSuRF** in complementing the filtering functionality of SEs. N.A. indicates that the SE under study is not applicable for a particular assessment. Information between square brackets, i.e., "[]", indicates initial results obtained in the empirical analysis conducted in Chapter 3. 81

B.1 Summary of datasets described in Chapter 3. 113

B.2 Summary of datasets described in Chapter 4. 113

B.3 Summary of datasets described in Chapter 5. 114

LIST OF FIGURES

1.2 Kizsearch retrieves no result for the query "breast tissue". 5

2.1 Resources retrieved for the query "coco movie". 11

3.1 Examples that showcase limitations of Kidrex and Kidzsearch in handing queries with terms relevant to children's educational health and science subjects. 21

4.1 Architecture of **KiSuRF**, which complements existing SEs by filtering resources with inappropriate content and prioritizing the ones that are suitable to a child in terms of readability, objectivity, and educational value. 34

4.2 Example of a hate-based content embedded in the meta tag of **NiggerMania**, a site known to promote hate-speech. 39

4.3 Cumulative scree plot showing the number of principal components suitable for reducing the dimension of TF-IDF vectors used in representing resources. 41

4.4 Correlation among analyzed data points in **KiSuRF**'s filtering strategy. Color represents the polarity of correlation, and bubble size indicates correlation strength. 44

4.5 Sample of pairs with *hasA* and *sameAs* relation type in $\mathbf{Edu_{KB}}$. 50

4.6 A description of Wikipedia category hierarchy [141]. 50

4.7 A sample of *hasA* relations among K–12 topics, concepts, and grade levels defined in $\mathbf{Edu_{KB}}$ for the topic "Geometry". 52
4.8 An example of educational topics extracted from a sample resource. . . . 54

5.1 Comparison of objectivity scoring strategy on different datasets. 69
5.2 Pattern of contextual similarity among pairs in $\mathbf{Edu_{KB}}$. 70
5.3 Sample of a grouping which shows contextually similar educational topics in $\mathbf{Edu_{KB}}$. 71
5.4 Comparison of educational value scores assigned to diverse information sources. Educational resources had the highest value. 74
5.5 Performance of examined regression models used in aggregating data points for ranking suitable resources. GBT: Gradient Boosting, LINEAR: Linear, MLP: MultiLayer Perception, and SVM: Support Vectors Machine. 79
5.6 Comparison of ranking performance when different aspects of suitability are considered for prioritizing resources. OBJ–Objectivity, RD–Readability, EDV–Education Value, and ALL–a combination of Objectivity, Readability, and Education Value. 84
5.7 Overall performance of **KiSuRF** in prioritizing suitable resources. 86

A.1 Interface for a child searching to indicate his grade level. 110
A.2 Interface where a child initiates his search and selects retrieved resource of interest. 110
A.3 Sample of a child's search session. 111

LIST OF ABBREVIATIONS

CCSS – Common Core State Standards

DMOZ – Directory.Mozilla.Org

ICS – Idaho Content Standards

IDLA – Idaho Digital Learning

K-12 – Kindergarten to 12^{th} grade

MRR – Mean Reciprocal Rank

NDCG – Normalized Discounted Cumulative Gain

NGSS – Next Generation Science Standards

NLP – Natural Language Processing

ODP – Open Directory Project

PoS – Part of Speech

P@K – Precision at K

RMSE – Root Mean Squared Error

SE – Search Engine

CHAPTER 1

INTRODUCTION

Popular **Search Engines** (**SEs**) like Google [65], Yahoo [161] and Bing [31] facilitate access to web resources. In response to a recent survey, more than half of Americans indicated that they turned to SEs at least once a day, with a majority reporting their preference for Google [77]. With **children** being introduced to the web at increasingly young ages [73, 118], the use of SEs is not limited to mature audiences.

According to Rowlands et al. [133], young children turn to SEs daily as their first "port of call for knowledge". Children's use of SEs goes beyond accessing online gaming, social networking and other sites that interest them, as they also utilize these tools for **school work**. In the USA, schools now regularly use SEs [84], as teachers assign information discovery tasks to their students both within and outside the classroom [138]. Early exposure to SEs can help children build foundational skills crucial in a knowledge-rich society [103]. However, search literacy is not always part of the K–12 curriculum [140] and popular SEs are not always equipped to guide children's searches.

Research has shown that children's use of SEs is different from adults, the population for whom SEs were designed [68, 70]. Due to their unique **search behaviors**, children cannot always complete successful searches. In addition to problems that may stem from poor query formulation, as children are known to struggle with it

[113, 149], notable challenges are attributed to children identifying resources that are relevant to their search. Unlike adults, who select retrieved resources in an unordered style as they can identify relevant resources by examining snippets, children's selection style is sequential, i.e., they scan from the top to the bottom of the result list [68]. Additionally, children favor resources on the first page, and barely notice that other result pages exist [50, 74]. Children's preference for linear resource selection, coupled with the fact that they may not have sufficient skills to analyze and identify relevant resources when using their favorite SEs [28, 70, 154], make it essential that resources relevant to their search are positioned higher in the ranking.

Another problem related to unsuccessful searches is the inclusion of resources in the result set that are not necessarily relevant. On one hand, this can be due to the inclusion of **ambiguous** terms in a query, which naturally leads to resources pertaining to diverse topics; not all of them targeting the information needs of children. For example, consider a child looking for a popular game using the query "sand castle". Among the top–10 resources retrieved by Google (shown in Figure 1.1), we find resources that refer to a water park, a movie, and a restaurant, positioning the resource that best reflects the query intent lower in the ranking. On the other hand, irrelevance can be attributed to lack of **comprehension**. Resources may not align with children's reading skills and developmental levels, making it difficult for children to understand their content. This is important to address, as children's experiences with SEs can affect their motivation to use the web, their skill to adequately use resources for their personal and educational interests, and their exposure to information beneficial for enhancing their mental capability [59].

We argue that to better serve children, it is imperative that SEs consider the purpose of a search, in addition to explicitly addressing resource suitability, and thus

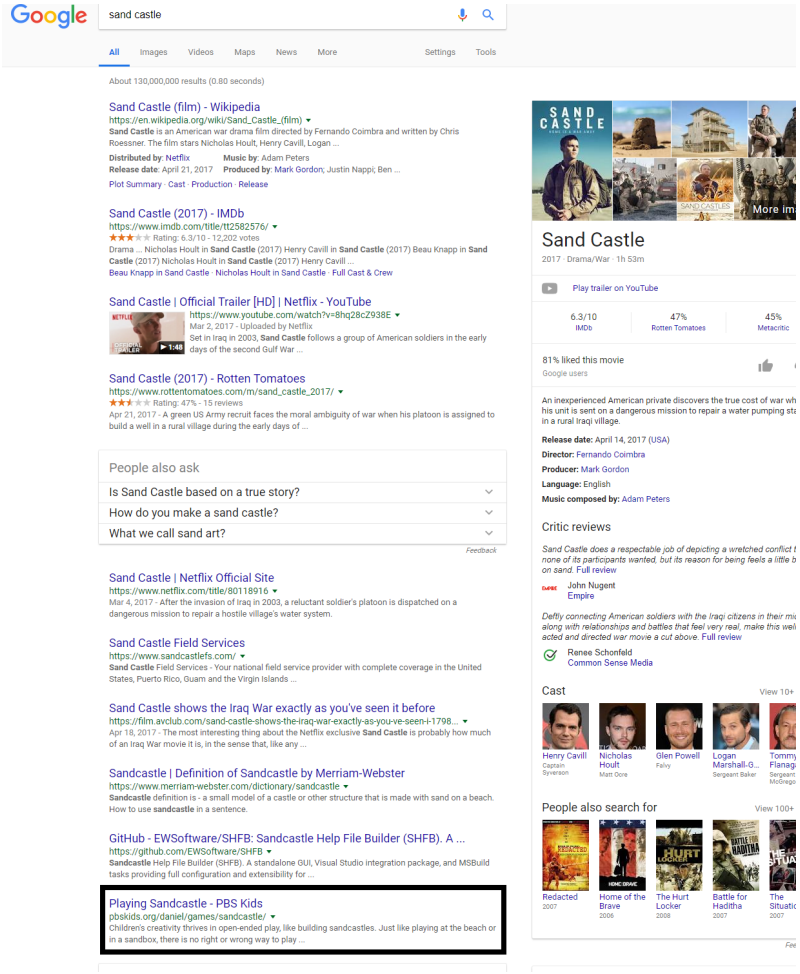


Figure 1.1: Screenshot of Google's result page for the query "sand castle". The black rectangle indicates a resource matching the query intent.[1]

prioritize resources that are relevant to a child's search context. In this thesis, we take resource **suitability** to be comprehensive, considering the fact that it differs per user and context. We treat as suitable resources that include appropriate contents, i.e., do not contain pornography or hate-speech content; a child can comprehend based on his / her reading skills; are objective; and are relevant for education-related searches. Examining the suitability of retrieved resources is not a trivial task. Differentiating

[1]Screenshot generated for the query "sand castle" in June 2018.

content relevant to a search context versus content relevant to a query is a common challenge in Information Retrieval [63, 160] which is yet to be solved from a child's perspective. For example, for the query "frozen", Google treats a Wikipedia page as the most relevant result: while the retrieved content matches the query term, it cannot be comprehended by a young child due to content complexity.

In order to offer a better search experience for children, popular SEs make available a *safe search* filter [42] that is meant to disregard resources with inappropriate contents, such as violence-related and sexually explicit materials. Unfortunately, in some scenarios, safe search may be too restrictive: it might filter resources that are adequate to an educational context but happen to include terms that might be misconstrued as unsafe. For example, Kidzsearch's safe search interprets searching for resources pertaining to "breast tissue" as inappropriate, and therefore does not retrieve any results (see Figure 1.2). This can be problematic when children are given school research assignments related to the human anatomy subject, as this would prevent them from locating the right resources. Moreover, some resources considered unbefitting for children slip through the filter. Alternatively, there are SEs designed exclusively for children which adopt safe search and aim at addressing problems children might encounter when conducting online searches. Among these child-oriented SEs, popular ones include Kidzsearch [98], Kidrex [97], Kiddle [96], and Yahooligans (now Yahoo! kids). Unfortunately, the same aforementioned safe search issues affect these child-oriented SEs. Additionally, in some cases, these SEs manually select resources to be indexed [55, 75], limiting the amount of resources they make available. Even though SEs explicitly designed for children may improve their search experience [90, 93], children still prefer to use popular ones [59]. Thus, it is imperative that popular SEs provide resources that are suitable in response to a

child's query [72].

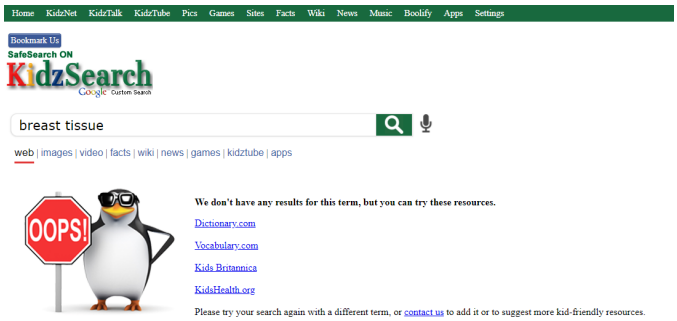


Figure 1.2: Kizsearch retrieves no result for the query "breast tissue".

Research focused on offering suitable web resources in response to a query is not novel. However, efforts dedicated for this purpose when it comes to children is limited, as it only addresses either resource readability levels [39] or inappropriate content [14, 16, 54, 88], thus not offering a comprehensive solution. For instance, if used in the classroom setting, SEs such as Google and Kidzsearch filter resources using their safe search options [24], but this means that only those resources containing sexually explicit or violence-related content are disregarded [66]. Yet, resource comprehension, degree of objectivity, and educational pertinence is not ensured.

In this thesis, we limit our scope to children in the **3rd - 5th** grades, as children within this range are known to exhibit similar search traits [59]. As opposed to leisure-related searches which cover a broader scope of topics, we focus on the retrieval of resources in response to information discovery tasks in the **classroom setting**. We first discuss the **empirical analysis** conducted on a number of SEs to understand how they fare in preventing children from accessing inappropriate resources, as well

as in prioritizing resources in response to queries that are meant to satisfy educational information needs. In response to the findings, we designed **K**i**S**u**RF**–**K**ids **S**uitable Resource **R**anker and **F**ilter. **K**i**S**u**RF** is a novel filtering and ranking strategy meant to complement the retrieval functionality of popular SEs. **K**i**S**u**RF** goes beyond traditional safe search by examining retrieved resources from multiple perspectives in order to quantify the degree to which their content is unsafe for children, retaining those that include terminologies common among children's school subjects, and prioritizing suitable ones.

In designing **K**i**S**u**RF**'s filtering strategy, we explore several *qualitative aspects* in order to determine if resources should indeed be filtered. For **K**i**S**u**RF**'s ranking strategy, we consider three perspectives for resource analysis: *Readability*, *Objectivity*, and *Educational Pertinence*. By considering the readability of resources, **K**i**S**u**RF** ensures that resources retrieved are of the reading level expected for children in the 3^{rd} – 5^{th} grades. In examining the degree of objectivity in retrieved resources, **K**i**S**u**RF** is able to prioritize those that are non-opinionated, as children may not have sufficient skills to judge the quality of resources [28, 154]. By quantifying the education relevance of resources, **K**i**S**u**RF** ensures that children access resources that respond to educational standards, such as Common Core State Standards [45], Next Generation Science Standards [44], and Idaho Content Standards [85], which outline educational topics defined for children in the K–12 grade levels. During the research process, we found that there is no existing information source such as knowledge bases, taxonomies, or ontologies, that allow us identify specific concepts and their relatedness in the educational domain. To respond to this limitation, we take advantage of a deep learning based architecture and create a novel education-domain knowledge base that captures relations among educational concepts and topics at different grade levels.

In order to validate the correctness and effectiveness of **KiSuRF**, we conducted an in-depth study. We performed several experiments to evaluate each perspective considered in the design of **KiSuRF**. To ensure that **KiSuRF** is able to eliminate inappropriate resources while retaining those with educational terminologies, we compared **KiSuRF**'s filtering strategy with safe search filters available on popular SEs, as well as child-oriented ones. We also demonstrate the overall effectiveness of **KiSuRF**'s ranking strategy by examining the rank assigned to resources known to be suitable for children's educational search tasks with and without using **KiSuRF** to complement the retrieval functionalities of popular SEs.

The main **contributions** of this thesis are fourfold:

- Limitations and lessons learned as a result of an **empirical analysis** conducted on selected SEs in terms of handing searches conducted by children in an educational setting.
- A **filtering and ranking strategy** that responds to the limitations found as a result of our empirical analysis, as well as compliments SE retrieval functionality by simultaneously considering multiple perspectives in order to offer children suitable web resources.
- A number of new datasets, including one comprised of **children's search sessions**, which can be of benefit to the research community as one primarily for children is not readily accessible.
- A novel **education-domain knowledge base** that includes K–12 educational topics and concepts relations at different levels, which is also a benefit to the research community as, to the best of our understanding, a knowledge base specific to the educational-domain does not exist.

The rest of this manuscript is organized as follows: In Chapter 2, we discuss background literature pertaining to children's use of SEs in the classroom setting, safe search filters, and related work on personalization and ranking strategies. In Chapter 3, we present the results of the empirical analysis we conducted to identify limitations of existing SEs when filtering and ranking resources in response to children's queries. Thereafter, in Chapters 4 and 5, we introduce **K**i**S**u**RF**, along with the experimental analysis conducted to validate its performance. Lastly, in Chapter 6, we offer some concluding remarks and directions for future work.

CHAPTER 2

BACKGROUND AND RELATED WORK

In this section, we provide an in-depth discussion regarding children's use of SEs in the school environment, as well as safe search filters. We also present related work on diverse aspects of personalization and ranking strategies.

2.1 Children's use of Search Engines in the Classroom

The use of SEs is now becoming a "valuable asset for children's education, as it encourages learning, enhances the class environment, and introduces children in the early stages of their lives to today's information society" [20]. In schools, SEs are often used by children to locate information for educational tasks: from looking up the meaning of words and finding math formulas, to addressing history-related inquiries for which it usually takes longer to find answers for when using printed books [101]. Unfortunately, children are known to fail to complete successful searches [67], which could lead to frustration and discourage them to continue their engagement with search tools [59]. The lack of search success is in part correlated to their insufficient skills in formulating effective queries or in identifying relevant resources [28, 70, 154]. Search literacy could help with the development of these skills, but unfortunately, it is not always part of the K–12 curriculum [140]. The use of SEs designed exclusively for children could also help them complete successful searches and thus improve their

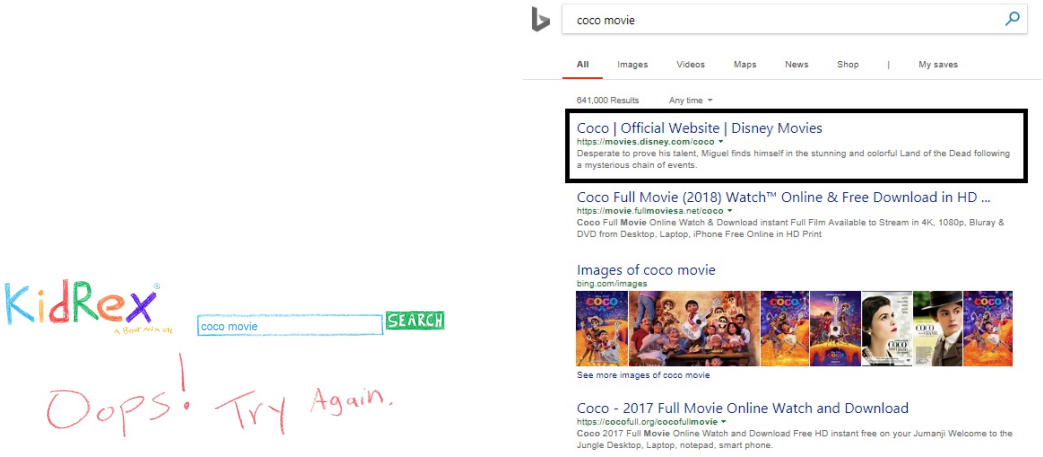
overall search experience. Yet, children still prefer to use popular SEs [59]. In fact, in a recent survey, 94% of participants who were teachers reveal that their students are likely to use Google for school assignments, as opposed to child-specific SEs [128]. This serves as an indication that children should either be provided with some form of assistance whenever they utilize their preferred SEs [59] or that instead mainstream SEs should be adapted so that they can offer children a better experience. This adaptation can be in the form of guidance for query creation or prioritization of resources that address their information needs in the educational context, which happens to be the focus of this thesis.

2.2 Safe Search

In order to prevent children from accessing inappropriate content when locating information of interest on the web, popular SEs along with their child-oriented counterparts, adopt safe search functionality [5, 7, 13]. Safe search is meant to filter resources with inappropriate content, such as pornography and hate-speech [7, 87], hence, offering a better and safer search environment. Traditional safe search filters may, however, be limited to blacklisted terms and URLs, which can be challenging to update: while blacklisted URLs may be ineffective as they keep changing, blacklisted terms may be deterred by "homographs."[1] Moreover, safe search is not always the perfect deterrent [51]. On the one hand, resources with inappropriate content pass through the filter, a fact we empirically verified in Chapter 3. On the other hand, safe search has been known to be too strict and filter resources that are relevant to users' information needs and the context of their searches. For instance, as of June

[1]Homographs are words that have the same spelling but may not be pronounced similarly and have different meanings.

2018, Kidrex's safe search filter prevents the retrieval of any resources for the query "coco movie". This should not be the case, especially given the number of available resources that are related to this particular information need and are suitable for children (see Figure 2.1).



(a) Kidrex. (b) Bing.

Figure 2.1: Resources retrieved for the query "coco movie".

2.3 Personalization of Web Resources

Personalizing the retrieval of resources in order to satisfy diverse user needs and their preferences is a challenging task, as evidenced by the body of work in this area [15, 30, 34, 38, 89, 99]. This typically requires researchers to analyze implicit feedback or user behavioral information, which depends upon the existence or inference of a user profile, that may include age or previous search sessions. Obtaining this information in the case of our target audience is difficult, as children's online privacy rules, e.g., the Children's Online Privacy Protection Act (COPPA) [61], Family Educational Rights and Privacy Act (FERPA)[150], and General Data Protection Regulation (GDRP) [57], prevent archiving identifiable information. Still, we argue that there

are a number of perspectives that can be explored to improve tailoring of resources so that they match children's needs and search context, even if that cannot be done on an individual basis. We discuss some attempts to consider such aspects in the quest for improving search tasks for children.

2.3.1 Readability

Results of an in-depth analysis of children's search sessions on Google demonstrate that most resources retrieved on the web have reading levels that target mature readers [26]. Based on a recent survey, Bilal and Boehm [27] identify that out of 300 results retrieved in response to queries written by seventh graders, only one match their reading level. These findings were further verified by Vajjala et al. [151] who compute the readability of the top–10 retrieved results on Google for 50 test queries and observe that the reading levels of these resources are relatively high [151]. These assessments offer supporting evidence for the need to retrieve resources that match children's reading levels.

Personalization based on reading levels has received attention in recent years [29, 40, 53, 147]. To compute text complexity of resources, existing works rely on html-based features [40]; which are known to offer limited perspective on the resource level. Researchers also take advantage of traditional readability formulas [29, 40, 53, 147]. Among popular readability formulas, we find Flesch reading ease [58], which examines shallow features such as the average length of words and sentences in a resource in order to assert its difficulty. This formula was enhanced in order to align text difficulty with grade levels in the well-known Flesch Kincaid formula [100]. Several other formulas, including Spache [145] and Dale-Chall [35], also utilize shallow features in addition to a list of keywords to estimate the complexity of a

resource. Most of these formulas, however, lack accuracy, especially in dealing with web resources [21, 142].

2.3.2 Educational Relevance

Traditionally, research on identifying documents relevant to a domain rely on techniques based on statistical models [134], decision trees [130], or support vector machines [159]. Advancements in this research area have led to new approaches that depend upon information sources such as ontologies [111, 112] and knowledge bases [139, 156], as these capture conceptual information in documents. Popular ontologies used include DBPedia [105] and Medical Subject Headings-MESH [82], whereas popular knowledge bases include Probase [158] and Freebase [33]. Neither of these sources target the K–12 audience, nor are known to directly map to educational standards, e.g., Common Core State Standards and Next Generation Science Standards. Furthermore, to the best of our knowledge, a knowledge base designed particularly for the educational domain is non-existent.

Strategies on automatically identifying the relevance of a given document to the educational domain has received limited attention from the research community [135, 159]. The authors in [159] outline particularities of educational materials and use a SVM classifier trained on TF-IDF weights extracted from labeled educational documents, to predict if a document is educational. Instead, the authors in [135] examine domain terms common among resources categorized for K–12 students and use affinity mitigation, a clustering-based-approach. The aforementioned strategies, however, center on analyzing individual terms in documents, which can fail to capture conceptual or semantic meanings [107]. Examining the latter can influence the process of correctly determining what a document as a whole is about and the degree to which

its content is educational. Moreover, classification of resources as non-educational is not applicable in the context of ranking of resources in response to a query (the focus of our work). In this case, it is necessary to instead quantify the degree to which a resource is educational and use that score to inform ranking strategies.

