Evaluating Relevance Feedback: An Image Retrieval Interface for Children

Sander Bockting, Matthijs Ooms, Djoerd Hiemstra, Paul van der Vet, Theo Huibers
University of Twente
Department of Computer Science
P.O.Box 217, 7500 AE Enschede, The Netherlands
{bockting,oomsm,hiemstra,vet,huibers}@ewi.utwente.nl

ABSTRACT

Studies on information retrieval for children are not yet common. As young children possess a limited vocabulary and limited intellectual power, they may experience more difficulty in fulfilling their information need than adults. This paper presents an image retrieval user interface that is specifically designed for children. The interface uses relevance feedback and has been evaluated by letting children perform different search tasks. The tasks were performed using two interfaces; a more traditional interface – acting as a control interface – and the relevance feedback interface. One of the remarkable results of this study is that children did not favor relevance feedback controls over traditional navigational controls.

1. INTRODUCTION

The Convention on the Rights of the Child [11] states that children should be able to access special types of information, such as "information and material of social and cultural benefit to the child" or "information that helps the development of the child's personality, talents and mental and physical abilities to their fullest potential" [7]. Children can use information retrieval systems for this purpose. However, one problem of information retrieval systems is that they are not aimed at little children. Children may find themselves experiencing problems that adults do not encounter when using such systems. Some examples of them are: the information is not suited for their age, children may be unable to formulate their queries without errors and interfaces not specifically designed for children. One way to aid children in letting them find information is more specialized information retrieval systems using interfaces that are more suitable for children.

Delving deeper into the interface problem, imagine a child sitting in front of a computer searching for images. He types in his query "mushrooms" and starts the search. The system returns a lot of images of mushrooms, possibly accompanied with metadata about the images. But what should a child do if he does not instantly find the result he hoped for? He may not know how to refine his query to retrieve better results. In this case, all he can do is browse through all result pages, hoping to find an image he was searching for.

In this study a new image retrieval interface, specifically designed for children, is proposed and evaluated. This interface uses relevance feedback (RF) to let children refine their search results. In this case, RF is applied to support image retrieval. It works by indicating which images look (dis)similar to the image a user is searching for. Using this indication, the image retrieval system is able to recompute a new result set that (hopefully) contains more images like the one a user is looking for.

A lot of effort has been put in investigating the effect of RF already [15, 2, 4]. It appears that adults do not like to use RF [14], but this may be different in the case of children. As children may not be able to formulate their query as well as adults, they may favor a method that helps them refine their query without using extra words. This paper contributes to the research field by studying how children can search using RF.

This investigation will be done by creating an image retrieval system that can work with RF. To test whether children can search better with RF, two interfaces were created. The first interface looks like a more traditional interface – like Google Images – and acts as a control interface. The second interface has RF incorporated and is the experimental interface. This interface will be called the RF interface.

In the following section research question are defined. Sect. 3 describes the prototype image retrieval system that was developed. It is followed by a section describing three parts of the evaluation: the design, the participants and the procedure. Sect. 5 describes the results of the evaluation. The paper ends with a discussion and a conclusion.

2. RESEARCH QUESTIONS

This study investigates whether an image retrieval interface with RF incorporated works better for small children than a more traditional interface without RF. As 'better' is quite vague, this term needs refinement. 'Better' can be reformulated to 'getting the results quicker'. But then again, 'quicker' is also susceptible to multiple interpretations, namely less time or less actions performed before finding an image. During testing of the interfaces, some children will need a lot of confirmation while searching or they are not as autonomous as others. Therefore, there may be a huge variance in the time children need to find an image, so we opted to look at the number of actions performed in order

DIR-2008 April 14-15, 2008, Maastricht, the Netherlands.