2.3.3 Objectivity

Resources on the web are not limited to reputable or fact-based sources [19]. As children do not always have sufficient skills to analyze the quality and reliability of resources [154], it is essential for resources presented to young users to be objective. This would help in removing the need for them to qualify the correctness of expressed opinions, i.e., they might not be able to properly judge if they are reading resources written by experts whose opinions are reliable.

Research on objectivity detection has received a lot of attention over the years [36, 69, 92, 95, 132]. Most work focuses on examining the objectivity of social media resources such as tweets, forums, and movie reviews [69, 95]. Researchers also dedicate efforts to detecting objectivity on web resources, such as blogs and news articles [36, 92, 132], which usually contain longer texts. Some of the existing techniques for objectivity detection rely on pre-defined lexicons to determine the degree to which resources contains vocabulary that map to opinion words [120]. Others depend upon a classification-based approach [36], where resources are labeled as being subjective or objective.

Although subjectivity / objectivity detection is prominent in the area of Natural Language Processing, especially given its correlation with opinion mining tasks, it has not been explicitly adopted to enhance SE retrieval tasks.

2.3.4 Target Audience

A number of researchers have dedicated efforts to detecting the right set of web resources to retrieve in response to queries formulated by young children [53, 54, 71, 72, 104, 123]. Eickoff et al. [53] analyze features such as text complexity, result presentation, and ease of navigation. The authors in [54] extend the work in [53] and include ethical information (i.e., the proportion of ads in a resource) for filtering non-child-friendly resources. To prevent children from accessing inappropriate content, such as pornography, hate speeches and vulgar terms, Patel et al. [124] use a term weighting technique to create features for classifying web pages. Gupta et al. [71] develop a search filter, in order to give a child age appropriate content. They extend their work in [72] by designing an algorithm that makes a child securely access a filtered content, by redirecting the query to the child's educational interest instead of blocking the retrieved resource.

None of the aforementioned strategies, however, simultaneously consider different aspects of child-friendliness and result inappropriateness when determining resources that specifically target young children.

2.4 Ranking Web Resources

As one of the major tasks of the search process, ranking strategies have been well studied [18, 41, 76, 79, 121, 144]. In addition to Google's Page Rank [121] algorithm that assigns importance to a web page by counting the number and quality of links pointing to it, other well-known ranking algorithms include Topical Page Rank [79] and Trust-Rank [76], which prioritize resources from a specific topic and minimize spam, respectively. It is also worth mentioning attempts to enhance ranking

algorithms for a number of domain-dependent tasks by considering features such as geographical references [144], the readability [41], or a combination of semantic and information theoretic techniques [18] for determining the relevance of retrieved resources. The aforementioned algorithms, however, focus on a more general audience.

Compared to adults, children seldom have the experience to identify relevant resources [28, 154], which makes it imperative for high quality, child-suitable resources to be ranked higher on the result list. Unfortunately, little research addresses the retrieval and ranking of high quality resources for children. Gyllstrom et al. [75] prioritize resources in response to a child's search query by designing a linked-based algorithm that ranks web pages according to their appropriateness for children. Furthermore, Bilal et al. [27] use an approach that involves children, mediators and researchers evaluating and ranking resources based on its relevance from a child's perspective. The aforementioned strategies, however, only consider ranking resources based on single aspects, unlike **K**i**S**u**RF**, which prioritizes resources for children by simultaneously considering a number of perspectives, such as their: readability, objectivity, and educational value.

CHAPTER 3

EXAMINING THE LIMITATIONS OF FILTERING AND RANKING STRATEGIES

A number of researchers have dedicated efforts to the design and development of strategies that aim to improve children's experience when performing search discovery tasks [20, 22, 49]. However, there is no empirical evidence in the literature of how existing SEs fare in handling the resources retrieved for children's queries, let alone, searches they conduct in the educational setting. This prompts us to ask questions which we use to guide our analysis on the limitations of ranking and filtering strategies of existing SEs: (i) are SEs too restrictive when it comes to locating resources in response to educational information discovery tasks?; (ii) does safe search always disregard sexually explicit resources?; (iii) are safe search filters effective in filtering web resources that potentially promote violence?; (iv) do SEs retrieve resources that match the reading abilities of school-aged users?; (v) are educational resources ranked high for education-relevant searches?; and (vi) do SEs prioritize specific web domains?

3.1 Search Engines

In conducting our empirical analysis, we consider six different SEs: Google, Bing, their safe search counterparts, Kidrex, and Kidzsearch. We examine Google and Bing, as previous research shows that they are favored by children [59]. We also consider

the safe search options offered by these SEs, as we are particularly interested in investigating how the filters available on children's preferred SEs handle both unsafe and educational resources. Kidrex and Kidzsearch are two popular SEs included among several child-oriented SEs [12], which is why we consider them in our empirical analysis.

Resources retrieved by Google, Bing, and their safe search counterparts were accessed through their API services [6, 9]. Being that Kidrex and Kidzsearch do not make available a search API, we wrote a web scraping script to perform search and retrieval tasks.

3.2 Data Sources for Empirical Analysis

For SE analysis purpose, we rely on a diverse collection of data sources.

- **DMOZ Open Directory Project (ODP)**. This is a collection of web resources categorized by human annotators. In this data source, resources have been categorized by age groups, i.e, kids, teenagers and adults, as well as information types such as news, politics, economy, weather, and sports.[1] We extracted resources in the following kids categories: health, news, entertainment, school, and sports. These categories were selected because they represent diverse information types that children seek for both their educational and leisure search inquiries [123].

- **IDLA Resources**. Upon collaboration with Idaho Digital Learning Academy (IDLA),[2] a K–12 educational institution in Idaho, we gathered a collection of

[1] http://dmoztools.net/
[2] https://idiglearning.net/

66,000 educational resources. These resources have been labeled with their subjects and grade levels by educators, and refer to subjects such as Mathematics, English, Health, Science, and Government. For analysis, we extracted the 4,000 resources that were related to children's health and science subjects.

- **List of Bad Words**. This list includes 1,400 keywords that have been identified as sexually explicit by Google.[3]

- **Hate-speech Dictionary**. We created a dictionary using 1,040 hate speech lexicons compiled by hateBase,[4] a repository of hate speech language. Additionally, we included a refined collection of hate-based and offensive language n-grams created by the authors in [47].

- **Hate-speech Resources**. We gathered a collection of 3,000 hatespeech web resources, which were compiled by Hate Speech Movement,[5] a site known to report websites that promote violence. We sampled 1,000 of these resources for analysis purposes.

- **Children's Queries**. We gathered a collection of 115 queries written by approximately 50 children performing search discovery tasks in an educational setting.[6] The children were in the 4^{th} - 5^{th} grade levels, and conducted all their searches under the supervision of their teacher at Garfield Elementary school in Boise, Idaho.

[3]https://code.google.com/archive/p/badwordslist/
[4]https://www.hatebase.org/
[5]https://nohatespeechmovement.org/
[6]IRB approval number: 131-SB16-103

3.3 Empirical Analysis and Discussion

We present and discuss the results of the empirical analysis conducted to understand the effectiveness of SEs in retrieving resources in response to search tasks initiated by school-aged children. In our experiments, we use a number of queries to simulate the search process and evaluate only the top–10 retrieved results. We only consider these results as existing research shows that children do not go beyond the first page, usually containing the 10 most relevant resources, when conducting search tasks [74].

3.3.1 Are SEs too restrictive when it comes to locating resources in response to educational information discovery tasks?

To investigate how SEs fare in terms of retrieving resources for educational searches, we simulate a context where a child would initiate a search that pertains to subjects taught at school. In order to create queries used in this analysis, we randomly sampled phrases (of up to tri-grams) relating to children's health and science subjects, extracted from the IDLA data source (introduced in Section 3.2). This resulted in $\mathbf{Ed_{kw}}$, a list of the top–1000 n-grams, selected according to their TF-IDF scores.[7] We then examine the number of times SEs under study fail to retrieve results in response to these searches.

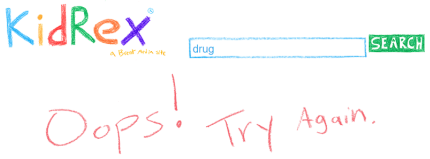
As shown in Table 3.1, child-oriented SEs, i.e., Kidrex and Kidzsearch, failed to retrieve resources for some of these searches. Google (with and without the safe search option) retrieved results for all of these searches, whereas Bing safe search retrieved no results for 1%. We found that some searches that led to no results were initiated with queries like "drug" or "breast tissue" (see Figure 3.1). Based on the K–12 curriculum,

[7]TF-IDF–is a weighting scheme that is used to reflect the importance of a word in a document, as well as among a collection of documents in a corpus.

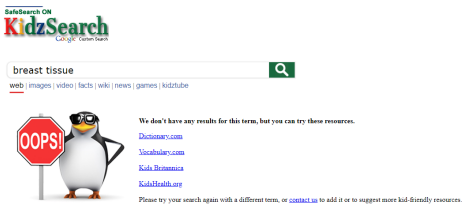
these searches could be relevant to children's biology, human anatomy, and science subjects. In designing SE retrieval algorithms, it would beneficial to instead effectively examine the content of the retrieved resources or consider the context of the search, when determining if a resource should be filtered.

Table 3.1: Percentage of searches initiated with queries in $\mathbf{Ed_{kw}}$ that led to no results.

Search Engine	**Searches that Retrieve No Results**
Bing	0%
Bing (Safe search)	1%
Google	0%
Google (Safe search)	0%
Kidrex	5%
Kidzsearch	3%



(a) Kidrex does not retrieve results for the search initiated with the query "drug", which relates to inquires on a child's health and science subject.



(b) Kizsearch retrieves no result in response to the query "breast tissue", which pertains to human anatomy subject, common in the K-12 curriculum.

Figure 3.1: Examples that showcase limitations of Kidrex and Kidzsearch in handing queries with terms relevant to children's educational health and science subjects.

3.3.2 Does safe search always disregard sexually explicit resources?

We investigate the performance of safe search filters available on SEs in handling sexually explicit content–one of the main goals of safe search [7]. In doing this, we initiate the search with queries containing sexually explicit terms, which we create by

randomly sampling 1,000 keywords from the list of bad words described in Section 3.2. We refer to these queries as $\mathbf{B_{kw}}$.

Being that there is no ground truth to determine if a website includes sexually explicit content, we turn to an online tool: Web Shrinker.[8] This tool assign categories, such as educational, adult content, news, and entertainment, to resources. We then compute the percentage of searches which retrieved at least one result categorized as "adult content". It is important to note that we exclude Google and Bing from this analysis, as they are not meant to filter explicit resources.

As showcased in Table 3.2, Kidzsearch and Kidrex performed better when compared to the other SEs in terms of not retrieving resources that contain sexually explicit materials. In fact, for 39% and 57% of the searches respectively, Kidzsearch and Kidrex retrieved no results. Note that retrieving no resource for queries in $\mathbf{B_{kw}}$ is an advantage, as it indicates that the SE accurately identifies the search context to be inappropriate. On the opposite spectrum, Google's safe search retrieved results for 95% of these searches, while Bing retrieved for approximately 71%. Having resources with sexually explicit content slip through the filter is problematic, as parents specifically turn on safe search options available on SEs for preventing their children from accessing inappropriate content.

Table 3.2: Assessment of SEs in responding to searches formulated for accessing sexually explicit content.

Search Engine	**Searches that**	
	include Sexually Explicit Content	**Retrieves No Results**
Bing (Safe search)	33%	29%
Google (Safe search)	16%	5%
Kidrex	14%	39%
Kidzsearch	9%	57%

[8]https://www.webshrinker.com/

3.3.3 Are safe search filters effective in disregarding resources that potentially promote violence?

We examine how safe search filters available on SEs under study perform when it comes to retrieving resources that contain hate speech and offensive language, as materials that promote violence are considered inappropriate for children [87]. For doing this, we use a number of queries with terms containing hate-based keywords to simulate the search process. We create these queries–which we refer to as $\mathbf{H_{kw}}$–by randomly selecting 1,000 lexical items from the hate speech dictionary introduced in Section 3.2. Similar to the assessment performed in Section 3.3.2, we compute the percentage of searches that retrieve at least one resource categorized as hate-based. We also exclude Google and Bing from this analysis, as they were originally not designed to disregard violence-related resources. To the best of our understanding, there is no ground truth that determines the degree to which a document is violence-related. As a result, we rely on the hate speech detection algorithm introduced in [47], in order to label resources that contain either offensive language or hate speech. Although this algorithm was originally designed to detect hate-speech or offensive language on tweets, we empirically verified that it was able to accurately label 95% of the Hate-Speech resources described in Section 3.2 to be hate-based. Moreover, it labeled all kids resources extracted from DMOZ as non-hate-based.

As shown in Table 3.3, safe search filters available in Google and Bing retrieved more resources labeled to be violence-related, when compared to Kidzsearch and Kidrex. Again, retrieving no resource for searches initiated using queries in $\mathbf{H_{kw}}$ is an advantage, as it indicates that the safe search filter on the corresponding SE is able to identify the context of the search to be inappropriate. Our results also

show that Google and Bing retrieved resources respectively for 95% and 98% of these searches, which was higher when compared to that of Kidzsearch and Kidrex. This prompts the need for further improvement in filtering strategies available on popular SEs, as children may be unwittingly exposed to content that promote violence, when conducting their educational or leisure searches.

Table 3.3: Assessment of SEs in handling searches conducted with the aim of accessing violence-related content.

Search Engine	**Searches that**	
	include Violence-related Content	**Retrieves No Results**
Bing (Safe search)	22%	5%
Google (Safe search)	23%	2%
Kidrex	15%	12%
Kidzsearch	11%	15%

3.3.4 Do SEs retrieve resources that match the reading abilities of school-aged users?

According to Lennon and Burke [106], a reader can comprehend a text when he / she understands 75% of its content. This is why it is imperative that the resources retrieved in response to children's queries match their reading abilities when they conduct searches. With this in mind, we investigate the reading levels of resources retrieved in response to searches conducted by children in an educational setting. To simulate a child's search process, we sampled 100 queries, which we refer to as $\mathbf{K_{qry}}$, from among the children's queries introduced in Section 3.2. To determine the average readability of the retrieved web results, we explore a number of readability formulas: Flesch Kincaid [100], Dale-Chall [35], Spache [145], and Smog [115]. We consider these formulas, as existing research have demonstrated their applicability in determining reading levels in web documents [25, 151].

As shown in Table 3.4, the average grade levels among the resources retrieved for queries in $\mathbf{K_{qry}}$ were relatively high. Most of them happen to be four to six grades above that of children that formulated these queries, i.e., children in the 4^{th} and 5^{th} grades. Results from this analysis demonstrate the need to consider the reading abilities of a child when prioritizing retrieved resources. Although a resource may be relevant to information needs expressed in the queries used to trigger the search, they may be completely irrelevant if the child can not understand its content.

Table 3.4: Average grade levels computed for top–10 results retrieved for queries in $\mathbf{K_{qry}}$.

Search Engine	**Readability Formulas**			
	Flesch Kincaid	**Smog**	**Dale-Chall**	**Spache**
Bing	9^{th}	8^{th}	9^{th}	5^{th}
Bing (Safe search)	9^{th}	8^{th}	9^{th}	5^{th}
Google	9^{th}	8^{th}	10^{th}	6^{th}
Google (Safe search)	9^{th}	8^{th}	10^{th}	6^{th}
Kidrex	11^{th}	8^{th}	9^{th}	5^{th}
Kidzsearch	11^{th}	8^{th}	9^{th}	5^{th}

3.3.5 Are educational resources ranked higher for education-related searches?

Recall that children tend to analyze results on the first page when conducting searches [74]. Thus, in the context of educational searches, it would be beneficial for relevant educational resources to positioned higher on the ranked list of resources. With this in mind, we investigate how SEs respond to educational searches–particularly how they position educational resources retrieved for these searches. This is different from the analysis conducted in Section 3.3.1, as here we focus on the ranking of the resources instead of the filtering.

In performing this analysis, we simulate the search process using queries we created by extracting the titles of resources known to be educational. We refer to these queries as $\mathbf{E_{qry}}$. For this purpose, we sampled 1,000 educational resources from the IDLA data source (introduced in Section 3.2). We treat as the ground truth the URLs of the web resources from which we extracted the titles. We then explore the percentage of searches that retrieve the URLs for the corresponding queries, i.e., their titles, as well as the average position the SEs assign to these URLs.

As showcased in Table 3.5, the child-oriented SEs underperformed in terms of retrieving the corresponding URLs for queries in $\mathbf{E_{qry}}$. In fact, Kidrex assigned a lower position to most of the educational resources, when compared to other SEs. We also observe that Google (with and without safe search) is less stable than its counterparts for positioning the education-relevant resources higher on the ranked list, as evidenced by the variance in the ranking assigned to these resources. This is problematic, being that Google is the SE most favored by children [128], when conducting educational search tasks.

Given that children are known to select resources in order, the relatively low average position assigned to educational resources is a concern. This increases the likelihood of children not even looking at educational-relevant resources retrieved by SEs due to their low ranking position. In the end this could prevent access to the right resources; in turn, leading to an unsuccessful educational search session.

Knowledge from this analysis informs us about the need for improving SE retrieval strategies, as the benefits of assigning a higher position to educational web resources for education-related searches is clear.

Table 3.5: Assessment of the effectiveness of SEs in handling educational searches initiated using queries in $\mathbf{E_{qry}}$.

Search Engine	**Searches that Retrieve the exact URL**	**Average position**	**Variance**
Bing	27%	5^{th}	1.58
Bing (Safe search)	27%	5^{th}	1.70
Google	26%	5^{th}	3.64
Google (Safe search)	27%	5^{th}	3.12
Kidrex	24%	5^{th}	1.59
Kidzsearch	23%	6^{th}	1.39

3.3.6 Do SEs prioritize specific web domains?

A myriad of information sources exist on the web, containing URLs that showcase their domain names and suffixes. Educational web domains use the domain suffix ".edu" [52], while commercial domains like Amazon or StackOverFlow use the suffix ".com" [8]. Other popular domain suffixes are ".org", which is used for educational and non-profit organizations, as well as ".net" and ".gov", used by Government entities and by organizations [8]. In order to examine the degree to which the aforementioned domains are prioritized in response to educational searches, we conduct an empirical analysis. To simulate the searches conducted in an educational environment, we use $\mathbf{K_{qry}}$ (introduced n Section 3.3.4), and compute the average position of the resources retrieved in response to these queries for each of the domain suffixes under study.

As shown in Table 3.6, across the SEs under study, educational resources were consistently ranked low, which is a problem as these searches were education-related. On the other hand, commercial websites were repeatedly positioned first or second on the results list, which was anticipated, as the domain suffix of the most visited sites on the web is ".com", based on the Alexa ranking statistics.[9] ".org" websites were

[9]Alexa is a well known organization that provides statistics on top-ranked websites world wide [4].

also more favored than ".edu", as they consistently appeared on the 2^{nd} position.

Table 3.6: Average ranking assigned to resources from different domains for searches initiated using queries in $\mathbf{K_{qry}}$.

Search Engine	**Average position of Searches with Domain from**					
	Wikipedia	**.org**	**.gov**	**.com**	**.edu**	**.net**
Bing	3^{rd}	2^{nd}	4^{th}	2^{nd}	18^{th}	12^{th}
Bing (Safe search)	3^{rd}	2^{nd}	4^{th}	2^{nd}	17^{th}	12^{th}
Google	5^{th}	2^{nd}	4^{th}	1^{st}	36^{th}	26^{th}
Google (Safe search)	5^{th}	2^{nd}	4^{th}	1^{st}	36^{th}	25^{th}
Kidrex	4^{th}	2^{nd}	6^{th}	1^{st}	30^{th}	21^{st}
Kidzsearch	3^{rd}	2^{nd}	14^{th}	2^{nd}	24^{th}	18^{th}

In addition to examining domain suffixes, we investigate the average position assigned to Wikipedia pages, being that this information source tends to appear for majority of online searches on Google and Bing [10]. As seen in Table 3.6, Wikipedia pages are ranked higher than the educational resources by all the SEs, which should not be the case as the latter may be more informative and suitable from a child's perspective. Existing research show that web resources appealing to children may contain few texts and more graphics [68], which is not common in Wikipedia resources. Consequently, as children seldom have sufficient skills to judge the validity of the information they access [28, 154], it would be useful to position educational resources higher than those from Wikipedia, for searches conducted by children in an educational setting.

3.4 Lessons Learned

We presented the results of an empirical analysis which informs us about limitations that might hinder the retrieval of resources relevant to school-aged users conducting searches in an educational environment. Among the investigated SEs, Kidrex and Kidzsearch were more effective than the general-purpose ones in filtering sexually

explicit and hate-speech content. However, these SEs underperformed when responding to educational searches, as they sometimes failed to retrieve resources for these searches and positioned educational resources low on the result list. This was also an issue we found with Google and Bing (including their safe search counterparts) which performed poorly in eliminating resources containing inappropriate content. Moreover, all the SEs under study retrieved resources that were of a higher reading level than for children that conducted the search (4^{th} and 5^{th} grades).

The knowledge gained from these experiments indicates the need for designing and developing a strategy that simultaneously and explicitly integrates several aspects in retrieving resources for children's educational searches, which is what we discuss in the next Chapter.

CHAPTER 4

KiSuRF: A MULTI-PERSPECTIVE STRATEGY FOR OFFERING SUITABLE WEB RESOURCES TO CHILDREN

We introduce **K**i**S**u**RF**–**K**ids **S**uitable Resource **R**anker and **F**ilter–a strategy that complements SE functionality by addressing the limitations of existing SEs when responding to searches pertaining to children's educational information discovery tasks. We first discuss tools and text processing operations applied in the design and development of **K**i**S**u**RF**. Thereafter, we present details of its architecture for filtering and ranking resources.

4.1 Programming Language and Tools

In designing **K**i**S**u**RF**, we use Python due to its versatility. We also relied on Seaborne, Matplotlib, and GGplot for visualizations, along with tools and libraries which we outline below.

Natural Language Toolkit (NLTK) [32]. This is a text processing and machine learning toolkit that offers several capabilities, including tokenization, stemming, part-of-speech tagging, and chunking.

PyEnchant [94]. The Pyenchant library spell-checks words written in several languages including English, German, Spanish, and French. This library also has the functionality to suggest the most likely correct spelling for a given misspelled word.

TLDExtract [102]. The TLDExtract toolkit offers capabilities to segment a URL based on its domain name, suffix, and sub domain name.

Sci-kit Learn [125]. Sci-kit learn is a machine learning toolkit that offers several data mining and analysis libraries, in addition to classification, regression, clustering, and dimension reduction algorithms.

Requests [129]. This is a Apache2 Licensed HTTP library that makes it possible to access the response data of retrieved resources in json and xml formats.

Beautiful Soup [131]. The Beautiful Soup library offers capabilities to iterate, search, and extract information from a URL, and thus, is suitable for web scraping.

TensorFlow [148]. This is an open source software library for high performance numerical computation, commonly used for machine learning and neural network applications.