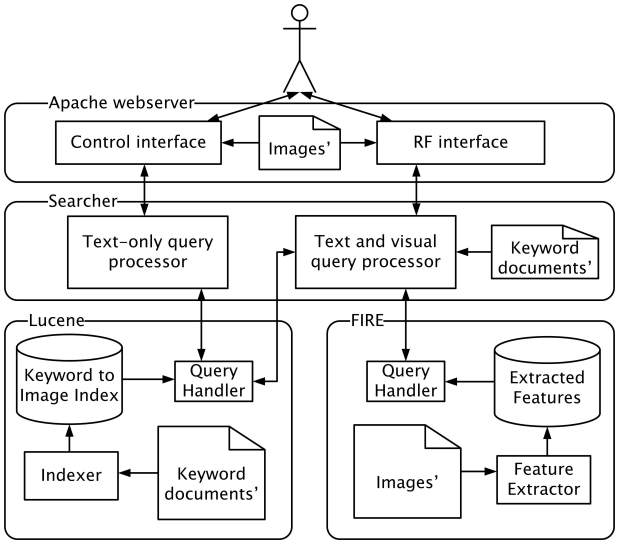

Figure 1: Lucire Architecture

to find the result. These actions are clicks on the controls of the interface.

This results in the following research questions:

1. Can children find images performing less actions using an image retrieval system that incorporates relevance feedback than while using a more traditional image retrieval system?

2. Do children value an image retrieval system with relevance feedback higher than a more traditional image retrieval system?

3. PROTOTYPE

To test a traditional image search interface acting as a control interface, an image retrieval system is required that is able to search using text. To test the RF interface, the retrieval system should also be able to handle RF. Instead of completely rebuilding what already exists, Apache Lucene [12] and the Flexible Image Retrieval Engine (Fire) [3] were used to build an image retrieval system. The new image retrieval system, creatively called Lucire, enables the possibility to use the same engine for both interfaces.

The architecture of the Lucire application is depicted in Fig. 1. The top layer in the diagram shows the web server. The web server contains the different interfaces that will be tested. The interfaces get the user's queries and present the search results. The web server also has a copy of the images of the test collection to present the results to the user. The middle layer describes the Search layer. The query processors in this layer connect the interfaces to the actual search engines and process the results from the search engines to generate the results that will be presented to the user. The bottom layers contain the actual search engines. The middle and bottom layers will be discussed in more detail in this section.

3.1 Test collection

As this study deals with children, the test collection needed to be child friendly. This was achieved by taking a general

Figure 2: Example of a butterfly image in the test collection.

selection of images from the Microsoft Office Online clip art[1]. The general collection consisted of a random selection of about 3000 images, varying from simple drawings to photos from hundreds of topics. Besides a more general selection, a deeper selection has been made for some categories: animals living under water, candies, cyclists, dogs, flags, orchids and palm trees. Each of those categories consisted of about 100 images. As it turned out, this deeper selection in categories was necessary because the images would otherwise be too easy to find with the image retrieval system; the Dutch keywords assignment – done by Microsoft – was performed quite well, making the images easy to find. A text file was produced for every image that contained the image file name and the keywords corresponding to that image.

Besides the Office Online clip art, 200 images of butterflies were added as well. These images originated from the Natural History Museum [2] and an example of such an image is shown in Fig. 2. Including the butterfly images, the total collection consists of 3921 images.

3.2 Lucene

Apache Lucene [9] is a well-documented text search engine written in Java. We selected it for its easy modifiability and capability to do full-text search in many files. By default, Lucene is able to index whole words in files. Using wildcard (*) and substitution (?) tokens, words beginning with a certain part of the keyword can be found, e.g. telephone can be found with tele* or teleph???. However, it does not work when the wildcards are used at the beginning of a word; telephone can not be found with *phone.

As this may reduce the effectiveness of keywords assigned to images, the Lucene Indexer was modified to also index subwords by iteratively removing the first letter of every word while adding it to the index. This has the major advantage that telephone can be found with the search term 'phone', but it also has a disadvantage. Originally existing subwords – especially if the words are short – may suddenly occur much more, giving it a higher term frequency in one document and in the whole document set. This results in lower scores when searching for that shorter term if it also

[1]The MS Office Online clip art can be found at `http://office.microsoft.com/nl-nl/clipart/default.aspx`

[2]The Cockayne collection can be found at `http://www.nhm.ac.uk/research-curation/projects/cockayne/`, date of last visit: March 13, 2007

is a subword of a longer word.