4.2 Text Extraction and Processing

In the context of our work, we focus on websites retrieved by SEs, which in addition to textual content, include HTML tags, cascading style sheet (CSS), and JavaScript.

For analysis of retrieved resources, we remove all non-textual contents, and examine information obtained from their main page, meta tags, and anchor tags. Furthermore, we perform a series of text processing operations, common in Information Retrieval, in order to make extracted raw content suitable for computer models to interpret.

Tokenization. Tokenization is the process of extracting sequences of strings, i.e., tokens, from a piece of text [46]. These tokens may be in the form of a word, punctuation, number, or acronym. To tokenize resources, we use the NLTK library (introduced in Section 4.1). We normalize the extracted tokens by removing all punctuations. An example of a tokenized sentence is showcased in Table 4.1.

Table 4.1: Example of text processing operations performed on a sample sentence extracted from the children's book "Alice's Adventures in Wonderland" written by Lewis Carroll, where NN–noun, VBD–verb in the past participle, VBG–verb in the present participle, JJ-adjective, and PRP–pronoun.

Sentence	ALICE was beginning to get very tired of sitting by her sister on the bank.															
Tokens	ALICE	was	beginning	to	get	very	tired	of	sitting	by	her	sister	on	the	bank	.
Remove Stop words and Punctuations	ALICE		beginning		tired		sitting			her		sister		bank		
Parts of Speech	NN		VBG		JJ		VBG			PRP		NN		NN		

Stop Words Removal. Stop words are function words that contribute to the structure of a sentence but have little meaning on their own. These words are commonly articles or prepositions, such as "the", "a", "for", and "it" [46]. By discarding stop words, we improve processing time, while retaining informative words. To identify and remove stop words, we depend upon the stop list provided by NLTK. An example of a sentence after removing stop words is seen in Table 4.1.

Part of Speech (PoS) Tagging. In Natural Language Processing, PoS tagging is the process of labeling a word in a text with its part-of-speech based on the context

in which the word appears in that text [46, 143]. We show an example of the words in a sentence tagged with their respective PoS in Table 4.1. We use the NLTK PoS tagger, which was originally trained on the Penn TreeBank tag set containing 36 PoS tags for the English Language.[1]

Stemming. Stemming is the process of reducing several variations of a word into a common root or stem [46]. These words may differ as a result of plurals, verb tenses, or suffixation. We consider stemming, as this helps to reduce the vocabulary size, which in turn improves processing time. To stem words, we use the snow ball stemmer made available by NLTK. We showcase an example of different words that map to a common stem in Table 4.2.

Table 4.2: Example of different variations of the term "stay" and their corresponding stem.

Original Word	stay	stays	staying	stayed
Stem	stay			

4.3 Design Overview of KiSuRF

KiSuRF aims to retrieve resources considered suitable for school-aged children. It does so by relying on filtering and prioritization steps based on supervised and semi supervised learning. **KiSuRF** first applies a novel filtering approach that does not only disregard resources considered unsafe for children, i.e., violence-related and sexually explicit content, but retains those that are relevant to a child's educational search context. Thereafter, **KiSuRF** simultaneously examines the remaining resources from multiple perspectives in order to prioritize the ones that match the reading ability

[1] https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html

of a child in the $\mathbf{3^{rd}}$ – $\mathbf{5^{th}}$ grades; that are objective; and of educational value. We outline the architecture of **KiSuRF** in Figure 4.1.

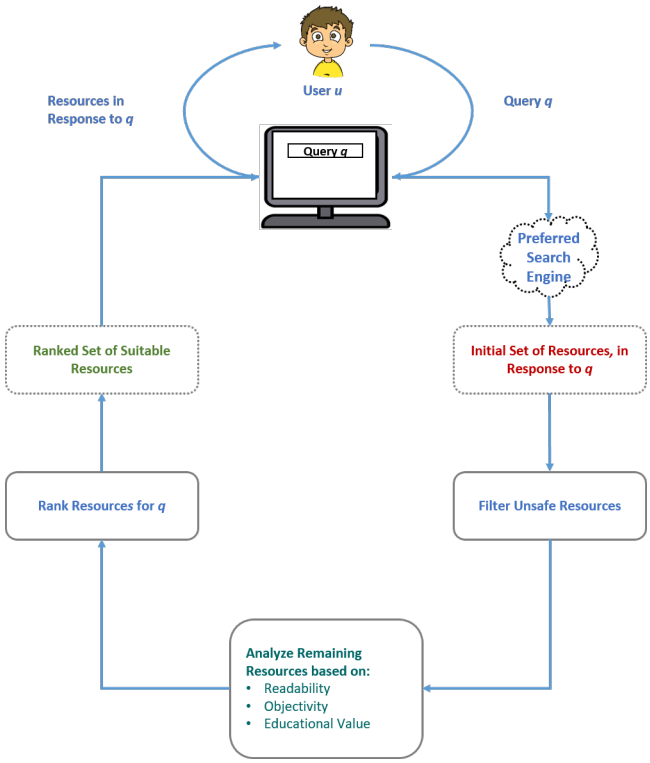


Figure 4.1: Architecture of **KiSuRF**, which complements existing SEs by filtering resources with inappropriate content and prioritizing the ones that are suitable to a child in terms of readability, objectivity, and educational value.

4.4 Data Sources used in the Design and Development of KiSuRF

In designing **KiSuRF**, we depend upon on a number of data sources for analysis.

- **Sexually Explicit Resources - $\mathbf{Exp_{res}}$.** We gathered 4,000 web resources that have been categorized as Sexually Explicit and Adult News on Alexa, an

online directory that assigns categories to resources based on their global traffic rank and the type of information they contain.

- **Hate-Speech Resources - $\mathbf{Hate_{res}}$.** We use the same set of hate-speech resources described in Section 3.2.

- **Kids Safe and Unsafe Resources - $\mathbf{K_{safe}}$.** We extracted 14,000 documents labeled as safe and unsafe for children using resources in $\mathbf{Exp_{res}}$ and $\mathbf{Hate_{res}}$ as the unsafe ones, as well 7,000 resources from DMOZ (described in Section 3.2) which we label as safe for children.

- **Wikipedia Resources - $\mathbf{Wiki_{res}}$.** This collection consist of 1,700,000 Wikipedia articles.

- **Educational and Non-educational Resources - $\mathbf{Diverse_{res}}$.** We extracted the descriptions of 3,000 children educational books from Teachers Pay Teachers which we label as educational resources. We treat as non-educational, 600 celebrity news articles extracted from InTouch Magazine [1], as well as 600 resources each pertaining to sports, weather, economy and politics made available in [91].

4.5 Eliminate Inappropriate Resources

KiSuRF first examines resources and applies a filtering strategy that disregards the ones with inappropriate content while retaining those that potentially match the educational information needs of children.

4.5.1 Quantify the Degree of Sexual Explicitness in Resources

In order to measure the degree to which a retrieved resource R exhibits sexual explicitness, we compute a number of *numerical data points* that examine R's main content R_{con}, meta tag R_{meta}, as well as the title attribute of all its anchor tags R_{anc_title}.

Sexual Explicitness in Resource Content. Resources with adult content are known to contain a high frequency of sexual explicit terms [124], which is why we capture this information using two data points:

DP_1 *Unique Count of Sexually Explicit Words*, which we measure as the distinct count of words in R_{con} that correlate with words in the list of bad words introduced in Section 3.2.

DP_2 *Proportion of Sexually Explicit Words*, which we measure as the total count of words in R_{con} that match terms in the aforementioned bad list, divided by the total number of words in R_{con}.

Upon analysis of resources in $\mathbf{Exp_{res}}$, we noticed misspellings in their content that map to inappropriate words, e.g., "phukk" and "masterba3". We argue that these websites intentionally spell inappropriate terms incorrectly, in order to pass through certain web filters, much like safe search. To capture this information, we consider two more data points:

DP_3 *Unique Count of Misspelled Sexually Explicit Words*, which we measure as the distinct count of misspelled[2] words in R_{con} that exist in a list of misspelled sexually explicit lexicons (introduced in Section 3.2).

[2]To identify misspelled words, we rely on the enchant library described in Section 4.1.

DP_4 *Proportion of Misspelled Sexually Explicit Words*, which we measure as the total number of misspelled words in R_{con} that exist in the aforementioned list, divided by the total number of words in R_{con}.

Sexual Explicitness in Resource Meta Tag. In building websites, designers often use the meta tag to provide an overall site description [11]. We explored $\mathbf{Exp_{res}}$ and verified the inclusion of representative terms in descriptions embedded in their meta tags. For instance, frequent keywords we found among resources in $\mathbf{Exp_{res}}$ include "sex" and "hardcore" in the meta tag. For this reason, and following the premise proposed by the authors in [108], we examine information embedded in this tag. In doing this, we create data points DP_5 and DP_6, which we compute as described in DP_1 and DP_2 but applied to R_{meta}.[3]

Sexual Explicitness in Resource Anchor-tag. A hyperlink, usually defined with an anchor tag, is used to link a website to other websites–which we refer to as target websites [153]. Common attributes present in the anchor tag are the "href" and "title". The former contains the URL of the target website, while the latter is often used by designers to include the main title or description of the target website [83]. Based on an initial analysis of websites in $\mathbf{Exp_{res}}$, we observed that they contained hyperlinks to other websites that conveyed the same type of information, e.g., educational websites mostly linked to other educational websites, whereas pornographic websites linked to other pornographic websites. Moreover, we found representative keywords of the target website in the title attribute, which is what prompts us to examine information extracted from this tag. We consider two

[3]We sampled resources in $\mathbf{Exp_{res}}$ and verified that the meta tags rarely included misspellings.

other data points focused on R_{anc_title}, DP_7 and DP_8, which are computed as in DP_1 and DP_2.

4.5.2 Determine the Degree to which Resources Promote Violence

Similar to the assessment outlined in Section 4.5.1, we examine R_{con}, R_{meta}, and R_{anct_title}, in order to quantify the degree to which retrieved resources potentially promote violence, i.e., contain hate-speech and offensive language.

Hate-speech in Resource Content. Hate speech is common among resources where users write comments on contradictory topics, including politics, relationships or race, as well as on social sites, such as Twitter and Facebook [47, 48]. These sites are known to include a high frequency of hate-words in their content [48], motivating us to examine two data points.

DP_9 *Unique Count of Hate-Based Words*, which we measure as the distinct count of words in R_{con} that correlate with words in the dictionary of hate-based words introduced in Section 3.2.

DP_10 *Proportion of Hate-Based Words*, which we measure as the total count of words in R_{con} that match terms in the aforementioned bad list, divided by the total number of words in R_{con}.

Hate-speech in Resource Meta-tag. When examining resources in $\mathbf{Hate_{res}}$, we observed that a high percent of them contained hate-based keywords in the meta tags (see Figure 4.2). With this in mind, we create two data points DP_11 and DP_12, computed as described in DP_5 and DP_6, but applied to R_{meta}.

```
<meta name="keywords" content="nigger, niggermania, racist, jokes, humor, kkk, niggers, Tom Shelly, Raptorman, AFN FAQ" />
<meta name="description" content="The Niggermania Forum for nigger jokes, facts, and bashing. All are welcome to join but no niggers !" />
```

Figure 4.2: Example of a hate-based content embedded in the meta tag of **Nigger-Mania**, a site known to promote hate-speech.

Hate-speech in Resource Anchor-tag. To account for representative hate-based terms embedded in title attributes of resource anchor tags, we also consider two data points DP_13 and DP_14, calculated as discussed for DP_5 and DP_6, but on information we extract from R_{anc_title}.

4.5.3 Retain Education-relevant Resources

One of the main motivations of **KiSuRF** is to ensure it retains educational resources even if they happen to contain terms that in isolation may be misconstrued as being inappropriate by safe search filters. To do so, **KiSuRF** quantifies the extent to which R is educational by analyzing terminology common among educational searches.

As a way to capture educational terminologies in R_{con}, we use the IDLA resources introduced in Section 3.2. From this collection, we only examine resources for children's health and science subjects, as they are most likely to include terms that are often misinterpreted as inappropriate by filters much like safe search, when compared to other subjects like Mathematics or Social Sciences. We then compute TF-IDF on terms extracted from this collection, in order to identify the top significant educational n-grams. We consider n-grams of size 1, 2, and 3, i.e., uni-grams, bi-grams and tri-grams, as they were sufficient for capturing educational contextual information without excessively increasing the feature dimension space. We treat the set of 10,000 most informative n-grams, i.e., the ones with the highest frequency across the IDLA

resources after removing stop words, as our educational vocabulary V.[4] Note that for IDF, we take advantage of an external collection $\mathbf{Wiki_{res}}$, so as to quantify the significance of the terms in V. Following this, we represent R_{con} as a TF-IDF vector representation of V, which we define as R_{con_vec}.

Due to the high dimensionality of R_{con_vec}, which is usually sparse, we reduce it into a small number of relevant data points. Doing this helps to reduce the computational cost for learning models without affecting the resource representation. We follow the premise by the authors in [136] by applying the truncated Singular Value Decomposition (SVD) technique [64], which has shown positive results for text mining tasks [37, 126, 136]. When using this technique, we are required to indicate the desired dimension, i.e., number of principal components which is also known as the eigenvalue, for the output vector. We determine this number by selecting the best "eigenvector" of the covariance matrix of the data, i.e., the eigenvectors corresponding to the largest eigenvalue. To select the eigenvalue, we first represent resources in $\mathbf{K_{safe}}$ as a matrix, $Matrix(\mathbf{K_{safe}})$ consisting of TF-IDF weights, as shown below.

$$Matrix(\mathbf{K_{safe}}) = \begin{pmatrix} x_{11} & x_{12} & ... & x_{1j} \\ . & . & & . \\ . & & . & . \\ x_{i1} & x_{i2} & ... & x_{ij} \end{pmatrix}$$

where x_{ij} is the TF-IDF weight of the j^{th} term for the i^{th} resource in $\mathbf{K_{safe}}$, i = 1, 2, ..., $|\mathbf{K_{safe}}|$, and j = 1, 2, ..., $|V|$.

We depend upon the sklearn implementation of truncated SVD, and initially set the desired number of principal components to its initial dimensions size. We then

[4]We depend upon the scikit-learn toolkit for implementing TF-IDF.

select the eigenvalue (number of components) with enough variance that explains $Matrix(\mathbf{K_{safe}})$. As seen in Figure 4.3, an eigenvalue of 1,000 was sufficient for this purpose, as the cumulative variance explained at this point is close to 1. Following this, we use this value as the new dimension size of R_{con_vec}, hence creating the data point below:

DP_15 *Vector Representation of R_{con} which Reflects Terms Common among Children School Subjects*, which we compute as the reduced dimension of R_{con_vec} using the truncated SVD.

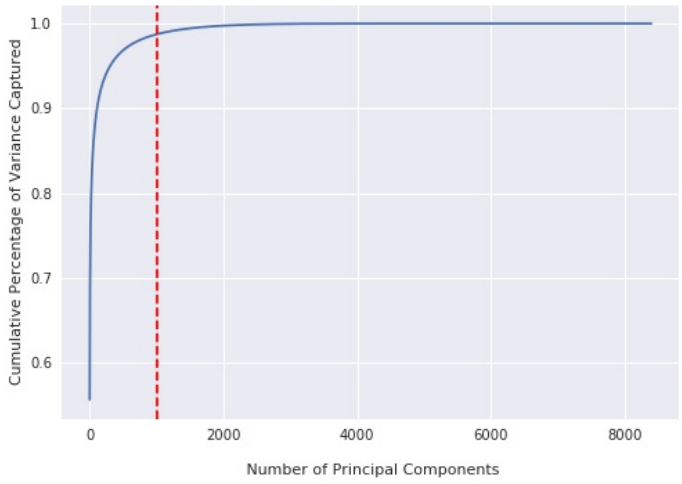


Figure 4.3: Cumulative scree plot showing the number of principal components suitable for reducing the dimension of TF-IDF vectors used in representing resources.

4.5.4 Aggregate Data Points for Filtering

To showcase the influence of the data points presented in Sections 4.5.1- 4.5.2 (and summarized in Table 4.3) in terms of determining whether a document is safe, we

computed Pearson's correlation among data points and the class (i.e., label) representing whether a resource is safe or not. For this analysis, we depend upon resources in $\mathbf{K_{safe}}$. We do not include DP_15 in the correlation analysis, as this would require that we present the TF-IDF vector of each resource as one single numerical value.

As illustrated in Figure 4.4, the data points that best correlate with the label (kid safe or unsafe) are DP_2, DP_6, and DP_10, which correspond to the hate speech and inappropriate words proportion in R_{con}, as well as the inappropriate words proportion in R_{meta}. These correlations indicate the relevance of the corresponding data points in terms of determining whether a document should be filtered or not. Furthermore, these data points are negatively correlated with the class, meaning that the more hate speech or inappropriate word in a resource content, or the more inappropriate words in its meta tag, the more likely it is considered unsafe. We also considered pairs of data points and observed that examined pairs do not exhibit high correlation among each other, indicating that they indeed describe resources from different aspects, and as a result, need to be considered in tandem in identifying (un)safe resources.

To aggregate all data points outlined in Section 4.5, **K**i**S**u**RF** uses Random Forest [127], a supervised learning algorithm that *"collectively combines tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest"* [127].[5] In our context, we represent R as a tuple of the form $\langle DP\_1, DP\_2, \ldots DP\_15 \rangle$. We then use this as input to the Random Forest classifier, and get as output a label that indicates if R should be filtered.

[5]We empirically verified in Section 5.2.6 that Random Forests was the model best suited for designing **K**i**S**u**RF**'s filtering strategy.

Table 4.3: Overview of **KiSuRF**'s filtering data points.

Data Point	Analysis Type	Definition
DP_1	Sexual explicitness	Unique count of sexually explicit words in resource content
DP_2	Sexual explicitness	Proportion of sexually explicit words in resource content
DP_3	Sexual explicitness	Unique count of misspelled sexually explicit words in resource content
DP_4	Sexual explicitness	Proportion of misspelled sexually explicit words in resource content
DP_5	Sexual explicitness	Unique count of sexually explicit words in resource meta tags
DP_6	Sexual explicitness	Proportion of sexually explicit words in resource meta tags
DP_7	Sexual explicitness	Unique count of sexually explicit words in resource anchor tags
DP_8	Sexual explicitness	Proportion of sexually explicit words in resource anchor tags
DP_9	Promotes violence	Unique count of hate-based words in resource content
DP_10	Promotes violence	Proportion of hate-based words in resource content
DP_11	Promotes violence	Unique count of hate-based words in resource meta tag
DP_12	Promotes violence	Proportion of hate-based words in resource meta tag
DP_13	Promotes violence	Unique count of hate-based words in resource anchor tag
DP_14	Promotes violence	Proportion of hate-based words in resource anchor tag
DP_15	Educational terms	Vector representation of resource content which reflects terms common among children school subjects

4.6 Prioritize Suitable Resources

In order to prioritize resources that are suitable for children, we analyze the remaining resources based on their readability, objectivity, and educational pertinence.

4.6.1 Estimate the Reading Level of Resources

To prioritize resources that school-aged children can comprehend, **KiSuRF** estimates the readability of R in DP_16. In doing this, we use the Spache reading index [145]. This readability formula examines shallow features, e.g., average sentence length, and depends upon a list of simple words,[6] in order to estimate the reading level of a resource. We consider this readability formula, as existing research has demonstrated its applicability in determining the reading levels of web resources [113], a fact which we empirically verified in Section 3.3.4.

[6]The Spache list of simple words is found in: http://www.readabilityformulas.com/articles/spache-formula-word-list.php

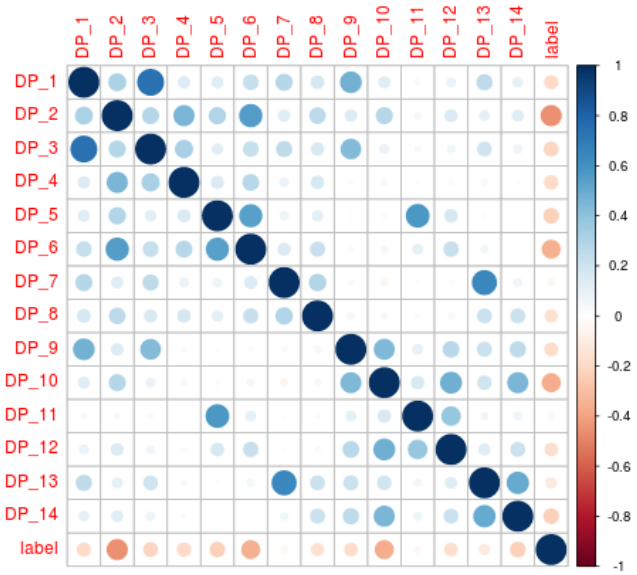


Figure 4.4: Correlation among analyzed data points in **KiSuRF**'s filtering strategy. Color represents the polarity of correlation, and bubble size indicates correlation strength.

DP_16 *Spache Readability Index of Resource Content,* R_{con}, which we measure by applying Spache formula described in Equation 5.13 to R_{con}.

$$Spache = (0.141 \times ASL) + (0.086 \times PDW) + 0.839 \tag{4.1}$$

where ASL is the average length of sentences in R_{con} and PDW is the percentage of difficult words.[7]

[7]The Spache formula determines a word to be difficult if it does not exist on a list of Spache's simple words.

4.6.2 Quantify the Degree of Objectivity in Resources

The proliferation of resources on the web have presented opportunities for people to voice their opinions, some of which do not convey fact-based information [19]. Popular resources known to contain opinionated content are blogs, online forums, as well as other popular social media sites such as Facebook and Twitter. Being that children may not have the right skills to evaluate if a resource contains the right information [154], we want to ensure that in responding to educational online quests, we prioritize resources with a high level of objectivity.

To detect R's degree of objectivity, we adapt the language-model-based approach introduced by the authors in [92]. The authors define objectivity scores for resources by examining their content in order to determine if they are more likely to contain vocabulary common among objective or subjective documents. In designing our objectivity scoring model, we build subjective and objective language models using Equation 4.2 and Equation 4.3.

$$\theta_S : \{P_{ML}(w_i|\theta_O) = \frac{c(w_i, C_S)}{\sum_{w_j \varepsilon V_{c_S}} c(w_j, C_S)} = \frac{C(w_i, C_S)}{|C_S|}\}_{i=1}^{|V_{C_S}|} \tag{4.2}$$

$$\theta_O : \{P_{ML}(w_i|\theta_O) = \frac{c(w_i, C_O)}{\sum_{w_j \varepsilon V_{c_O}} c(w_j, C_O)} = \frac{C(w_i, C_O)}{|C_O|}\}_{i=1}^{|V_{C_O}|} \tag{4.3}$$

where C_S and C_O represent subjective and objective collections, and V_{C_S} and V_{C_O} represent the respective vocabulary of these collections.