Fig. 1 shows that Lucene uses keyword documents, which are files containing keywords for images with the filename of the image with an extension appended to it, e.g. `img1.jpg.txt`. The Indexer indexes them using the aforementioned approach and stores it in an index. The Query Handler in its turn can search this index for search terms and then returns the filenames of the images having these terms as keywords.

3.3 Fire

Fire is a content-based image retrieval system [3], which enables searching for images by image content. It does this by extracting features from images and storing them, so they can later be used to find images with similar properties like the query image.

As it was not our aim to build the best image retrieval system ever, we opted to analyze the images with Fire's default feature extraction tools, thereby extracting the following features: color histogram, global texture feature, invariant feature histogram and the tamura texture histogram. Analyzing all images from the collection took about 90 hours with three AMD Athlon 64 and one Intel Pentium Mobile processors with an average clock speed of 2GHz.

Just as Lucene, Fire also has a Query Handler, which accepts queries in the form of positive and negative images. A query may look like this: `retrieve +img1.jpg -img2.jpg`. Fire searches images that are similar to the positive images and dissimilar to the negative images and returns a predefined fixed number of best matching images.

3.4 The Searcher module

The Searcher module, as shown in Fig. 1, connects the two interfaces to the two retrieval engines. In fact, the Searcher module consists of only one query processor, but to better indicate the difference between the two interfaces, they have been separated in the diagram. The two search methods will be discussed separately.

3.4.1 Search using the Control Interface

When a user searches with a set of search terms, the query is modified, before sending it to Lucene's Query Handler. The example query `test queries` is expanded to `test^3 (test*)^2 queries^3 (queri*)^2` by performing the following steps:

- Boost the original terms by a factor of three, making it the most important terms.
- If a term is longer than 5 characters, the last two characters are removed. The result is appended with a *, to find words starting with the (possibly reduced) search term. This step simulates word stemming a bit [16]. The "stem" is boosted by a factor of two.

This query expansion method was the result of some quick experimentation and yielded satisfying results with the queries. It should by no means be seen as a scientifically valid way to produce the best results, but this was not the intention.

The expanded query is sent to Lucene's Query Handler which calculates a set of scores for every image in the collection. The scores in the result set are normalized, sorted and sent to the interface by the query processor. The interface presents the results and when a user wants to perform a new query, the whole process is repeated.

3.4.2 Search using the RF Interface

The RF interface differs in two aspects from the control interface: it also searches using visual aspects and it incorporates RF. The search process can be characterized by the following steps:

- The query expansion is the same as with the control interface, except when two or more positive images have been marked. In that case, the query is also expanded with extra terms. The query processor pairwise compares all keywords associated with the positive images and every cooccurring term is added to the query, without any factor boosts or other modifications. By doing this, the RF uses both image features and keywords to compute a result set. The expanded query is sent to the Lucene Query Handler which produces a result set .
- Every time a search is performed, the RF set of positive and negative images is sent to Fire's Query Handler, except when the RF set is empty. A result set is computed, consisting of a predefined number of best matches (as otherwise the whole set of images can be returned). In this case the predefined number is set to 300.
- Both results from the Query Handlers are normalized to 1 and merged using the following formula:

  $$\text{score}(i) = 0.5 \cdot \text{lucene\_score}_i + 0.5 \cdot \text{fire\_score}_i$$

  This score is calculated for every image i that occurs in at least one of the two result sets.

4. EVALUATION

4.1 Design

The evaluation of the interfaces is performed by using a between-subjects design, which means that every subject does a certain exercise with only one interface [5]. It is not possible to use a within-subjects design [6], as doing exercises with one interface may give clues for doing the same exercise with the other interface.

The between-subjects design has the advantage that the scores are independent of other scores, as each subject is measured only once. However, it also has a major disadvantage, namely more variance due to individual differences of the participants which means more participants are required.

There were six exercises, but as every exercise could be answered with or without the use of RF, effectively there were twelve exercises. Each subject performed three exercises. One or two exercises were done using the RF interface and the other using the control interface. A subject never performed the same exercise on both interfaces – as that would make it a within-subjects design – and the exercises were divided in such a way that every exercise was performed about equally often with as without the use of RF.

The six exercises were the following:

1. Find the following image, where an image with a pile of red and white striped candies was shown.
2. Find the following image, where an image of a checkered (racing) finish flag was shown.