We also build a language model for the resource content, R_{con}, on-the-fly using Equation 4.4.

$$P(w_i|\hat{\theta}_{R_{con}}) = \frac{c(w, R_{con}) + \alpha}{\sum_{w_i \varepsilon Z} c(w_i, R_{con}) + \alpha\,|Z|} \tag{4.4}$$

where $P(w_i|\hat{\theta}_{R_{con}})$ is the probability distribution for all words in R_{con}, α is a smoothing parameter and Z is the length of R_{con}'s vocabulary.[8]

We therefore use as DP_17, the objectivity score of R_{con}, which we compute as Equation 4.5.

$$DP\_17 = -(D(\theta_o||\theta_{R_{con}}) - D(\theta_s||\theta_{R_{con}})) \tag{4.5}$$

where $D(\theta_o||\theta_{R_{con}})$ and $D(\theta_s||\theta_{R_{con}})$ are computed using KL_divergence [152], as shown in Equation 4.6. KL_divergence quantifies the similarity between the probability distribution of R_{con} and the reference language models, i.e., subjective or objective.

$$D(\Theta_1||\Theta_{R_{con}}) = \sum_{w \epsilon V} P(w|\theta_1) log \frac{P(w|\theta_1)}{P(w|\theta_{R_{con}})} \tag{4.6}$$

where θ_1 represents the reference subjective or objective language models and $\theta_{R_{con}}$ represents R_{con}'s language model.

4.6.3 Quantify the Degree to which Resources are Educational

In order to prioritize resources that satisfy educational information needs, we also examine R's educational value. To accomplish this, we take advantage of a number of data sources that allow us to examine relations in R_{con} from educational and non-educational perspectives. Thereafter, we compute a number of qualitative indicators which we use to determine the degree to which R is educational.

[8] This computation was different from the one defined by the authors, who instead use a two-stage smoothing approach for building the resource language model.

Examined Data Sources for Educational Pertinence

To inform the design process for mapping resources to the K–12 curriculum, we rely on Probase [158], as well as on a knowledge base that captures K–12 curriculum terminology.

1. **Education Domain Knowledge Base.** As discussed in Section 2.3.2, to identify information relevant to a specific domain, researchers rely on ontologies, taxonomies, or knowledge bases. While these sources allow us to capture conceptual information in documents, they do not explicitly exist for the educational domain. To address this limitation, we create a novel educational knowledge base: $\mathbf{Edu_{KB}}$.

   Design. A knowledge base is known as a conceptual or semantic network, which contains relations among words and their respective concepts. By using an education-specific knowledge base, we hypothesize that we would identify concepts and topics relevant to the K–12 curriculum. In this context, we refer to educational concepts as abstract or general ideas that are representative of K–12 topics. For instance, "shapes" and "angles" are concepts that are relevant to the educational topic "Geometry".

   Educational Topic and Concept Pairs. In designing $\mathbf{Edu_{KB}}$, we first need resources so as to leverage information that is relevant to the K–12 curriculum. In doing this, we rely on **educational standards** as guidelines.

   - ***Common Core State Standards* - CCSS.** Established in 2009, the common core is a set of college and career-ready standards for children in

kindergarten through 12th grade in English Language and Mathematics [45]. Some English Language topics include Phonetics and Metaphors, while some Mathematics topics are Geometry and Ratios.

- ***Next Generation Science Standards* - NGSS.** The NGSS are science content standards that set expectations of what children should know in science related subjects such as Chemistry, Biology, and Solar System [44]. The NGSS enables teachers the flexibility to design classroom experiences that enhance children's interests in science.

- ***Idaho Content Standards* - ICS.** ICS details what children in Idaho Public schools should know at the end of each grade in subjects including Sciences, English, and Social Studies [85]. We depend on ICS to identify topics relating to children's Social Studies subjects.

We use the aforementioned educational standards to identify educational topics specific to each grade level, as well as educational concepts aligned with these topics. We then use this information to manually create *hasA* and *sameAs* topic-concept pairs for different K–12 grade levels. [9] We explicitly create 4,100 relationships of the form:

- Grade level "hasTopic" Topic

- Topic "hasConcept" Concept

- Concept "sameAs" Concept

[9] "*hasA*" is a relationship type that is used to indicate that an instance is a reference to another instance of the same class or category. On the other hand, "*sameAs*", as defined by "owl", is used to declare two items or instances to be identical [78].

We include in Figure 4.5, a sample of relationships used for creating $\mathbf{Edu_{KB}}$. We are aware that relations in $\mathbf{Edu_{KB}}$ identified using the aforementioned standards may not be sufficient to explicitly infer connections, e.g., while "biodiversity" may be relevant to "science" and "biology", this connection may not explicitly exist. To this effect, we also rely on WikiBooks, an external knowledge base that allows us identify concepts relevant to children elementary school subjects.[10] WikiBooks is a Wikipedia community hosted for the purpose of creating a free library of educational text books. Connections in WikiBooks are inferred based on page categories, as illustrated in Figure 4.6. In this case, web pages about "stars" and "universe", are related based on their higher level page category "planet", which would also be related to "science". Thus, we use this information to indicate the relevance of concepts "stars" and "planet" to "science". In using WikiBooks, we first extract all the web pages categorized for Kids and Elementary School Students. Following this, we also create 1,672,000 *hasA* relations among these web pages based on their inter-connectivity, i.e., how they link among each other. After creating the link pairs, we extract their page titles. To ensure that we have informative relation pairs, we first extract tokens from their titles and strip them of stop words, characters, and words not in the English dictionary. Finally, we use these concept pairs to inform relations already in $\mathbf{Edu_{KB}}$.

Contextually Relate Topics and Concepts. We take advantage of a deep learning based methodology in order to relate concepts and topics that occur in the same context. For instance, the concept "shapes" should be contextually closer to the topic "geometry" than to the topic "ratio", even though these topics

[10]https://en.wikibooks.org/wiki/Wikijunior

A	RelationType	B
Grade 3	**hasTopic**	Geometry
Geometry	**hasConcept**	Angle
Angle	**sameAs**	Angles

Figure 4.5: Sample of pairs with *hasA* and *sameAs* relation type in $\mathbf{Edu_{KB}}$.

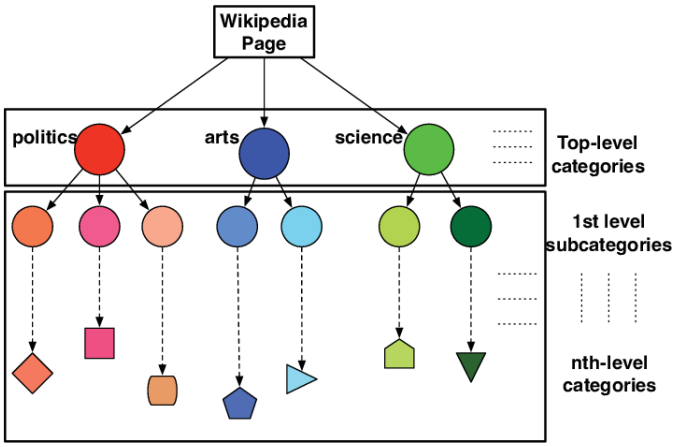


Figure 4.6: A description of Wikipedia category hierarchy [141].

are both related to the subject "Mathematics". To capture contextual relationships among pairs in $\mathbf{Edu_{KB}}$, we depend upon a model based on **word2vec** [117], which we use to create an embedding[11] for each concept and topic in $\mathbf{Edu_{KB}}$. Word2vec is a state-of-the-art algorithm introduced by Mikolov et al. [117], that can be used to create numerical embeddings for representing words in a textual document [162]. Word2vec relies on a two-layer neural network, for processing un-labeled documents. We use the skip-gram architecture for building our word2vec model. In this case, the word2vec model takes topics or concepts as an input, and then tries to predict its neighbors or the surrounding context topics and concepts. An important benefit of using word2vec is that it

[11] An embedding is a mapping from discrete objects, such as words, to n-dimensional vectors of real numbers.

allows us to group vectors of similar topics or concepts together.

In order to train the educational word2vec model, we rely on the pairs that we used in designing $\mathbf{Edu_{KB}}$. We use each of the pairs as input to the model, so as to get as an output, each topic or concept, along with its respective embedding.[12] In our work, we capture contextual similarity between relation pairs in $\mathbf{Edu_{KB}}$ using the cosine similarity [23] of their embeddings. By doing this, we are able to identify contextually similar educational topics and concepts, as well as the K–12 grade level they relate to. We show an example of a sample topic and concept relation using $\mathbf{Edu_{KB}}$ in Figure 4.7.

2. **General Perspective Knowledge Base.** In our work, we take advantage of Probase [158], which is a general-purpose knowledge base, in order to identify concepts in resources from a general perspective. Probase is a probabilistic semantic network containing over 12 million instances, i.e., terms, 5 million unique concepts and 85 million isA relationships. This knowledge base was built from public data–billions of web pages, and is made available for only research purposes. Compared to other popular knowledge bases, such as WordNet [62] and FreeBase [33], we use Probase because it provides more fine grained concepts and covers a variety of worldly facts, as it is trained on a huge number of web data [158].

Setup and Candidate Topics and Concepts Generation Process

In this section, we detail the strategy we use in extracting relevant topics and concepts from R.

[12] It is important to note that we set the number of dimensions for our embeddings to 300, as prior work have empirically verified this dimension size to be effective for capturing contexts [155, 157].

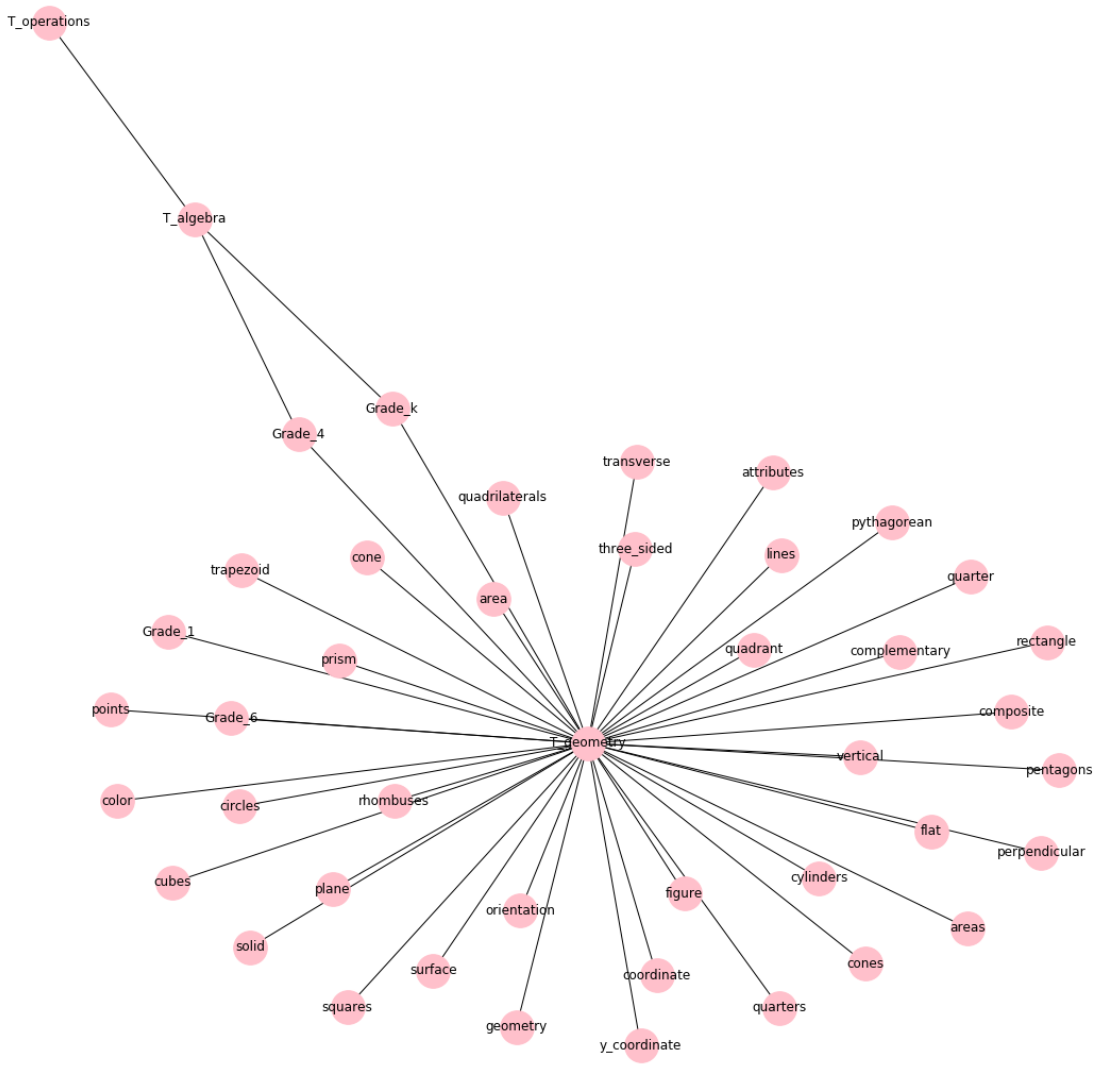


Figure 4.7: A sample of *hasA* relations among K–12 topics, concepts, and grade levels defined in $\mathbf{Edu_{KB}}$ for the topic "Geometry".

Candidate Term Extraction. In order to extract educational topics and concepts from R using $\mathbf{Edu_{KB}}$, it is imperative that we first identify informative terms from R_{con}, which allows us map R to the K–12 curriculum. To do so, we treat nouns from R_{con} as the representative terms for capturing the content.

This has proven to be an effective strategy, which we empirically verified in Section 5.2.4. To identify the nouns, we pre-process R_{con}, extract all its tokens, and tag them with their respective PoS. We then use the candidate selection approach proposed by the authors in [162]. To reduce the number of possible term candidates, the authors compare the input (representative terms in our case) against an index of all their relevant entities (in our case we compare these terms against entries in $\mathbf{Edu_{KB}}$). We know that some terms may not explicitly exist in $\mathbf{Edu_{KB}}$ as a result of non-exact match or in having different variations, i.e., plurals vs. singulars, and verb tenses. To address this limitation, we normalize each term extracted from R_{con}, by replacing it with the one in $\mathbf{Edu_{KB}}$ that share the same root form with it. For instance, if the term "dimensional" is extracted from R_{con} and does not exist in $\mathbf{Edu_{KB}}$, whereas "dimension" does, based on our normalization strategy, "dimensional" will be replaced with "dimension". Following this, we select the top-20 most frequent informative terms in R_{con}. [13]

- **Educational Topic Extraction.** In using $\mathbf{Edu_{KB}}$, our aim is to extract relevant K–12 educational topics from R_{con}. To do so, we extract the most similar candidate topics for each informative term extracted from R_{con} using the cosine similarity between the embedding of each term and that of topics in $\mathbf{Edu_{KB}}$. We treat as the most similar topics, those that have a similarity score with respect to each term, above 0.85. We set the similarity threshold $\alpha = 0.85$, as we found that the contextual similarity among most relations in $\mathbf{Edu_{KB}}$ were between 0.85 and 1.

[13] We empirically verified in 5.2.4 that using the top-20 nouns was best suited for this strategy.

To ensure that we capture relevant topics out of the candidate topics for each term, we select the ones that are inter-connected, i.e., the ones that had overlaps among the topic candidates identified for each term. We show an example of the process in which we extract educational topics in Figure 4.8.

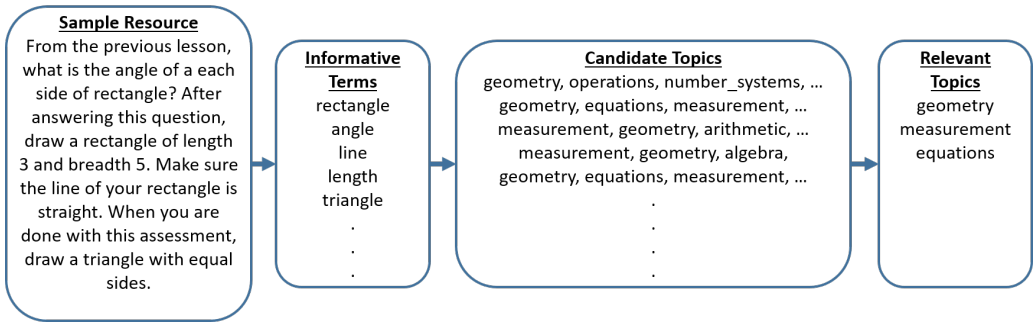


Figure 4.8: An example of educational topics extracted from a sample resource.

Educational Concept Extraction. Other than being able to identify educational topics, $\mathbf{Edu_{KB}}$ allows us to map resources to concepts, i.e., their main ideology from an educational perspective. To identify concepts in R_{con}, we also use the informative terms extracted from R_{con}, and extract the most similar concepts each of these terms map to. We then select the top-20 most frequent concepts across the total collection of candidate concepts identified for these terms.

Probase Concept Extraction. As discussed previously, we depend upon $\mathbf{Edu_{KB}}$ solely to identify K–12 topics and concepts from resources. Using $\mathbf{Edu_{KB}}$ alone may not be comprehensive enough to allow us to identify what a resource as a whole is about, as it focuses on just the educational domain. We know that resources may have more general concepts, hence, it is important that we quantify the degree to which general concepts identified in a resource are

more dominant than the educational ones. To identify these general concepts, we turn to Probase. Unlike our word2vec-based strategy that is able to identify concepts using the contextual similarity approach, Probase identifies concepts using a probabilistic approach. In this case, concepts for the word "orange", such as "fruit", "movie", and "color", are selected and ranked based on the probability of each of them being identified with the word "orange". For instance, the probability scores of "fruit", "movie", and "color" are 0.4, 0.1 and 0.5 respectively.

In using Probase, we identify the top concepts identified by their probabilistic scores for each term extracted from R_{con}. It is important to note that we removed noisy candidate concepts, i.e., high level concepts such as "objects", "elements", and "thing", that tend to appear for all instances in Probase. We then follow the same concepts selection strategy detailed for $\mathbf{Edu_{KB}}$, in order to extract general concepts from R_{con} using Probase.

Qualitative Indicators for Examining Resource Educational Value

To quantify the degree to which R pertains to the education-domain, we create several qualitative data points based on R_{con}. Based on an assessment we conducted using resources in $\mathbf{Diverse_{res}}$ (see Section 4.4), we found that in most scenarios, non-educational resources did not map to educational topics, which was also the case for educational concepts. This prompted us to create four data points:

DP_18 *Unique Number of Educational Topics*, which we measure as the distinct count of educational topics extracted from R_{con} using $\mathbf{Edu_{KB}}$.

DP_19 *At Least One Topic*, which is a binary data point indicating if an educational topic was extracted from R_{con}. In this case, 1 indicates that a topic was extracted, and 0 indicates otherwise.

DP_20 *Unique Number of Concepts*, which we compute as the distinct count of educational concepts identified from R_{con} using $\mathbf{Edu_{KB}}$.

DP_21 *Proportion of Non-Educational Key Terms*, which we measure as the number of informative terms extracted from R_{con} that did not map to concepts in $\mathbf{Edu_{KB}}$ divided by the total number of its informative terms.

We create three additional data points by examining R_{con} using Probase.

DP_22 *Unique Number of Probase Concepts that are Educational*, which we compute as the number of concepts extracted from R_{con} using Probase, that mapped to concepts in $\mathbf{Edu_{KB}}$.

DP_23 *Significance of R_{con} from Educational domain versus General Perspective*, which we compute as the number of Probase concepts extracted from R_{con} that mapped to concepts in $\mathbf{Edu_{KB}}$ divided by the number of Probase concepts extracted from R_{con} that did not map to $\mathbf{Edu_{KB}}$.

DP_24 *Proportion of Probase Concepts that are Educational*, which we measure as the number of of concepts extracted from R_{con} using Probase that exist in $\mathbf{Edu_{KB}}$ divided by the total number of Probase concepts extracted from R_{con}.

4.7 Integrate Qualitative Aspects

To simultaneously integrate evidences from each aforementioned perspective for ranking purposes (summarized in Table 4.4), **KiSuRF** adopts Gradient Boosting Regres-

sion [60], a supervised machine learning algorithm useful for both classification and regression tasks. Gradient Boosting Regression refers to an algorithm where decision trees are used with a boosting strategy. A decision tree is a predictive modeling approach that uses a tree-like graph or model of decisions in order to make decisions. Each tree-node represents a question and the branches are the consequences of answers to that question. Like the decision tree, each node in the Gradient Boosting strategy is assigned a label, which informs the decision process for prediction or classification purposes.

Table 4.4: Overview of **KiSuRF**'s ranking data points.

Data Point	**Analysis Type**	**Definition**
DP_16	Readability	Spache readability index of resource content
DP_17	Objectivity	Objectivity score of resource content
DP_18	Educational value	Unique number of educational topics in resource content
DP_19	Educational value	At least one topic in resource content
DP_20	Educational value	Unique number of concepts extracted from resource content
DP_21	Educational value	Proportion of non-educational key terms in resource content
DP_22	Educational value	Unique number of probase concepts extracted from resource content that are educational
DP_23	Educational value	Significance of resource content from educational domain vs. general perspective
DP_24	Educational value	Proportion of probase concepts extracted from resource that are educational

In the Gradient Boosting Regression, predictors on each node of the decision tree are made sequentially, as opposed to independently, like the Random Forests [127]. The idea behind this algorithm is that subsequent predictors learn from the previous mistakes of previous ones, hence, taking less time to reach close to the actual predictors. As a result, each observation have unequal probabilities of occurring in subsequent models. Some of the benefits of Gradient Boosting models as mentioned in [60], include that they:

- Are efficient

- Reduce bias
- Achieve high accuracy when compared to other machine learning techniques
- Generate sequential predictors, i.e., a prediction task occurs by predicting the next items of a sequence [146]

As part of the ranking process, **KiSuRF** represents each resource as a tuple capturing all qualitative data points considered for ranking (shown in Equation 4.7), in order to build a Gradient Boosting Regression model through training.[14]

$$R = \langle DP\_15, DP\_16, DP\_17, ..., DP\_24 \rangle \tag{4.7}$$

where DP_15, DP_16, ..., DP_24 represent data points that **KiSuRF** considers for ranking purposes. The output from the Gradient Boosting model, is a predicted score, which **KiSuRF** uses for prioritizing suitable resources. Based on this, resources with higher scores will be positioned earlier on the ranked list.

[14]We empirically verified in Section 5.2.7 that a ranking model trained on Gradient Boosting best suited **KiSuRF** for ranking suitable resources.

CHAPTER 5

EVALUATION

In order to validate the correctness and effectiveness of **KiSuRF**, we performed a comprehensive study using several metrics and datasets, which we summarize in Table 5.1. We evaluated each aspect considered in the design of **KiSuRF** in order to minimize errors in each design step, and compared **KiSuRF**'s filtering strategy with SEs under study, as well as their safe search counterparts (introduced in Section 3.3). We conducted a study to empirically verify the usefulness of simultaneously integrating evidences from multiple perspectives for ranking suitable resources. Finally, we performed an empirical analysis on the overall ranking strategy of **KiSuRF** using a dataset comprised of search sessions for children in the $\mathbf{4^{th}}$ and $\mathbf{5^{th}}$ grades that conducted searches in the school setting. In doing this we investigated the position assigned to suitable resources on a ranked list of resources, with and without complementing Google and Bing with **KiSuRF**.

5.1 Experimental Setup

5.1.1 Evaluation Strategy

We depend upon the **Train-Test Split** strategy on datasets used in designing and developing **KiSuRF**. The train-test split [46] refers to a technique used for ensuring that the evaluation conducted on a model when using a dataset will generalize well to

Table 5.1: Summary of the datasets used in validating the effectiveness and correctness of **KiSuRF**.

Dataset	Sources	Type	Size
Diverse$_{DS}$	kaggle.com/patjob/articlescrape intouchweekly.com teacherspayteachers.com/	Web Resources	5,000
Edu$_{TOPICS}$	teacherspayteachers.com/	Text Book Descriptions	3,000
Edu_Search$_{DS_A}$ **Edu_Search$_{DS_B}$**	irlab.boisestate.edu/admincsv.php dmoz.org	Relevant & Irrelevant Set of Web Resources	1,120
Movie$_{DS}$	cs.cornell.edu/people/pabo/movie-review-data/	Web Resources	10,000
Read$_{DS}$	idiglearning.net/	Web Resources	40,000
Kids_Safe$_{DS}$	alexa.org dmoz.org nohatespeechmovement.org/	Web Resources	14,000
WebObj$_{DS}$	datacommons.org/docs/download.html kaggle.com/snapcrack/all-the-news u.cs.biu.ac.il/∼koppel/BlogCorpus.htm	Web Resources	240, 000

unknown independent data. This technique randomly splits the data into train and test sets, each of which contains $N\%$ and (100 - $N\%$) of the data instances. For every experiment in this Chapter, we consistently use 80% for training, and the remaining 20% for testing purposes.

5.1.2 Metrics

For validating **KiSuRF**'s design, we depend upon well known metrics which we discuss below.

- ***Accuracy*** [46], which measures the fraction of correctly classified instances out of the observed cases, i.e., true positive and true negatives from observed instances and is computed using Equation 5.1.

$$Accuracy = \frac{CorrectlyClassifiedInstances}{TotalInstances} \tag{5.1}$$

where an instance in our case is a resource that is treated as being correctly classified if the predicted label matches the actual label of that resource.

- ***Mean Reciprocal Rank (MRR)*** [46], which is the average of the reciprocal rank at which the first relevant resource was retrieved, for a given set of queries. It quantifies the average position of resources a user scans through before locating the first relevant one, as shown in Equation 5.2.

$$MRR = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{rank_i} \tag{5.2}$$

where $\frac{1}{|Q|}$ is a normalization factor, Q is a sample of queries, and $rank_i$ is the position of the first relevant resource for the i^{th} query.

- ***Normalized Discounted Cumulative Gain (NDCG)*** [46], which measures the precision and ranking of a given strategy, while penalizing relevant resources that are positioned lower on a ranked list. The NDCG is computed as in Equation 5.3.

$$nDCG = \frac{DCG}{IDCG} \tag{5.3}$$

DCG denotes the total gain accumulated at a particular rank N and is computed using Equation 5.4, while IDCG is the ideal DCG (rank) which is used as a normalization factor, and is computed using the same Equation.

$$DCG = rel_1 + \sum_{i=2}^{N} \frac{rel_i}{log_2(i)} \tag{5.4}$$

where $log_2(i)$ is a penalization factor, N is a ranking position, rel_i is the graded relevance of a resource retrieved at position i. Note that rel_i is 1 if the resource is relevant and 0 otherwise.

- ***Recall and Precision@K*** [46], where precision@k is the proportion of resources in a top-k set (ranked list) that are relevant and is computed using Equation 5.5.

$$Precision@K = \frac{NumberOfRelevantResources@K}{TotalNumberOfResourcesAtK} \tag{5.5}$$

Recall is the fraction of relevant resources that are retrieved and is computed using Equation 5.6.

$$Recall = \frac{|\{RelevantResources\} \bigcap \{RetrievedResources\}|}{|\{TotalRelevantResources\}|} \tag{5.6}$$

where $\bigcap$ indicates an intersect / overlap.

- ***Hits*** [81] is a common effectiveness metric used in Information Retrieval, which measures the match between a relevant and reference set. In this case, the hit value is 1 if at least one relevant item or resource is found on a list of resources, and 0 otherwise.

- ***F-Measure*** [46] is an effectiveness measure which is based on precision and recall that is used for evaluating classification performance. It is defined as the harmonic mean of precision and recall and computed as in Equation 5.7.

$$F - Measure = \frac{2 \times Recall \times Precision}{(Recall + Precision)} \tag{5.7}$$

where Recall is computed as in Equation 5.6 and precision is computed using Equation 5.8

$$Precision = \frac{|\{RelevantResources\} \bigcap \{RetrievedResources\}|}{|\{TotalRetrievedResources\}|} \tag{5.8}$$

- ***False Negative Rate (FNR)*** [46], which is the measure of the percentage of positive outcomes that were wrongly predicted as negatives, in a classification task.

- ***Root Mean Squared Error (RMSE)*** [46], which is a quadratic scoring formula that measures the difference between predicted and actual scores, of which the values are then squared and averaged over the given sample [46].

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(p_i - a_i)^2} \tag{5.9}$$

where p_i and a_i are the predicted and actual values respectively, and n is the number of samples.

5.1.3 Test of Statistical Significance

In order to validate the performance of **KiSuRF**'s filtering and ranking strategies, we use the paired t-test of **statistical significance** [114]. The t-test is used for comparing the difference between the mean of two samples, as well as the significance of these differences.

5.2 Examine Correctness of KiSuRF

To determine the correctness of each strategy considered in the design of **K**i**S**u**RF**, we conducted several assessments. We do this in order to prevent error mitigation at different stages of **K**i**S**u**RF**'s design.

5.2.1 Which readability formula performs best for estimating the reading levels of web resources?

Since **K**i**S**u**RF** requires prioritizing resources that children can comprehend, we performed this experiment in order to demonstrate the validity of the readability prediction strategy **K**i**S**u**RF** adopts for estimating the reading levels of resources. We explored a number of traditional readability formulas: Flesch Kincaid [100], Dale-Chall [35], SMOG[115], and Spache [145], which have demonstrated applicability for web readability prediction tasks [25, 39, 151]. The Flesch formula [58] uses shallow features, such as average length of words and average length of sentences, in order to estimate the level of difficulty of a document. This formula was later adapted to the American educational grading system, as Flesch Kincaid [100] (Equation 5.10), in order to estimate the complexity of resources with respect to a child's grade level.

$$FK_R = 0.39 \times (\frac{TW}{TS}) + 11.8 \times (\frac{TW_{syllables}}{TW}) - 15.59 \tag{5.10}$$

where TW refers to the total number of words in R, TS refers to its total number of sentences, and $TW_{syllables}$ refers to its total number of syllables.

The SMOG[115] readability formula (Equation 5.11) estimates the reading level of a document by using a non-linear strategy that combines the number of complex

terms in a document along with the number of sentences, in order to determine its complexity. It treats complex terms as those ones that have three or more syllables.

$$SMOG_R = 1.0430 \times \sqrt{TCW \times \frac{30}{TS}} + 3.1291 \tag{5.11}$$

where TCW is the total number of complex words in R, i.e., words having more than three syllables, and TS refers to its total number of sentences.

Other alternatives such as Dale-Chall [35] and Spache [145], rely on shallow features, in addition to a pre-defined list of simple words. Dale-Chall (Equation 5.12) uses a list of manually generated 3,000 simple terms. This formula estimates the frequency of simple terms in a document, as well as the average sentence length as an indicator of its complexity.

$$DC_R = 15.79 \times (\frac{NDW}{TS}) + 0.0496 \times (\frac{TW}{TS}) \tag{5.12}$$

where NDW refers to the number of words in R that does not exist in the Dale-Chall pre-defined simple words list, TS refers to the total number of sentences in R, and TW refers to the total number of words in R.

Spache [145] (Equation 5.13) estimates the reading level of a document based on the sentence length and number of difficult words. In this case, difficult words are the ones that do not exist in a defined list of Spache's 3,000 easy words.

$$SP_R = (0.121 \times ASL) + (0.082 \times PDW) + 0.839 \tag{5.13}$$

where ASL is the average sentence length in R and PDW is the percentage of words in R that do not exist in the Spache list of simple words.

Given the lack of standard datasets for this analysis, using the IDLA educational data source [86] described in Section 3.2, we created $\mathbf{Read_{DS}}$, a dataset comprised of 40,000 educational resources that have been labeled with their respective reading levels (in the K-12 range). To measure the predictive performance of the aforementioned readability formulas, we used each of them to estimate the reading level of all resources in $\mathbf{Read_{DS}}$, and depend upon the RMSE metric (described in Section 5.1.2) to measure their error rate. In using this metric for validation, the actual label, i.e., a_i, is the pre-defined grade level for these resources, while the predicted label, i.e., p_i, is the reading level estimated by each of these readability formulas.

Although the Spache formula was prone to errors in estimating reading level of resources in $\mathbf{Read_{DS}}$, it outperformed its counterparts (see Table 5.2). We anticipated that this was going to be the case, as prior research has demonstrated the effectiveness of Spache in predicting the readability of web resources [113]. This analysis offers supporting evidence on the usefulness of Spache in estimating the reading level of web resources, and hence, the formula of choice in designing **KiSuRF**.

Table 5.2: Comparison of the correctness of traditional readability formulas in estimating the reading level of resources, using the RMSE.

Readability Formula	**RMSE**
Dale-Chall	6.350
Flesch Kincaid	5.918
SMOG	4.959
Spache	**4.213**

5.2.2 Can KiSuRF identify objective resources?

Recall that one of the main qualitative aspects that **KiSuRF** considers is in examining the objectivity of resources. We therefore conducted this study to validate the

performance of our strategy for prioritizing objective resources over subjective ones. Being that we adopt the language-model-based strategy introduced by the authors in [92], we replicated their experiment and investigated the effectiveness of this strategy in assigning objectivity scores in labeled datasets. The authors use the same reference language model described in Equation 4.2 and Equation 4.3, but instead implement a two staged smoothing method for the resource language model using Equation 5.14.

$$P(w_i|\hat{\theta}_d) = (1-\lambda)\frac{c(w,d)+\mu P(w|C)}{\sum_{w_i \varepsilon Z} c(w_i,d)+\mu} + \lambda P(w|C) \tag{5.14}$$

where for a document d, Z is its vocabulary, $P(w|C)$ is the probability of word w in a reference collection C by maximum likelihood estimation, while μ and λ are smoothing parameters. In their experiment, $\mu = 8$ and $\alpha = 0.9$.

For conducting this assessment, we rely on two datasets comprised of subjective and objective resources. $\mathbf{Movie_{DS}}$, is a dataset created by the authors in [122], which is comprised of 5,000 movie review sentences that they label as subjective, as well as sentences extracted from movie plots labeled to be objective. Since in our case we examine web resources, which most likely have longer texts and different vocabulary type than movie reviews and plots, we were interested in a dataset consisting of web resources. We found no labeled dataset for this purpose, and thus created one instead, which we refer to as $\mathbf{WebObj_{DS}}$. In this dataset, we include 120,000 blog articles we extracted from a blog authorship corpus [137], containing posts written by 19,320 bloggers at blogger.com. We labeled these blog articles as subjective, as blog posts are usually opinionated. We also include in $\mathbf{WebObj_{DS}}$ a number of objective resources: 50,000 news articles from 15 American publications compiled on a News dataset [91], 5,000 fact checked resources created by the authors in [17], 20,000 Wikipedia articles

and 40,000 educational resources out of the IDLA data source introduced in Section 3.2.[1]

Based on the aforementioned objectivity scoring strategy, we trained a language model using instances in $\mathbf{Movie_{DS}}$, which we refer to as $\mathbf{lm_{mr}}$, and tested on the remaining ones (we use 80:20 split ratio as detailed in Section 5.1). To quantify the ranking performance of this strategy, we depend on the precision@K metric (described in Section 5.1.2). We use precision@K since it allows us to capture the proportion of objective resources identified as we explore a ranked list of resources. While $\mathbf{lm_{mr}}$ was effective in assigning objectivity scores to the instances in $\mathbf{Movie_{DS}}$, we observed that this model yielded poor performance when applied to resources in $\mathbf{WebObj_{DS}}$ (see Figure 5.1), i.e., it was not able to distinguish and rank higher known objective resources. We infer that this was due to limited vocabulary in $\mathbf{Movie_{DS}}$; diverse topical information; as well as longer texts in resources in $\mathbf{WebObj_{DS}}$. As a result, we trained a new language model, $\mathbf{lm_{wb}}$, using instances in $\mathbf{WebObj_{DS}}$. For this purpose, we split $\mathbf{WebObj_{DS}}$ into train and test sets. As shown in Figure 5.1, $\mathbf{lm_{wb}}$ obtained better results when used to estimate objectivity scores on instances in $\mathbf{WebObj_{DS}}$–nearly as good as that of the ideal ranking. In most cases, $\mathbf{lm_{wb}}$ is able to position objective resources higher than the subjective ones (which aligns with our aim). From this experiment, we empirically showcase the validity of our approach that examines word distributions in resources in order to quantify the likelihood of them being objective.

[1]We made sure to exclude this data for other analysis in order to reduce bias.

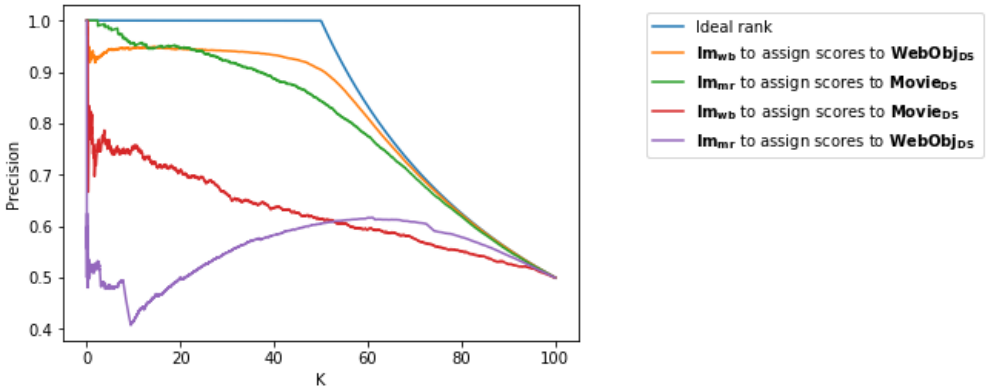


Figure 5.1: Comparison of objectivity scoring strategy on different datasets.

5.2.3 Do topics and concepts in $\mathbf{Edu_{KB}}$ abide by pre-defined contextual similarities?

We conducted an empirical study to validate the correctness of $\mathbf{Edu_{KB}}$, in relating topics and concepts that are contextually similar by subjects or grade levels they relate to. In addition to quantifying closeness among pairs, this experiment is also essential for evaluating the quality of our embeddings. We rely on the cosine similarity between embeddings for pairs in $\mathbf{Edu_{KB}}$, in order to capture their similarity.

As shown in Figure 5.2, the similarity score among pairs were skewed towards 1.0, thus, indicating their closeness in the vector space. This was anticipated, as concepts and topics that occur in the same context will usually have a high level of similarity. To gain more insights into the closeness of contextually similar topics, we investigated how they were clustered. This led us to empirically verify that contextually similar topics are grouped together, when compared to other topics. As shown in Figure 5.3, topics such as "Geometry", "Rate", "Proportions", and "Measurement", which are relevant to Mathematics subject, are closer to each other, when compared to topics such as "Earth Systems", "Solar", and "Chemical", that are relevant to children's

Science subjects. This analysis informs us about the quality of the embeddings, as well as the correctness of using $\mathbf{Edu_{KB}}$ for extracting concepts and topics that are relevant to the educational domain.

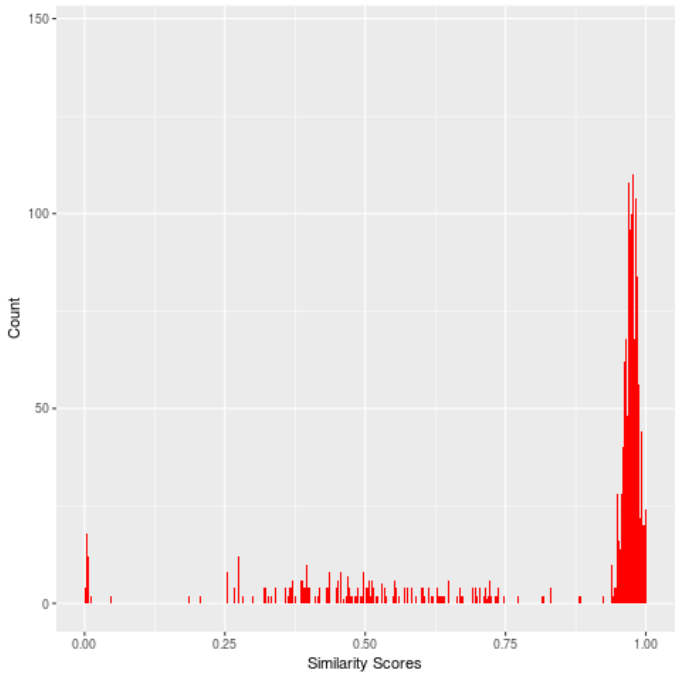


Figure 5.2: Pattern of contextual similarity among pairs in $\mathbf{Edu_{KB}}$.

5.2.4 Is KiSuRF able to identify relevant educational topics in resources?

Being that a key aspect that **KiSuRF** considers for prioritizing resources is based on their educational value, it is imperative that we extract relevant K–12 topics from resources. Thus, we conduct this experiment to validate our educational topic extraction process. In achieving this task, we extract educational topics from resources, using the strategy discussed in Section 4.6.3.

A suitable dataset for conducting this experiment would be one that is comprised of resources labeled with corresponding K–12 educational topics. However, to the

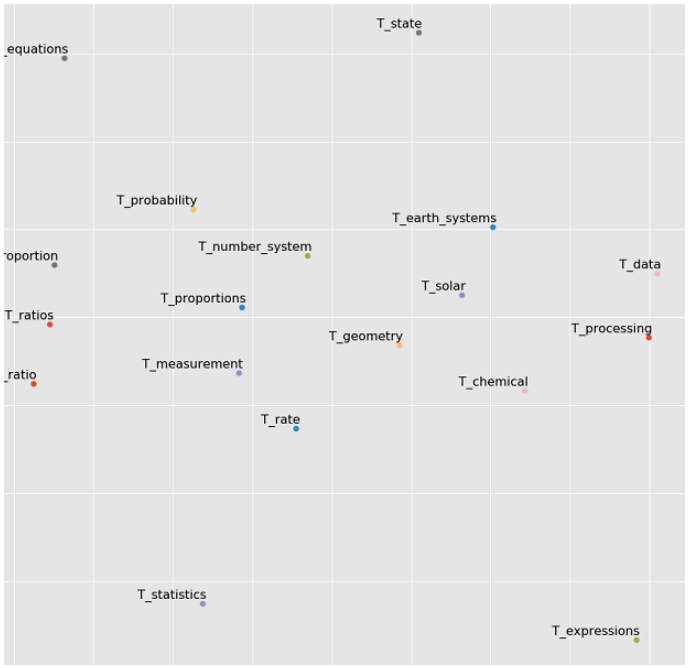


Figure 5.3: Sample of a grouping which shows contextually similar educational topics in $\mathbf{Edu_{KB}}$.

best of our understanding, such a dataset does not exist, prompting us to create one. To do so, we rely on information from Teachers-Pay-Teachers (TPT), an educational resource [3]. TPT includes over 3 million school resources (identified based on educational standards such as CCSS and NGSS) that have been created by K – 12 educators. In using this resource, we extracted descriptions of educational books designed for children in the K – 12 curriculum, along with their pre-defined subjects, topics, grade levels, and CCSS concepts, which we use to build a dataset: $\mathbf{Edu_{TOPICS}}$. We only used book descriptions for analysis purposes, as full content was available only for a fee.

Based on the topic extraction strategy detailed in Section 4.6.3, an initial step requires identifying representative terms in resources, which we would then use to

select candidate topics. For this purpose, we examined several keyword extraction strategies, including TagMe [56] and DBpedia Spotlight [116]. These strategies are known to annotate resources and map them to concepts, which they define as the corresponding Wikipedia pages the annotated terms link to. Upon analyzing TagMe and Spotlight, we found that they did not always perform well with longer texts and were not efficient in annotating texts. We compared these strategies with simply extracting top-N nouns. To validate these approaches for leading to the extraction of relevant educational topics, we used the Recall and Hits metric (defined in Section 5.1.2) as the suitable measures, being that they are applicable for evaluating the identification of relevant items in a list. In this case, we treat as ground-truth all educational topics assigned for each instance in $\mathbf{Edu_{TOPICS}}$. As shown in Table 5.3, using 20 informative keywords, i,e., nouns, outperformed both TagMe and Spotlight in terms of efficiency. The average time required to extract informative terms using nouns was 99.5% and 99.7% more efficient than Spotlight's and TagMe's, respectively. Moreover, by using nouns, we were able to find more relevant educational topics, when compared to the aforementioned keyword extraction strategies, in terms of Recall and Hits. We infer that this was the case, as nouns are known to capture representative information in documents. Evidences from this analysis showcase why we rely on nouns for identifying representative terms from resources, as well as the correctness of our strategy for identifying relevant educational topics in resources.

Table 5.3: **K**i**S**u**RF**'s performance in extracting relevant educational topics from resources.

	Number of Nouns Extracted					**Keyword Extraction Strategies**	
	Top 2	**Top 5**	**Top 10**	**Top 15**	**Top 20**	**Spotlight**	**TagMe**
Recall	0.07	0.25	0.43	0.55	**0.61**	0.47	0.36
Hits	0.08	0.30	0.52	0.64	**0.71**	0.57	0.45
Average Time (secs)	**0.0254**	**0.0254**	**0.0254**	**0.0254**	**0.0254***	5.09	7.89

5.2.5 Are educational resources assigned higher educational value scores when compared to non-educational resources?

In order to ensure that we accurately quantify the degree to which resources are educational, we conduct an empirical study. To do so, we first create a dataset comprised of diverse resources, which we refer to as $\mathbf{Diverse_{DS}}$. Among these resources, we included 3,000 educational resources, which we extract from $\mathbf{Edu_{TOPICS}}$ (described in Section 5.2.4). We treat as non-educational resources, 600 celebrity news that we extracted from InTouch Magazine [1]. We also include 600 resources each pertaining to Sports, Economy, Weather, and Politics, that we extracted from the dataset in [2]. This dataset is comprised of diverse news information from sources such as CNN, Economic times and NyPost.

To investigate the performance of our strategy in assigning educational value scores to resources in $\mathbf{Diverse_{DS}}$, we compare the average educational value scores we assign to educational resources, as opposed to the non-educational ones. In this context, we treat the educational resources in $\mathbf{Diverse_{DS}}$ as the relevant ones, and aim to assign them with higher educational value scores, as in a real world scenario, they should be ranked higher than resources that do not satisfy educational information needs.

We first extract the qualitative data points described in Section 4.6.3, and split $\mathbf{Diverse_{DS}}$ into train and test sets. For this experiment, we explored several aggregation models: Linear Regression [119], Support Vector Machines [43], Gradient Boosting Regression [60], and MultiLayer Perception Regression [80]. We evenly distributed resources for each source type, i.e., sports, magazine, education, weather, economy, and politics, in the test set. We then compute the average educational value score assigned to each source type by the aggregation models. As shown in

Figure 5.4, educational resources were consistently assigned higher educational value scores. Other source types with high educational value scores pertained to Weather and Economy, which was expected, as these resources usually include informative content and keywords that relate to school work. As anticipated, resources relating to Magazine and Politics had the lowest scores, which we infer was the case, as these sources usually have content that are entertainment-based and usually do not relate to the educational domain. Results from this assessment provides insights on the type of information sources that are educative and non-educative for kids. Moreover, we empirically verified the performance of our strategy in identifying resources of high educational value.

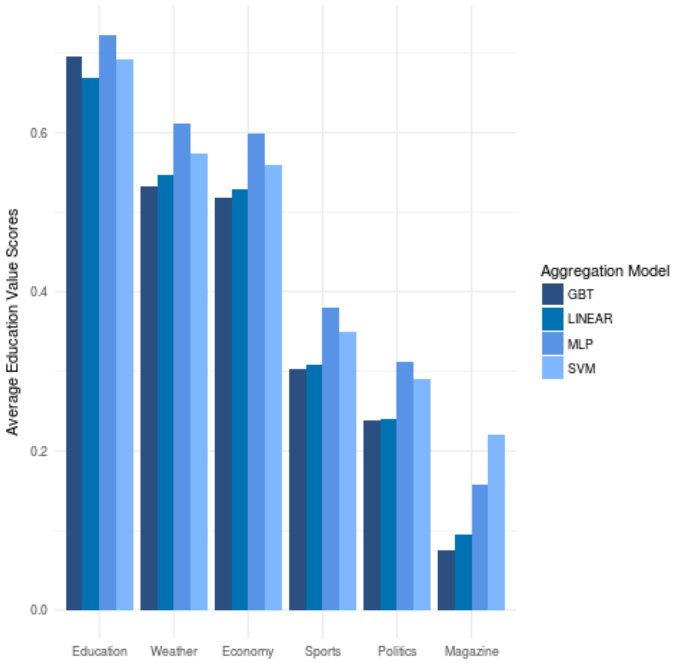


Figure 5.4: Comparison of educational value scores assigned to diverse information sources. Educational resources had the highest value.

5.2.6 Which model is best suited to aggregate data points for KiSuRF's filtering purposes?

In designing **KiSuRF**'s filtering strategy, we explored different models that have been successfully used for web filtering tasks: Logistic Regression, Random Forests, MultiLayer Perception classifier, and Gradient Boosting Trees. To conduct this experiment, we use the same labeled kids safe and unsafe resources described in Section 4.4, to build a dataset, comprised of 7,000 labeled resources, which we refer to as **Kids_Safe$_{\mathbf{DS}}$**. We train each of the compared models using the train:test split ratio discussed in Section 5.1, and predict for each test instance, a label that indicates if they should be filtered or retained.

We compare the performance of the aforementioned models in filtering inappropriate resources, while retaining the ones that contain educational terminologies. In this case, kids safe resources in **Kids_Safe$_{\mathbf{DS}}$** are treated as the relevant ones and should be retained, while resources containing hate-speech and sexually explicit content are meant to be filtered. Based on these evaluations, we found that a model based on Random Forests best suited this task (see Table 5.4). In fact, this model yielded an accuracy of 91.6% and f-score of 92%. Although the Logistic Regression and Multi-layer Perception models achieved a higher accuracy and equally had an f-score of 92% (same as the Random Forest), they however, resulted in higher false-negative rate than that of Random Forest. In this context, a lower false-negative rate indicates that a model is less likely to disregard resources that are kid-safe and education-relevant. On the other hand, the Gradient Boosting Tree classifier obtained the lowest performance across all metrics. The improvement in Random Forests, in terms of accuracy and f-score are not statistically significant with respect

to counterparts considered in this study, as determined using a paired t-test with a confidence of $p < 0.05$. However, given that Random Forests is able to retain resources that are both safe for children and education-relevant, we determined that this model best suits **K**i**S**u**RF**.

Table 5.4: Performance of different models considered in the design of **K**i**S**u**RF**'s filtering strategy.

Model	**Accuracy**	**FNR**	**F-Score**
Gradient Boosting Tree	91.4%	3%	91%
Logistic Regression	**92%**	4%	**92%**
Multi-layer Perceptron	**92%**	4%	**92%**
Random Forest	91.6%	**2%**	**92%**

5.2.7 Which model is best suited for simultaneously integrating aspects considered for ranking?

To ensure that we prioritize suitable resources, it is essential that we use an aggregation model that best fits **K**i**S**u**RF**'s ranking task. For doing so, we examined several learning models applicable in existing works for ranking web resources: Linear Regression [119], Support Vector Machines [43], Gradient Boosting Regression [60], MultiLayer Perception Regression [80], and learning to rank [109, 110]. In this experiment, a relevant dataset would be one comprised of queries pertaining to children's education searches, along with their corresponding relevant and irrelevant resources. Being that such datasets do not exist, we created two of these for our assessment. The first one we created, which we refer to as $\textbf{Edu\_Search}_{\textbf{DS\_A}}$, is comprised of 100 queries written by children conducting searches in the school setting, along with the resources they selected, which we treat as the ground-truth suitable resources. The other dataset, $\textbf{Edu\_Search}_{\textbf{DS\_B}}$, consists of 100 education-relevant URLs that we extracted from DMOZ (described in Section 3.2), as well as their

corresponding titles. In this case, we treat the titles as the education-relevant queries and their respective URLs as relevant resources. To create the non-relevant resources for queries in the aforementioned datasets, we simulate the search process on Bing using each query and extract the top-4 resources retrieved other than the relevant ones. Following this, each query in both **Edu_Search$_{\text{DS_A}}$** and **Edu_Search$_{\text{DS_B}}$** have their corresponding relevant resources, as well as 4 non-relevant ones (see Table 5.5 for the distribution of resources in **Edu_Search$_{\text{DS_A}}$** and **Edu_Search$_{\text{DS_B}}$**).

For this study, we investigate the performance of aforementioned models in terms of MRR and NDCG, in prioritizing suitable resources for queries in the aforementioned datasets. In this context, for each query in **Edu_Search$_{\text{DS_A}}$** and **Edu_Search$_{\text{DS_B}}$**, we examine the position **KiSuRF** assigns to the corresponding relevant resource, on the ranked list consisting of all resources associated with these queries, when using each of the compared models.

Table 5.5: Resource distribution on **Edu_Search$_{\text{DS_A}}$** and **Edu_Search$_{\text{DS_B}}$**.

Dataset	**Relevant**	**Non-relevant**	**Size**
Edu_Search$_{\text{DS_A}}$	119	501	620
Edu_Search$_{\text{DS_B}}$	100	400	500
Total	**219**	**901**	**1,120**

We initially considered the Learning-To-Rank (LTR) model [109], a supervised machine learning strategy commonly used in the construction of ranking models for Information Retrieval systems. For training purposes, this model requires a dataset containing query-resource pair with a numerical or ordinal score. When we created a dataset of this kind (using training instances constructed from IDLA URLs and their corresponding titles in a similar manner as **Edu_Search$_{\text{DS_B}}$**), and tested on **Edu_Search$_{\text{DS_A}}$** and **Edu_Search$_{\text{DS_B}}$**, we found that LTR model performed poorly in ranking suitable resources. Upon an in-depth analysis of the results, we

observed that ascertaining suitable resources that should be ranked over the other was a challenge, as some resources among the non-relevant set for their corresponding queries were education-relevant as well, hence, leading to noise for the LTR model. As a result, we turn to aggregation models for estimating the relationships among resources that are known to be relevant and irrelevant to education searches.

To train the models considered in this study (i.e., Linear, Gradient Boosting, SVM, and MLP Regression), we use all labeled instances extracted from $\mathbf{Diverse_{DS}}$. We represent each of the instances in $\mathbf{Diverse_{DS}}$ using data points examined by **KiSuRF**'s ranking strategy (discussed in Section 4.6). Gradient Boosting Regression is the model that achieves best results among the models, regardless of the dataset used (see Figure 5.5). The improvements of Gradient Boosting Regression over counterparts are statistically significant in terms of both MRR and NDCG, as determined using a paired t-test having a confidence of $p < 0.05$. Support Vector Machines obtained the lowest performance, which was surprising, as they are widely used for ranking purposes in existing work [46]. It is important to note the performance pattern across all models based on the dataset used, in terms of both NDCG and MRR. All models achieve better performance on $\mathbf{Edu\_Search_{DS\_B}}$, as opposed to $\mathbf{Edu\_Search_{DS\_A}}$. We attribute this to what we believe to be noisy information which we found in $\mathbf{Edu\_Search_{DS\_A}}$: children seemed to pick resources that they liked and those that were positioned earlier (a known fact about children search practices [74]), rather than the ones that were contextually related to their search tasks, making some of the suitable resources to be treated as unsuitable ones. On the opposite spectrum, instances in $\mathbf{Edu\_Search_{DS\_B}}$ contained resources that were either suitable or not, thus, leading to better results across all models. Given the consistency of the ranking performance of Gradient Boosting among different datasets and when compared to its

counterparts, we determined that this model was best suited for **KiSuRF**'s ranking task.

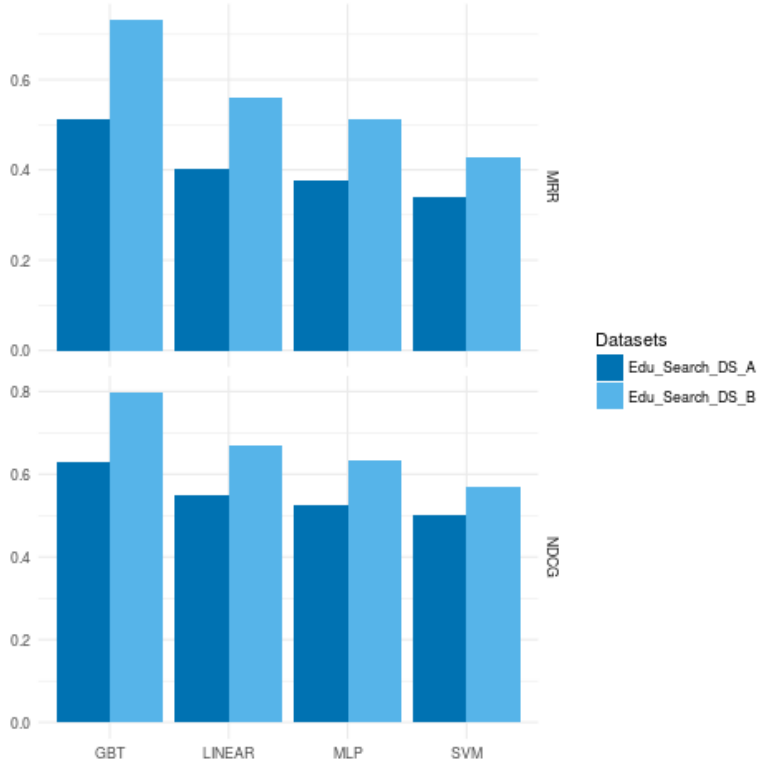


Figure 5.5: Performance of examined regression models used in aggregating data points for ranking suitable resources. GBT: Gradient Boosting, LINEAR: Linear, MLP: MultiLayer Perception, and SVM: Support Vectors Machine.

5.3 Overall Performance Analysis of KiSuRF

For offering suitable resources, **KiSuRF** has two main tasks: filtering resources with unsafe content and then ranking the remaining ones. To demonstrate the effectiveness of **KiSuRF**, first we validate its ability to filter unsafe resources. We then assess its ranking strategy and demonstrate the need to integrate evidences from different

perspectives in determining suitable ones. Finally, we evaluate **KiSuRF**'s overall strategy in ranking suitable resources.

5.3.1 How does KiSuRF's filtering strategy perform when compared to safe search filters on SEs under study?

So far, we examined the correctness of **KiSuRF**'s design strategy in terms of models and qualitative data points it considers. It is, however, imperative that we evaluate its performance in filtering unsafe resources, when compared to safe search filters available in SEs under study, which is why we conduct this experiment. For doing so, we replicated the experiments in Section 3.3.2. In this case, we have **KiSuRF** filter resources retrieved by Google and Bing (with no safe search functionality) for searches simulated using the hate-based, educational, and sexually explicit queries (i.e., queries in $\mathbf{H_{kw}}$, $\mathbf{Ed_{kw}}$, and $\mathbf{B_{kw}}$ introduced in Section 3.3). We also compared **KiSuRF**'s performance in filtering the unsafe resources that scaled through the filter on Kidrex and Kidzsearch. For this experiment,[2] we used all instances in $\mathbf{Kids\_Safe_{DS}}$ (introduced in Section 5.2.6) to train the Random Forests model.

As shown in Table 5.6, **KiSuRF** performs comparably to Kidzsearch and Kidrex in terms of handling web resources that contain sexually explicit and violence-related content, based on our empirical study in Section 3.3.2. Prior to applying **KiSuRF** on these SEs, Kidrex and Kidzsearch retrieved resources for 39% and 57% of searches initiated by queries in $\mathbf{B_{kw}}$, respectively. These numbers, however, reduced significantly to 4% and 1% respectively, after complementing these SEs with **KiSuRF**, illustrating the ability of **KiSuRF** to capture sexually explicit materials, even those that often

[2]We exclude child-oriented SEs from the assessment conducted to validate **KiSuRF** in retrieving educational resources when applied to these SEs, being that we are unable to access resources they filter for respective searches, as this was beyond our control.

Table 5.6: Assessment of **KiSuRF** in complementing the filtering functionality of SEs. N.A. indicates that the SE under study is not applicable for a particular assessment. Information between square brackets, i.e., "[]", indicates initial results obtained in the empirical analysis conducted in Chapter 3.

		KiSuRF applied on results retrieved by			
		Bing	**Google**	**Kidzsearch**	**Kidrex**
Educational Resource Analysis using $\mathbf{Ed_{kw}}$	No result	0.7% [1%]	0.4% [0%]	N.A.	N.A.
Sexually Explicit Content using $\mathbf{B_{kw}}$	No result	18% [29%]	15% [5%]	1% [57%]	4% [39%]
	Retrieves sexually explicit content	4% [33%]	2% [16%]	4% [9%]	3% [14%]
Hate Based Content using $\mathbf{H_{kw}}$	No result	2% [5%]	1% [2%]	1% [15%]	1% [12%]
	Retrieves violence-related content	1% [22%]	1% [23%]	6% [11%]	7% [15%]

bypass safe search filters. **KiSuRF** was also effective in identifying sexually explicit materials that Google and Bing overlooked, as it retrieved these type of resources for at most 4% of the simulated searches. In the case of Google and Bing, this number goes up to at least 16% and 33%, which is high, especially when considering that these percentages were yielded using the safe search version of Google and Bing. Moreover, **KiSuRF** outperformed traditional safe search filters in disregarding resources with violence-related content. We observed that the percentage of searches meant to retrieve resources with violence-related content, reduced by 93% and 92%, respectively, after using **KiSuRF**'s filtering strategy on Kidrex and Kidzsearch. Furthermore, **KiSuRF** was able to capture violence-related materials on the popular SEs, i.e., Google and Bing (retrieved only 1% of hate-based resources), which we find to be beneficial when used in a real life scenario, as parents and educators can be assured that their children do not access resources that promote violence when they utilize their preferred SEs. Most importantly, **KiSuRF** performed similarly to Google and Bing in terms of retrieving resources for nearly all the searches initiated using queries in $\mathbf{Ed_{kw}}$. This was what we envisioned, given that these searches are meant to satisfy

educational information needs.

Based on the analysis of the results of our experiments using $\mathbf{Ed_{kw}}$, $\mathbf{B_{kw}}$ and $\mathbf{H_{kw}}$, we are able conclude that **KiSuRF** is adequately strict when it comes to deterring any resource containing sexually explicit or violent content while still offering resources in response to queries meant to initiate a quest for education-related resources, even if these queries include terms that, while educational, can be misconstrued.

5.3.2 Is considering multiple perspectives essential for prioritizing suitable resources?

In this manuscript, we argued that simultaneously examining resources from multiple perspectives is essential for prioritizing suitable ones. To empirically verify this hypothesis, we performed an empirical study to investigate the usefulness of different aspects for ranking suitable resources. In doing this, we explore how considering individual perspectives, i.e, *objectivity*, *readability*, and *educational value*, as well as multiple aspects affects the prioritization of relevant resources in the aforementioned datasets. In this case, we aim to rank known suitable resources high on the result list.

For conducting this study, we rely on $\mathbf{Edu\_Search_{DS\_A}}$ and $\mathbf{Edu\_Search_{DS\_B}}$ (introduced in Section 5.2.7). For aggregating multiple data points (as discussed in Section 4.6), we use our final ranking model designed using the Gradient Boosting Regression algorithm, and trained on all instances in $\mathbf{Diverse_{DS}}$. Following this, for each query in $\mathbf{Edu\_Search_{DS\_A}}$ and $\mathbf{Edu\_Search_{DS\_B}}$, we predict the ranking score for all resources associated with it, after which we rank these resources based on their assigned scores. Based on the ground truth defined for $\mathbf{Edu\_Search_{DS\_A}}$ and $\mathbf{Edu\_Search_{DS\_B}}$, we compute the MRR and NDCG, being that: (i) we want to limit

the number of resources a user (a child in our case) should select from in response to their educational information need, and (ii) we want to measure the quality of each ranking strategy and ensure that suitable resources are consistently positioned high.

Individual data points, such as *readability* (RD) and *objectivity* (OBJ), had the lowest results for prioritizing suitable resources both in $\mathbf{Edu\_Search_{DS\_A}}$ and $\mathbf{Edu\_Search_{DS\_B}}$ (see Figure 5.6). This was anticipated, as resources of the right reading levels may not necessarily convey education-relevant information for children. It is also probable that resources that are objective may potentially not be of a child's reading level, as well as not being relevant to a his / her educational information need. We found that examining *educational value* (EDV) of retrieved resources was especially important for prioritizing suitable resources. Ranking by this aspect alone obtained good results than other individual ones, and was able to improve the ranking when aggregated along side the others. Moreover, the best results were achieved when all perspectives (i.e., EDV, OBJ, and RD) were simultaneously aggregated for ranking resources in $\mathbf{Edu\_Search_{DS\_A}}$ and $\mathbf{Edu\_Search_{DS\_B}}$. Based on this analysis, we have been able to demonstrate that the perspectives we consider for **K**i**S**u**RF**'s ranking are necessary for prioritizing suitable resources. We can also see that examining resources from a single aspect alone (common in existing works) is insufficient in determining those that should be ranked high for children educational searches.

5.3.3 How is the overall performance of KiSuRF in handling children education-related searches?

All the experiments we performed so far evaluated the effectiveness of **K**i**S**u**RF**'s filtering strategy, as well as the informativeness of perspectives considered for prioritizing resources. To put these in context, we examine **K**i**S**u**RF**'s overall performance,

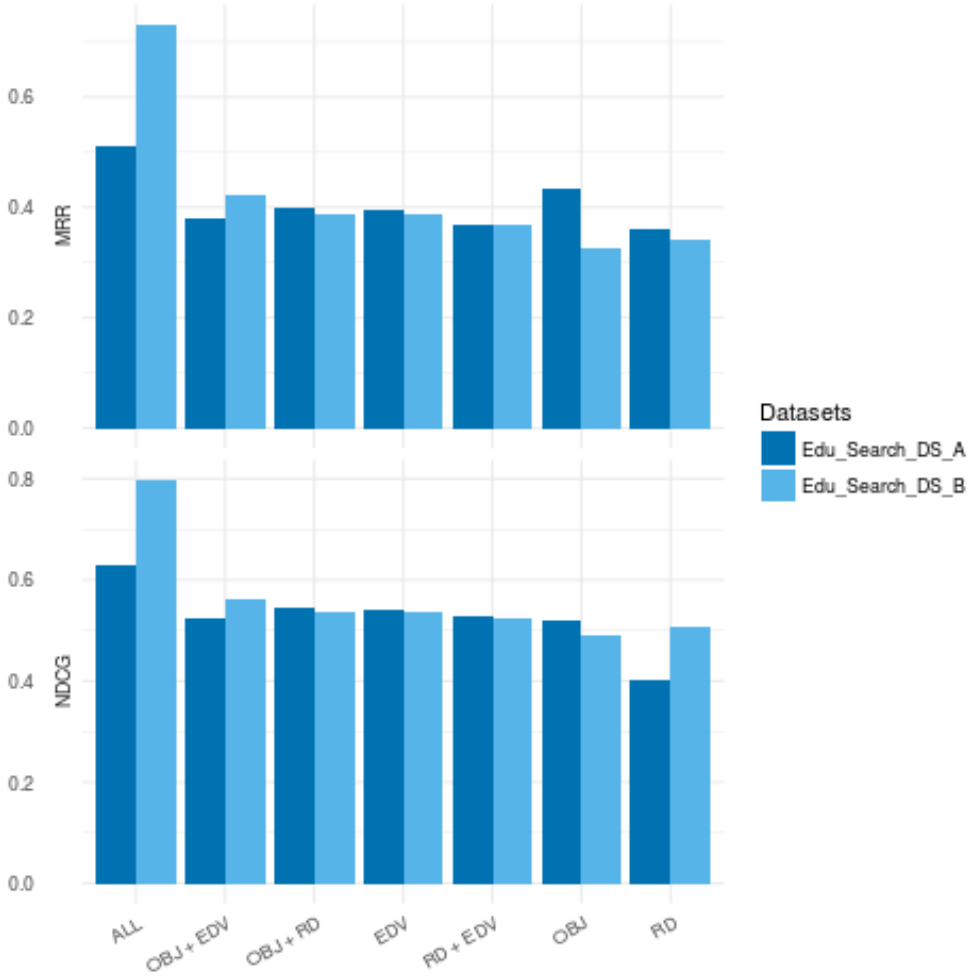


Figure 5.6: Comparison of ranking performance when different aspects of suitability are considered for prioritizing resources. OBJ–Objectivity, RD–Readability, EDV–Education Value, and ALL–a combination of Objectivity, Readability, and Education Value.

when compared to Google and Bing–the two popular SEs favored by children. The aim of this experiment is to investigate how **KiSuRF** fares in handling known child-suitable resources, as opposed to the aforementioned SEs. To validate the ranking performance of **KiSuRF**, we depend upon the MRR and NDCG metric, and treat resources known to be suitable to children educational searches as the relevant ones.

In conducting this experiment, we only rely on **Edu_Search$_{\mathbf{DS_A}}$**, as they were specifically written by children in the school setting for locating education relevant resources. To create the relevant set for queries in **Edu_Search$_{\mathbf{DS_A}}$**, we use these queries to simulate the search on Google and Bing. When investigating the ranking of the relevant URLs on these SEs, we were however, only able to find the corresponding relevant URLs for 11 queries on Google, and 20 queries on Bing, with most of them positioned as low as 50 on the results list.[3] As a result, we only include these queries, along with the top-50 resources retrieved for them on Google and Bing. For each of these queries, we first pass their top-50 resources through **KiSuRF**'s filter. We then rank the remaining ones using **KiSuRF**'s ranking model that is based on the Gradient Boosting Regression.

As shown in Figure 5.7, **KiSuRF** is able to consistently position suitable resources higher on the ranked list both in terms of NDCG and MRR, when compared to Google's and Bing's ranking. Based on Bing's results, on average (before applying **KiSuRF**), a child would have to locate the first suitable resource at the 7^{th} position on the ranked list (in terms of MRR). On the other hand, **KiSuRF** is able to rank the suitable resource at the 4^{th} position, which is a significant improvement, being that we limit the number of resources the child would have to select from (as children are known to select resources sequentially). We also obtained a significant improvement from the ranking assigned by Google for the suitable resources, as on average, while Google initially ranked these resources on the 13^{th} position, we instead rank the them at the 6^{th} position. Moreover, based on NDCG, which we use to measure the quality of our ranking (i.e., how consistent we are able to position suitable resources higher),

[3]The queries in **Edu_Search$_{\mathbf{DS_A}}$** were written in April 2017,which we attribute to the reason for the low ranking on Google and Bing.

we find that **KiSuRF** outperformed Google and Bing (we obtained an improvement of 22% and 16%, respectively). This empirically verifies that when compared to these SEs, we position most of the suitable resources higher. Recall that children usually do not go beyond the first result page when conducting searches [74], which makes it essential that we position suitable resources among the top–10 results. This motivated us to investigate the percentage of searches for which we rank the suitable resources among the top–10 results. From our analysis, we find that for a number of these searches **KiSuRF** surpassed Google and Bing for this purpose (**KiSuRF** obtained an improvement of 33% and 22%, respectively). **KiSuRF**'s results in terms of MRR and NDCG after complementing the retrieval functionality of Google and Bing, was statistically significant than the performance of these SEs before applying it. We determined this by the paired t-test with a confidence of $p < 0.05$.

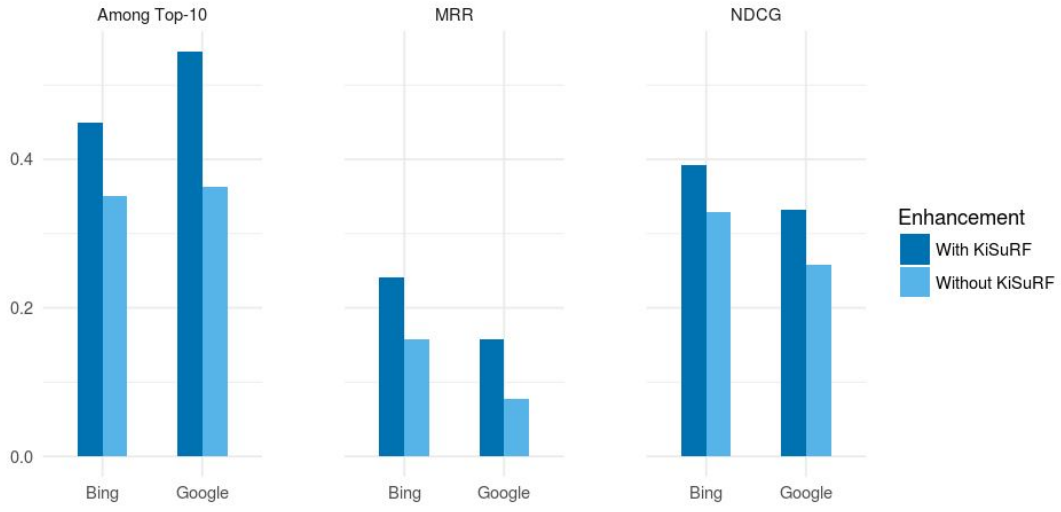


Figure 5.7: Overall performance of **KiSuRF** in prioritizing suitable resources.

Based on the results from this experiment, and all assessments in this Chapter, we were able to empirically verify the validity of **KiSuRF**'s filtering and ranking

performance. We also empirically demonstrate the need to simultaneously examine resources from multiple perspectives, in order to prioritize suitable ones. While experimental results were promising for prioritizing suitable resources when compared to counterparts, we observed that for some searches, children seemed to select resources that were positioned among the top-3 results, even if other ones were more related to an educational context and thus more relevant to their search. This led to some suitable resources being treated as the non-relevant ones. We expect this to be the reason why computed NDCG and MRR scores are within the forty and twenty percentiles, respectively. With this in mind, as we continue our research work in this area, we will conduct a live experiment (an A / B testing) that will allow us to capture resource selection in real time in order to quantify children successful searches, with and without augmenting functionality of their preferred SE with **K**i**S**u**RF**.

CHAPTER 6

CONCLUSIONS AND FUTURE WORK

We presented **K**i**S**u**RF**, a novel filtering and ranking strategy, which offers suitable resources in response to queries written by children conducting searches in a classroom setting. **K**i**S**u**RF** goes beyond traditional safe search and addresses limitations of existing search engines (SEs) in terms of satisfying the information needs that arise when children conduct educational information discovery tasks. **K**i**S**u**RF** eliminates inappropriate resources, i.e., containing hate-speech and sexually explicit content, while retaining resources with education-relevant terminologies. Moreover, **K**i**S**u**RF** takes advantage of a strategy that simultaneously considers the *readability*, *objectivity*, and *educational pertinence* of resources and uses a Gradient Boosting Trees model in order to prioritize resources and present suitable ones to users.

During the research process, we found a lack of evidence in the literature that empirically demonstrates the deficiencies of existing search engines in supporting search tasks in the education environment. As a result, we conducted an empirical analysis on the ranking and filtering strategies on selected SEs, which informed us about the limitations of these SEs in prioritizing education-relevant resources, filtering inappropriate resources, as well as in offering resources that are of the reading levels of children in the 3^{rd}–5^{th} grades. This analysis, to the best of our knowledge, is the first of its kind in the literature. Another important contribution of our work was the

design of **KiSuRF**, which directly responds to gaps identified based on our empirical analysis and complements the retrieval functionality of existing SEs. Given a set of resources retrieved in response to a child's query, we first filtered inappropriate resources by examining their content, meta tags, and anchor tags, using several qualitative indicators. Following this, we simultaneously examined the remaining resources from three different perspectives, when in fact most existing work either serve average users or try to accommodate children from a single perspective. These include: *Readability*, ensuring that children can comprehend their content, *Objectivity* for prioritizing resources that are objective, and *Educational value* ensuring that we favor those that align with the K–12 curriculum. In determining the educational value of resources, our strategy is based on examining contextual information in their content. For this purpose, we created a novel educational knowledge base, by taking advantage of educational standards such as CCSS, NGSS, and ICS, in order to identify K–12 topics and concepts, which is another significant contribution of our work.

We performed different offline experiments in order to validate the correctness and effectiveness of **KiSuRF**'s filtering and ranking strategy. To minimize error mitigation in **KiSuRF**'s design, we evaluated **KiSuRF**'s individual qualitative data points. In doing so, we have been able to demonstrate the degree of influence of each of these data points in terms of informing the performance of **KiSuRF**. Along the way, we also show the promising outcomes obtained by using Spache for estimating the reading level of web resources; our own knowledge base ($\mathbf{Edu_{KB}}$) for identifying the degree to which resources map to the K–12 curriculum; and in explicitly exploring resource objectivity. By comparing **KiSuRF**'s filtering with safe search filters available on SEs under study, we showed that **KiSuRF** is able to disregard inappropriate resources, while retaining education-relevant ones. Furthermore, results

from a number of experiments allow us to assert that simultaneously considering multiple perspectives fosters prioritization of resources that are suitable in response to searches conducted in the classroom setting. Lastly, we empirically verified the applicability of our proposed work by conducting experiments using Google and Bing enhanced with **KiSuRF**.

A major challenge we had to overcome during the research process was in gathering data sources to serve as ground truth for our empirical analysis, as well as in acquiring relevant datasets. For doing so, we explored several textual resources and created a number of datasets, that together permitted the assessment of the performance of the individual components of **KiSuRF**, as well as its overall strategy. These datasets will be made available to the research community as byproducts of this thesis.

6.1 Applicability

We anticipate that the availability of **KiSuRF** can be beneficial for designing a plugin that can be used to complement the functionality of existing SEs for the purpose of offering children suitable resources. Upon the deployment of such a plugin, this can benefit *Educators* and *Parents*, as they can be assured that their students or children are able to access suitable resources, irrespective of the SE being used to conduct their search tasks.

6.2 Future Directions

Although we responded to the impediments of SEs outlined in Chapter 3 through the design of **KiSuRF**, we are aware that there are several additional aspects that we still need to consider in order to determine resource suitability.

Based on an initial analysis on resources known to be designed exclusively for children such as gaming and educational sites extracted from DMOZ, we found that the ones for younger children had a different style of design, i.e., they contained fewer texts, more graphics, and simpler font style. These properties are useful indicators for identifying the target audience of retrieved resources. In the future, to quantify the readability of resources, we will not only depend on the text-based readability, but also analyze resources based on components such as the color contrast, font style, font size, headings, white space, animation and graphics. Qualitative aspects we considered in designing **K**i**S**u**RF**'s filtering strategy for disregarding resources with sexual explicit content were also text-based, which we found to be insufficient for filtering purposes. We observed that a number of examined pornographic websites contained provocative images, while others had inappropriate conceptual meanings that could not be explicitly identified from its individual terms. In the future, we will introduce more novel data points which would include examining graphics, as well as term relatedness in content of resources in order to determine if they are appropriate for a child. We also propose to examine other resource appropriateness aspects, e.g., resources with fake news, as we want to ensure that children access the right level of resources.

We also plan to incorporate temporal aspects in the process of identifying suitable resources, in the future. We are aware that resources relevant to educational searches are constantly updated, hence, the recency of resources should also be considered to inform resource suitability for a child's education information needs. We will examine if the intent of the child's search has a temporal value, and if it happens to be the case, we will prioritize resources by its date of publication. For instance, if a child searches for "President of the United States", we will prioritize a resource with the

name of the current president, whereas if he searches for "volume of a cylinder" recent resources will not be important as the information remains the same over time. In this study, we only utilize $\mathbf{Edu_{KB}}$ for mapping resources to the K–12 curriculum. In the future, we propose to extend our strategy, so that we can determine the grade level of resources as well.

Due to the scope of this thesis, we only considered children in the 3^{rd} – 5^{th} grade levels as they are known to exhibit similar search traits. We plan to extend our proposed solution to the broader K–12 population. Furthermore, we learned that in some cases, children tend to select resources in order, resources that they liked, rather than what was indeed suitable to their search. For instance, in analyzing resources in a dataset comprised of searches explicitly conducted by children in the classroom setting, for the query "where is the organ pipe cactus national monument", some kids selected resources from "answers.com", and ignored those from "kids.britannica.com", even if the latter was more education-relevant and child-suitable. Although beyond the scope of this study, in the future, we plan to conduct a live user study (e.g., an A / B testing) with children, where they directly use **K**i**S**u**RF**, as this would allow us to both offer them a more guided search and to qualify their search success, as opposed to only relying on an offline study.

REFERENCES

[1] Intouch. https://www.intouchweekly.com/, Accessed on: May, 2018.

[2] News and blog data crawl. https://www.kaggle.com/patjob/articlescrape, Accessed on: May, 2018.

[3] Teachers pay teachers. https://www.teacherspayteachers.com/Browse/PreK-12-Subject-Area/, Accessed on: May, 2018.

[4] Alexa Top Sites. https://www.alexa.com/topsites, Accessed: July 30, 2018.

[5] Bing Safe Search. https://www.bing.com/account, Accessed: July 30, 2018.

[6] Bing Web Search Api. https://azure.microsoft.com/en-us/services/cognitive-services/bing-web-search-api/, Accessed: July 30, 2018.

[7] Block explicit results on Google using SafeSearch. https://goo.gl/22Brm5, Accessed: July 30, 2018.

[8] Evaluating Internet Information. https://www.usg.edu/galileo/skills/unit07/internet07_08.phtml, Accessed: July 30, 2018.

[9] Google Custom Search. https://developers.google.com/custom-search/json-api/v1/overview, Accessed: July 30, 2018.

[10] Googles Love Affair with Wikipedia Far More Serious Than Bings [Study]. https://www.conductor.com/blog/2012/05/googles-love-affair-with-wikipedia-far-more-serious-than-bings-study/, Accessed: July 30, 2018.

[11] The Document-level Metadata element. https://developer.mozilla.org/en-US/docs/Web/HTML/Element/meta, Accessed: July 30, 2018.

[12] Top Best Safe Search Engine For Kids. http://www.ilovefreesoftware.com/28/featured/safe-search-engine-for-kids.html, Accessed: July 30, 2018.

[13] Yahoo Search Safety Guide. https://safety.yahoo.com/SafetyGuides/Search/index.htm, Accessed: July 30, 2018.

[14] Nitin Agarwal, Huan Liu, and Jianping Zhang. Blocking Objectionable Web Content by Leveraging Multiple Information Sources. *Association for Computing Machinery's Special Interest Group on Knowledge Discovery and Data Mining (ACM SIGKDD) Explorations Newsletter*, 8(1):17–26, 2006.

[15] Eugene Agichtein, Eric Brill, and Susan Dumais. Improving web search ranking by incorporating user behavior information. In *Proceedings of the 29th annual international Association for Computing Machinery Special Interest Group on Information Retrieval (ACM SIGIR) conference on Research and development in information retrieval*, pages 19–26. ACM, 2006.

[16] Satoshi Ando, Yutaro Fujii, and Takayuki Ito. Filtering Harmful Sentences Based on Multiple Word Co-occurrence. In *Proceedings - 9th IEEE/ACIS International Conference on Computer and Information Science, ICIS 2010*, 2010.

[17] Alexios Mantzarlis Ting Cai Cong Yu Andrew Moore, Bill Adair and R.V. Guha. Academia, publishers and tech come together to open up fact check data. `https://www.datacommons.org/docs/download.html`, May 2018.

[18] Kemafor Anyanwu, Angela Maduko, and Amit Sheth. Semrank: Ranking Complex Relationship Search Results On The Semantic Web. In *Proceedings of the 14th International Conference on World Wide Web*, pages 117–127, 2005.

[19] John Atkinson, Gonzalo Salas, and Alejandro Figueroa. Improving opinion retrieval in social media by combining features-based coreferencing and memory-based learning. *Information Sciences*, 299:20–31, 2015.

[20] Ion Madrazo Azpiazu, Nevena Dragovic, Maria Soledad Pera, and Jerry Alan Fails. Online searching and learning: Yum and other search tools for children and teachers. *Information Retrieval Journal*, pages 1–22.

[21] Ion Madrazo Azpiazu and Maria Soledad Pera. Is readability a valuable signal for hashtag recommendations? In *In Proceedings of the ACM Conference on Recommender Systems (ACM RecSys) posters*, 2016.

[22] Leif Azzopardi, Richard Glassey, Mounia Lalmas, Tamara Polajnar, and Ian Ruthven. Puppyir: Designing an open source framework for interactive information services for children. In *Proceedings of the Annual Workshop on Human-Computer Interaction and Information Retrieval*, volume 44. Citeseer, 2009.

[23] Ricardo Baeza-Yates, Berthier Ribeiro-Neto, et al. *Modern information retrieval*, volume 463. ACM press New York, 1999.

[24] Rakesh Chandra Balabantaray, Monalisa Swain, and Bibhuprasad Sahoo. Evaluation of Web Search Engines based on Ranking of Results and Features. *International Journal of Human Computer Interaction*, 4(3):117–127, 2013.

[25] Dania Bilal. Children's Use of the Yahooligans! Web Search Engine: I. Cognitive, Physical, and Affective Behaviors on Fact-Based Search Tasks. *Journal of the Association for Information Science and Technology*, 51(7):646–665, 2000.

[26] Dania Bilal. Comparing google's readability of search results to the flesch readability formulae: A preliminary analysis on children's search queries. *In proceedings of the Association for Information Science and Technology*, 50(1):1–9, 2013.

[27] Dania Bilal and Meredith Boehm. Towards New Methodologies for Assessing Relevance of Information Retrieval from Web Search Engines on Children's Queries. *Qualitative and Quantitative Methods in Libraries (QQML)*, 2:93–100, 2017.

[28] Dania Bilal and Rebekah Ellis. Evaluating Leading Web Search Engines on Children's Queries. *Lecture Notes in Computer Science*, 6764:549–558, 2011.

[29] Dania Bilal and Jacek Gwizdka. Children's eye-fixations on google search results. *Proceedings of the Association for Information Science and Technology*, 53(1):1–6, 2016.

[30] Yevgen Biletskiy, Hamidreza Baghi, Igor Keleberda, and Michael Fleming. An adjustable personalization of search and delivery of learning objects to learners. *Expert Systems with Applications*, 36(5):9113–9120, 2009.

[31] Bing. Bing Search Engine. `http://www.bing.com`, Accessed: July 30, 2018.

[32] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.", 2009.

[33] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250. AcM, 2008.

[34] Fei Cai, Shuaiqiang Wang, and Maarten de Rijke. Behavior-based personalization in web search. *Journal of the Association for Information Science and Technology*, 68(4):855–868, 2017.

[35] Jeanne Sternlicht Chall and Edgar Dale. *Readability revisited: The new Dale-Chall readability formula*. Brookline Books, 1995.

[36] Iti Chaturvedi, Erik Cambria, Soujanya Poria, and Rajiv Bajpai. Bayesian deep convolution belief networks for subjectivity detection. In *Data Mining Workshops (ICDMW), 2016 IEEE 16th International Conference on*, pages 916–923. IEEE, 2016.

[37] Heeryon Cho and Sang Min Yoon. Improving sentiment classification through distinct word selection. In *Human System Interactions (HSI), 2017 10th International Conference on*, pages 202–205. IEEE, 2017.

[38] Aleksandr Chuklin, Ilya Markov, and Maarten de Rijke. Click models for web search. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 7(3):1–115, 2015.

[39] Kevyn Collins-Thompson, Paul N Bennett, Ryen W White, Sebastian De La Chica, and David Sontag. Personalizing Web Search Results by Reading Level. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, pages 403–412, 2011.

[40] Kevyn Collins-Thompson, Paul N Bennett, Ryen W White, Sebastian De La Chica, and David Sontag. Personalizing web search results by reading level. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 403–412. ACM, 2011.

[41] Kevyn Collins-Thompson and Jamie Callan. A language modeling approach to predicting reading difficulty. 2004.

[42] Common Sense Media. Kid-Safe Browsers and Search Sites. `https://www.commonsensemedia.org/lists/kid-safe-browsers-and-search-sites`, Accessed: July 30, 2018.

[43] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, Sep 1995.

[44] National Research Council et al. *Developing assessments for the next generation science standards*. National Academies Press, 2014.

[45] Craft B. and Ideas D. Common Core State Standards for English Language Arts and Literacy in History/Social Studies, Science, and Technical Subjects. `http://www.corestandards.org/assets/CCSSI_ELA%20Standards.pdf`, Accessed: July 30, 2018.

[46] W Bruce Croft, Donald Metzler, and Trevor Strohman. Search engines: Information retrieval in practice. 283, 2010.

[47] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. 2017.

[48] Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. Hate speech detection with comment embeddings. In *Proceedings of the 24th international conference on world wide web*, pages 29–30. ACM, 2015.

[49] Nevena Dragovic, Ion Madrazo Azpiazu, and Maria Soledad Pera. Is sven seven?: A search intent module for children. In *ACM SIGIR*, pages 885–888. ACM, 2016.

[50] Sergio Duarte Torres, Djoerd Hiemstra, and Pavel Serdyukov. Query log analysis in the context of information retrieval for children. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 847–848. Association for Computing Machinery (ACM), 2010.

[51] Benjamin Edelman. Empirical analysis of google safesearch (2003). *Berkman Center for Internet & Society*.

[52] EduCause. Eligibility for the .edu Domain. `https://net.educause.edu/faq/eligibility`, Accessed: July 30, 2018.

[53] Carsten Eickhoff, Pavel Serdyukov, and Arjen P de Vries. Web page classification on child suitability. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, pages 1425–1428. Association for Computing Machinery (ACM), 2010.

[54] Carsten Eickhoff, Pavel Serdyukov, and Arjen P De Vries. A combined topical/non-topical approach to identifying web sites for children. In *Proceedings of the fourth ACM International Conference on Web Search and Data Mining*, pages 505–514. Association for Computing Machinery (ACM), 2011.

[55] Elana Broch. Childrens Search Engines from an Information Search Process Perspective. `http://www.ala.org/aasl/sites/ala.org.aasl/files/content/aaslpubsandjournals/slr/vol3/SLMR_ChildrensSearchEngines_V3.pdf`, Accessed: July 30, 2018.

[56] Paolo Ferragina and Ugo Scaiella. Tagme: on-the-fly annotation of short text fragments (by wikipedia entities). In *Proceedings of the 19th ACM international*

conference on Information and knowledge management, pages 1625–1628. ACM, 2010.

[57] Interinstitutional File. Proposal for a regulation of the european parliament and of the council on the protection of individuals with regard to the processing of personal data and on the free movement of such data (general data protection regulation). *General Data Protection Regulation*, 2012.

[58] Rudolf Flesch. A New Readability Yardstick. *Journal of Applied Psychology*, 32(3), 1948.

[59] Elizabeth Foss, Allison Druin, Robin Brewer, Phillip Lo, Luis Sanchez, Evan Golub, and Hilary Hutchinson. Children's Search Roles at Home: Implications for Designers, Researchers, Educators, and Parents. *Journal of the Association for Information Science and Technology*, 63(3):558–573, 2012.

[60] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[61] Federal Trade Commission (FTC). Complying with COPPA: Frequently Asked Questions. `Retrievedfrom:http://www.ftc.gov/tips-advice/business-center/guidance/complying-coppa-frequently-asked-questions`, Accessed: March 2017.

[62] George A. Miller. WordNet: A Lexical Database for English. `http://wordnet.princeton.edu/wordnet/download/current-version/`, Accessed: July 30, 2018.

[63] Ayse Göker and Hans I Myrhaug. User context and personalisation. In *Workshop proceedings for the 6th European Conference on Case Based Reasoning*, 2002.

[64] Gene H Golub and Christian Reinsch. Singular value decomposition and least squares solutions. *Numerische mathematik*, 14(5):403–420, 1970.

[65] Google. Google Search Engine. `http://www.google.com`, Accessed: July 30, 2018.

[66] Google. Google's safety tools-Make safety choices that fit your family. `https://www.google.com/safetycenter/families/start/basics/`, Accessed: July 30, 2018.

[67] Tatiana Gossen. *Search engines for children: search user interfaces and information-seeking behaviour*. Springer, 2016.

[68] Tatiana Gossen, Thomas Low, and Andreas Nürnberger. What are the real differences of children's and adults' web search? In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1115–1116. Association for Computing Machinery (ACM), 2011.

[69] Theodosis Goudas, Christos Louizos, Georgios Petasis, and Vangelis Karkaletsis. Argument extraction from news, blogs, and social media. In *Hellenic Conference on Artificial Intelligence*, pages 287–299. Springer, 2014.

[70] Leah Graham and Panagiotis Takis Metaxas. Of course it's true; I saw it on the Internet!: Critical Thinking in the Internet Era. *Communications of the ACM*, 46(5):70–75, 2003.

[71] Neha Gupta and Saba Hilal. Analysis of Web Content Filtering Factors and the Impact of Sieve Coupons. *International Journal of Engineering and Technology*, 4(4), 2012.

[72] Neha Gupta and Saba Hilal. Algorithm to Filter & Redirect the Web Content for Kids. *International Journal of Engineering and Technology*, 5, 2013.

[73] Aviva Lucas Gutnick, Michael Robb, Lori Takeuchi, Jennifer Kotler, Lewis Bernstein, and Michael H Levine. Always Connected. *Retrieved from: http://www.joanganzcooneycenter.org/wpcontent/uploads/2011/03/–jgcc_alwaysconnected.pdf*, 2011.

[74] Jacek Gwizdka and Dania Bilal. Analysis of children's queries and click behavior on ranked results and their thought processes in google search. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*, pages 377–380. ACM, 2017.

[75] Karl Gyllstrom and Marie-Francine Moens. Wisdom of the Ages: Toward Delivering the Children's Web with the Link-Based Agerank Algorithm. In *Proceedings of the 19th Association for Computing Machinery (ACM) International Conference on Information and Knowledge Management*, pages 159–168, 2010.

[76] Zoltn Gyöngyi, Hector Garcia-Molina, and Jan Pedersen. Combating Web Spam with Trustrank. In *Proceedings of the Thirtieth International Conference on Very large data bases-Volume 30*, pages 576–587, 2004.

[77] Alexander Halavais. *Search engine society*. John Wiley & Sons, 2017.

[78] Harry Halpin and Patrick J Hayes. When owl: sameas isn't the same: An analysis of identity links on the semantic web. *LDOW*, 628, 2010.

[79] Taher H Haveliwala. Topic-Sensitive Pagerank. In *Proceedings of the 11th International Conference on World Wide Web*, pages 517–526, 2002.

[80] Simon S Haykin et al. *Kalman filtering and neural networks*. Wiley Online Library, 2001.

[81] Jorge R Herskovic, M Sriram Iyengar, and Elmer V Bernstam. Using hit curves to compare search algorithm performance. *Journal of biomedical informatics*, 40(2):93–99, 2007.

[82] Angelos Hliaoutakis. Semantic similarity measures in mesh ontology and their application to information retrieval on medline. *Master's thesis*, 2005.

[83] HTML.com. HTML Anchor Title Attribute. `https://html.com/attributes/a-title/`, Accessed: July 30, 2018.

[84] Walayat Hussain, Osama Sohaib, Atiq Ahmed, and M Qasim Khan. Web Readability Factors Affecting Users of All Ages. *Australian Journal of Basic and Applied Sciences*, 5(11):972–977, 2011.

[85] Idaho State Department of Education. Idaho Content Standards. `https://www.sde.idaho.gov/academic/standards/`, Accessed: July 30, 2018.

[86] IDLA. Idaho Digital Learning Academy (IDLA). `https://idiglearning.net/`, Accessed: July 30, 2018.

[87] Varghese Jacob, Ramayya Krishnan, Young U Ryu, Ramaswamy Chandrasekaran, and Sungchul Hong. Filtering objectionable internet content. In *Proceedings of the 20th International Conference on Information Systems*, pages 274–278. Association for Information Systems (AIS), 1999.

[88] Varghese S Jacob, Ramayya Krishnan, and Young U Ryu. Internet Content Filtering using Isotonic Separation on Content Category Ratings. *ACM Transactions on Internet Technology (TOIT)*, 7(1):1, 2007.

[89] Adam Jatowt, Kouichi Akamatsu, Nimit Pattanasri, and Katsumi Tanaka. Towards more readable web: measuring readability of web pages based on link structure by adam jatowt, kouichi akamatsu, nimit pattanasri, and katsumi tanaka with ching-man au yeung as coordinator. *ACM SIGWEB Newsletter*, (Winter):4, 2012.

[90] Jennifer Johnson. Kid Friendly Searching. http://jenniferjohnson.bhasd.org/welcome/kid-friendly-searching/, Accessed: July 30, 2018.

[91] Kaggle. All the news. https://www.kaggle.com/snapcrack/all-the-news, June-12-2018.

[92] Samaneh Karimi and Azadeh Shakery. A language-model-based approach for subjectivity detection. *Journal of Information Science*, 43(3):356–377, 2017.

[93] Kassondra Granata. Tech in the Classroom: KidRex. http://www.educationworld.com/a_tech/tech-in-the-classroom/kidrex.shtml, Accessed: July 30, 2018.

[94] Ryan Kelly. Pyenchant a spellchecking library for python, 2016.

[95] Aamera ZH Khan, Mohammad Atique, and VM Thakare. Combining lexicon-based and learning-based methods for twitter sentiment analysis. *International Journal of Electronics, Communication and Soft Computing Science & Engineering (IJECSCSE)*, page 89, 2015.

[96] Kiddle. Web Search Engine for Kids. http://www.Kiddle.io, Accessed: July 30, 2018.

[97] Kidrex. Safe Search for kids. http://www.KidRex.org, Accessed: July 30, 2018.

[98] KidzSearch. Web Search Engine for Kids. http://www.Kidzsearch.com, Accessed: July 30, 2018.

[99] Youngho Kim, Ahmed Hassan, Ryen W White, and Imed Zitouni. Modeling dwell time to predict click-level satisfaction. In *Proceedings of the 7th ACM international conference on Web search and data mining*, pages 193–202. ACM, 2014.

[100] J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training Command Millington TN Research Branch, 1975.

[101] Simon Knight. Finding knowledge–what is it to knowwhen we search? 2014.

[102] John Kurkowski. Tldextract. https://pypi.org/project/tldextract/, Latest release - 2017.

[103] kurzweiledu. 5 Positive Effects Technology Has On Teaching and Learning. `https://www.kurzweiledu.com/blog/2015/02-12-2015.html`, Accessed: July 30, 2018.

[104] Thomas Largillier, Guillaume Peyronnet, and Sylvain Peyronnet. Efficient filtering of adult content using textual information. *Vanessa Murdock, Charles LA Clarke, Jaap*, page 14, 2016.

[105] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, Sören Auer, et al. Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web*, 6(2):167–195, 2015.

[106] Colleen Lennon and Hal Burdick. The Lexile Framework As An Approach For Reading Measurement And Success. *Electronic Publication on http://www.lexile.com*, 2004.

[107] Yiming Li, Baogang Wei, Liang Yao, Hui Chen, and Zherong Li. Knowledge-based document embedding for cross-domain text classification. In *Neural Networks (IJCNN), 2017 International Joint Conference on*, pages 1395–1402. IEEE, 2017.

[108] Shuhua Liu and Thomas Forss. Text classification models for web content filtering and online safety. In *Data Mining Workshop (ICDMW), 2015 IEEE International Conference on*, pages 961–968. IEEE, 2015.

[109] Tie-Yan Liu et al. Learning to rank for information retrieval. *Foundations and Trends® in Information Retrieval*, 3(3):225–331, 2009.

[110] Tie-Yan Liu, Jun Xu, Tao Qin, Wenying Xiong, and Hang Li. Letor: Benchmark dataset for research on learning to rank for information retrieval. In *Proceedings of SIGIR 2007 workshop on learning to rank for information retrieval*, volume 310, 2007.

[111] S Logeswari, R Gomathi, and B Gomathy. Concept based text document summarization using domain ontology. 2017.

[112] Jian Ma, Wei Xu, Yong-hong Sun, Efraim Turban, Shouyang Wang, and Ou Liu. An ontology-based text-mining method to cluster proposals for research project selection. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 42(3):784–790, 2012.

[113] Ion Madrazo Azpiazu, Nevena Dragovic, Oghenemaro Anuyah, and Maria Soledad Pera. Looking for the movie seven or sven from the movie frozen?:

A multi-perspective strategy for recommending queries for children. In *Proceedings of the 2018 Conference on Human Information Interaction&Retrieval*, pages 92–101. ACM, 2018.

[114] Christopher D Manning, Christopher D Manning, and Hinrich Schütze. *Foundations of statistical natural language processing*. MIT press, 1999.

[115] G Harry Mc Laughlin. Smog grading-a new readability formula. *Journal of reading*, 12(8):639–646, 1969.

[116] Pablo N Mendes, Max Jakob, Andrés García-Silva, and Christian Bizer. Dbpedia spotlight: shedding light on the web of documents. In *Proceedings of the 7th international conference on semantic systems*, pages 1–8. ACM, 2011.

[117] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[118] National Association for the Education of Young Children. Technology and young children. `https://www.naeyc.org/content/technology-and-young-children/infants-and-toddlers`, Accessed: July 30, 2018.

[119] John Neter, Michael H Kutner, Christopher J Nachtsheim, and William Wasserman. *Applied Linear Statistical Models*, volume 4. Irwin Chicago, 1996.

[120] Sylvester O Orimaye, Saadat M Alhashmi, and Eu-Gene Siew. Performance and trends in recent opinion retrieval techniques. *The Knowledge Engineering Review*, 30(1):76–105, 2015.

[121] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab, 1999.

[122] Bo Pang and Lillian Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the ACL*, 2004.

[123] Deepshikha Patel and Prashant Kumar Singh. Kids safe search classification model. In *Communication and Electronics Systems (ICCES), International Conference on*, pages 1–7. The Institute of Electrical and Electronics Engineers (IEEE), 2016.

[124] Deepshikha Patel and Prashant Kumar Singh. Kids Safe Search Classification Model. In *Communication and Electronics Systems: International Conference on Communication and Electronics Systems (ICCES)*, pages 1–7, 2016.

[125] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

[126] Marcin Pietroń, M Wielgosz, M Karwatowski, and K Wiatr. A study of parallel techniques for dimensionality reduction and its impact on the quality of text processing algorithms. *Measurement Automation Monitoring*, 61, 2015.

[127] Anita Prinzie and Dirk Van den Poel. *Random Multiclass Classification: Generalizing Random Forests to Random MNL and Random NB*, pages 349–358. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007.

[128] Kristen Purcell, Lee Rainie, Alan Heaps, Judy Buchanan, Linda Friedrich, Amanda Jacklin, Clara Chen, and Kathryn Zickuhr. How teens do research in the digital world. *Pew Internet & American Life Project*, 2012.

[129] Kenneth Reitz. Requests: Http for humans. `http://docs.python-requests.org/en/master/`, Latest release - 2018.

[130] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144. ACM, 2016.

[131] Leonard Richardson. Beautiful soup (html parser). `https://pypi.org/project/beautifulsoup4/`, Latest release - 2017.

[132] Sara Rosenthal and Kathleen McKeown. Detecting opinionated claims in online discussions. In *Semantic Computing (ICSC), 2012 IEEE Sixth International Conference on*, pages 30–37. IEEE, 2012.

[133] Ian Rowlands, David Nicholas, Peter Williams, Paul Huntington, Maggie Fieldhouse, Barrie Gunter, Richard Withey, Hamid R Jamali, Tom Dobrowolski, and Carol Tenopir. The google generation: the information behaviour of the researcher of the future. In *Aslib proceedings*, volume 60, pages 290–310. Emerald Group Publishing Limited, 2008.

[134] Timothy N Rubin, America Chambers, Padhraic Smyth, and Mark Steyvers. Statistical topic models for multi-label document classification. *Machine learning*, 88(1-2):157–208, 2012.

[135] María Helena Mejía Salazar. Towards creating a map of online educational resources based on textual content. 2016.

[136] Satoshi Sanjo and Marie Katsurai. Recipe popularity prediction with deep visual-semantic fusion. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 2279–2282. ACM, 2017.

[137] J Schler, M Koppel, S Argamon, and J Pennebaker. Effects of age and gender on blogging. aaai spring symposium on computational approaches for analyzing weblogs, 2006.

[138] Scholastic. Welcoming the Internet Into Your Classroom. `https://www.scholastic.com/teachers/articles/teaching-content/welcoming-internet-your-classroom/`, Accessed: July 30, 2018.

[139] Michael Schuhmacher and Simone Paolo Ponzetto. Knowledge-based graph document modeling. In *Proceedings of the 7th ACM international conference on Web search and data mining*, pages 543–552. ACM, 2014.

[140] Thomas J. Scott and Michael K. O'Sullivan. analyzing student search strategies: making a case for integrating information literacy skills into the curriculum. *Teacher Librarian*, 33(1):21 – 25, 2005.

[141] Xin Shuai, Xiaozhong Liu, Tian Xia, Yuqing Wu, and Chun Guo. Comparing the pulses of categorical hot events in twitter and weibo. In *Proceedings of the 25th ACM conference on Hypertext and social media*, pages 126–135. ACM, 2014.

[142] Luo Si and Jamie Callan. A statistical model for scientific readability. In *Proceedings of the tenth international conference on Information and knowledge management*, pages 574–576. ACM, 2001.

[143] Advaith Siddharthan. Christopher d. manning and hinrich schutze. foundations of statistical natural language processing. mit press, 2000. isbn 0-262-13360-1. 620 pp. $64.95/£ 44.95 (cloth).-. *Natural Language Engineering*, 8(1):91–92, 2002.

[144] Mrio J Silva, Bruno Martins, Marcirio Chaves, Ana Paula Afonso, and Nuno Cardoso. Adding Geographic Scopes to Web Resources. *Computers, Environment and Urban Systems*, 30(4):378–399, 2006.

[145] George Spache. A new readability formula for primary-grade reading materials. *The Elementary School Journal*, 53(7):410–413, 1953.

[146] Ron Sun and C Lee Giles. Sequence learning: from recognition and prediction to sequential decision making. *IEEE Intelligent Systems*, 16(4):67–70, 2001.

[147] Chenhao Tan, Evgeniy Gabrilovich, and Bo Pang. To each his own: personalized content selection based on text comprehensibility. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 233–242. ACM, 2012.

[148] Google Brain team. Tensorflow: An open source machine learning framework for everyone. `https://www.tensorflow.org/`, Latest update - 2018.

[149] Sergio Duarte Torres, Djoerd Hiemstra, Ingmar Weber, and Pavel Serdyukov. Query recommendation in the information domain of children. *JASIST*, 65(7):1368–1384, 2014.

[150] U.S. Department of Education (DoE). Family educational rights and privacy act (ferpa). `Retrievedfrom:https://ed.gov/policy/gen/guid/fpco/ferpa/index.html.`, Accessed: June 2016.

[151] Sowmya Vajjala and Detmar Meurers. On the applicability of Readability Models to Web Texts. In *Proceedings of the 2nd Workshop on Predicting and Improving Text Readability for Target Reader Populations*, pages 59–68, 2013.

[152] Daniel Valcarce, Javier Parapar, and Álvaro Barreiro. Additive smoothing for relevance-based language modelling of recommender systems. In *Proc. of CERI*, pages 9:1–9:8. ACM, 2016.

[153] W3 Schools. HTML Anchor Tag. `https://www.w3schools.com/tags/tag_a.asp`, Accessed: July 30, 2018.

[154] Henry M Walker. Homework Assignments and Internet Sources. *Association for Computing Machinery (ACM) Inroads*, 4(4):16–17, 2013.

[155] Peng Wang, Bo Xu, Jiaming Xu, Guanhua Tian, Cheng-Lin Liu, and Hongwei Hao. Semantic expansion using word embedding clustering and convolutional neural network for improving short text classification. *Neurocomputing*, 174:806–814, 2016.

[156] Tao Wang, Yi Cai, Ho fung Leung, Zhiwei Cai, and Huaqing Min. Entropy-based term weighting schemes for text categorization in vsm. *2015 IEEE 27th International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 325–332, 2015.

[157] Laura Wendlandt, Jonathan K Kummerfeld, and Rada Mihalcea. Factors influencing the surprising instability of word embeddings. *arXiv preprint arXiv:1804.09692*, 2018.

[158] Wentao Wu, Hongsong Li, Haixun Wang, and Kenny Q Zhu. Probase: A probabilistic taxonomy for text understanding. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, pages 481–492. ACM, 2012.

[159] Tian Xia. Support vector machine based educational resources classification. *International Journal of Information and Education Technology*, 6(11):880, 2016.

[160] Biao Xiang, Daxin Jiang, Jian Pei, Xiaohui Sun, Enhong Chen, and Hang Li. Context-aware ranking in web search. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 451–458. ACM, 2010.

[161] Yahoo. Yahoo! Search Engine. `http://www.yahoo.com`, Accessed: July 30, 2018.

[162] Stefan Zwicklbauer, Christin Seifert, and Michael Granitzer. Doser-a knowledge-base-agnostic framework for entity disambiguation using semantic embeddings. In *International Semantic Web Conference*, pages 182–198. Springer, 2016.

APPENDIX A

FRAMEWORK FOR ARCHIVING CHILDREN'S SEARCH SESSIONS

In this section, we provide a description of the framework we designed in order to archive children's search session. We were particularly interested in identifying queries specifically written by children, as well as retrieved resources preferred by them. In Information Retrieval, researchers often depend upon extracting children's queries from the popular AOL query log by identifying searches that retrieved resources that map to known children websites. Unfortunately, some of these queries may have been mis-labeled as some adults may have similar search patterns as children.

We designed our framework using .PHP and we store all search sessions in a MYSQL database server. Information we archive include: the child's grade level, search query, links selected, position of the links selected, as well as the time-stamp of the search. In using this frame work, a child first anonymously select his / her grade level from a drop down list (see Figure A.1) and is redirected to the search page after clicking on the submit button. The grade levels are pre-defined according to the K–12 curriculum.

As shown in Figure A.2, the search interface was designed to look like Google, so the child would have a similar search experience as they would when they searching on their preferred search engine. We show a sample of an archived search session in Figure A.3. Information we extracted from the query log was paramount for creating the gold standard that we used in evaluating the overall ranking strategy of **K**i**S**u**RF**.

Figure A.1: Interface for a child searching to indicate his grade level.

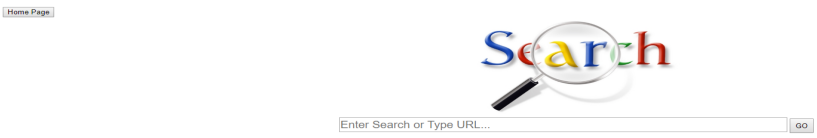


Figure A.2: Interface where a child initiates his search and selects retrieved resource of interest.

```
<session>
  <sessionid>{CF38BCC2-7C08-1A5E-D594-EC3EFD958E2B}</sessionid>
  <timestamp>2017-04-18 20:41:35</timestamp>
  <grade>
    <name>5th Grade</name>
    <queryterms>
      <query>
        <searchterm>
          when was the last time this volcano erupted in new mexico
        </searchterm>
        <trigger>none</trigger>
        <usedsuggestion>false</usedsuggestion>
        <timestamp>2017-04-18 20:43:18</timestamp>
        <clickedlinks>
          <clickedlink>
            <url>
              https://www.scientificamerican.com/article/recent-east-coast-volcano/
            </url>
            <position>0</position>
            <timestamp>2017-04-18 20:43:25</timestamp>
          </clickedlink>
        </clickedlinks>
      </query>
      <query>
        <searchterm>when was the last time capulin erupted</searchterm>
        <trigger>when was the last time capulin</trigger>
        <usedsuggestion>true</usedsuggestion>
        <timestamp>2017-04-18 20:46:18</timestamp>
        <clickedlinks/>
      </query>
      <query>
        <searchterm>how many people died in johnstown flood</searchterm>
        <trigger>how many people died in johnst</trigger>
        <usedsuggestion>true</usedsuggestion>
        <timestamp>2017-04-18 20:50:13</timestamp>
        <clickedlinks/>
      </query>
      <query>
        <searchterm>what is the name of that desert organ</searchterm>
        <trigger>none</trigger>
        <usedsuggestion>false</usedsuggestion>
        <timestamp>2017-04-18 20:53:24</timestamp>
        <clickedlinks/>
      </query>
      <query>
        <searchterm>
          island that is not in the united states of america roosevlt
        </searchterm>
        <trigger>none</trigger>
        <usedsuggestion>false</usedsuggestion>
        <timestamp>2017-04-18 21:01:44</timestamp>
        <clickedlinks/>
      </query>
    </queryterms>
  </grade>
</session>
```


Figure A.3: Sample of a child's search session.

APPENDIX B

SUMMARY OF DATASETS

We discuss below all the datasets we used for analysis and experimental purposes in Chapters 3, 4, and 5, which are summarized in Tables B.1, B.2 and B.3, respectively.

Table B.1: Summary of datasets described in Chapter 3.

Name	**Section**	**Description**	**Training (Tr) / Testing (Te) / Search Simulation (S) / Analysis (A)**
$\mathbf{Ed_{kw}}$	3.3.1 5.3.1	Queries with educational keywords	S
$\mathbf{B_{kw}}$	3.3.2 5.3.1	Queries with sexually explicit keywords	S
$\mathbf{H_{kw}}$	3.3.3 5.3.1	Queries with hate based keywords	S
$\mathbf{k_{qry}}$	3.3.4 3.3.6	Queries written by children conducting searches in the educational environment	S
$\mathbf{E_{qry}}$	3.3.5	Pairs of the form ⟨ title, URL ⟩, where the "titles" are treated as the educational queries and the "URLs" are treated as the relevant resources for the corresponding queries	S

Table B.2: Summary of datasets described in Chapter 4.

Name	**Section**	**Description**	**Training (Tr) / Testing (Te) / Search Simulation (S) / Analysis (A)**
$\mathbf{Exp_{res}}$	4.5.1	A dataset that consists of labeled sexually explicit web resources	A
$\mathbf{Hate_{res}}$	4.5.2	A dataset that contains labeled hate-based resources	A
$\mathbf{K_{safe}}$	4.5.3 4.5.4	A dataset that contains labeled kids-safe and kids-unsafe resources	Tr and Te
$\mathbf{Wiki_{res}}$	4.5.3	Wikipedia resources	Tr
$\mathbf{Diverse_{res}}$	4.6.3	A dataset that consists of diverse resources pertaining to education, celebrity magazine news, as well as weather, sports, economy, and political news	A

Table B.3: Summary of datasets described in Chapter 5.

Name	Section	Description	Training (Tr) / Testing (Te) / Search Simulation (S) / Analysis (A)
$\textbf{Edu}_{\textbf{TOPICS}}$	5.2.4	Resources labeled with respective K-12 topics	A, Tr, and Te
$\textbf{Edu\_Search}_{\textbf{DS\_A}}$	5.2.7 5.3.3	Tuples of the form ⟨ query, URL, label ⟩, where the "queries" were written by children at school, "URL" refers to resources associated with each query, and label indicates the relevance of each URL to the corresponding query	A and S
$\textbf{Edu\_Search}_{\textbf{DS\_B}}$	5.2.7	Tuples of the form ⟨ query, URL, label ⟩, where the "queries" were extracted from the titles of educational resources, "URL" refers to resources associated with each query, and label indicates the relevance of each URL to the corresponding query	A and S
$\textbf{Diverse}_{\textbf{DS}}$	5.2.5	A dataset comprised of 3,000 resources pertaining to education, celebrity magazine news, as well as weather, sports, economy, and political news	Tr and Te
$\textbf{Movie}_{\textbf{DS}}$	5.2.2	A dataset that consists of 10,000 movie reviews and plots that have been labeled as subjective and objective resources, respectively	Tr and Te
$\textbf{Read}_{\textbf{DS}}$	5.2.1	A dataset comprised of 40,000 educational resources, along with their respective K-12 grade levels	A
$\textbf{WebObj}_{\textbf{DS}}$	5.2.2	A dataset that includes 240,000 diverse objective and subjective web resources	Tr and Te
$\textbf{Kids\_Safe}_{\textbf{DS}}$	5.2.6	A dataset that consists of 14,000 resources labeled as safe and unsafe for children	Tr and Te