3. Find three images of animals living in water, where an image with a huge number of fishes in the ocean was shown.

4. Find three images of flowers or trees with pink leaves.

5. Sometimes butterflies have a pattern on their wings that looks like eyes. An image of such a butterfly was shown (see Fig. 2). The question was to find three images of butterflies with such a pattern on their wings.

6. Find the following image, where an image of stunting/jumping bikers was shown.

4.2 Participants

As this research is aimed at children, children are the preferred group of test users. The teachers at the RC Paulus primary school in Enschede, the Netherlands, permitted us to test the interfaces with children of their classes. In consultation with the teachers, the children of grade four belonged to the youngest suitable group, as they are just able to type words and have a limited vocabulary. This reasoning is supported by Hanna et al [8]. By selecting children from one grade, a common group of children is obtained, of which children on average have equal experience with computers as their peer group.

The test group consisted of 14 girls and 6 boys, with ages ranging from 7 to 9 years and an average age of 8.2 years. Seven children indicated having searched for images before, using Google or Kennisnet (a portal for children).

4.3 Procedure

The tests took place in a separate closed room next to the children's classroom, minimizing the surrounding noise and reducing the chance of distraction. The children were asked to sit in front of a laptop, one at a time. Two people were sitting beside the child, an interviewer and an observer. The interviewer explained the system, asked questions and tried to help the child where needed. The interviewer also helped with spelling errors, as we assume spelling checkers exist that can correct typo's made by the child. The child's click behaviour, search terms, search results and remarkable actions were written down by the observer. At the end of the test, the child was asked how he valued both interfaces by asking which interface he liked best, and which one he found most easy. Also the question was asked whether the child would use the system if it would be available on the Internet.

Fig. 2 shows a screenshot of the RF interface. Its functionality was explained by doing an example exercise. During the explanation, the following things were explained about the search process:

- The search is started by typing one or more search terms and pressing the search button.

- The system shows ten images and ten other images can be selected by pressing a number on the number bar.

- If a happy smiley is clicked on, the system tries to find more images that look like the image belonging to that smiley. If an unhappy smiley is clicked on, the system tries to find images that do not look like the belonging image.

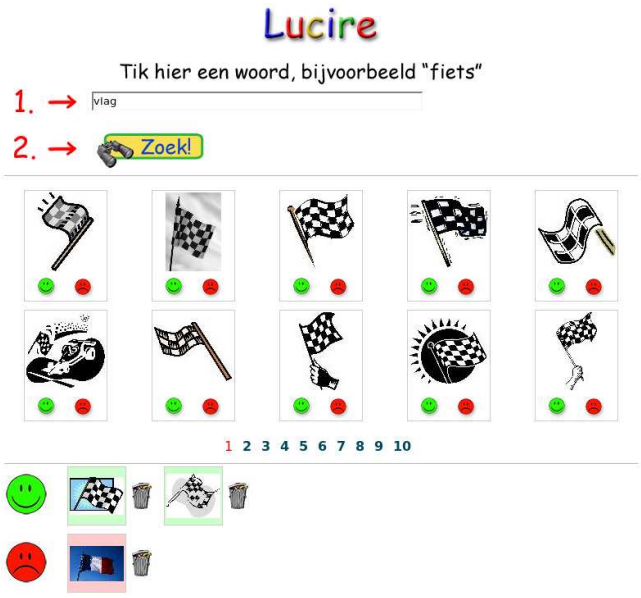

Figure 3: Screenshot of Lucire RF interface. The control interface looks similar, but has no smileys and misses the bottom part of the screen.

- If the result is not as expected, a happy or unhappy click could be undone by pressing the garbage can.

5. RESULTS

> Can children find images performing less actions using an image retrieval system that incorporates relevance feedback than while using a more traditional image retrieval system?

In order to define a null hypothesis corresponding to this first research question, some measures are defined. As stated in Sect. 2, the number of actions performed can be measured by counting the clicks needed to get to the result. The observer logged the click behaviour per subject. Clicks on Relevance Feedback controls (smileys and trash cans) and Clicks on Navigational Controls (the page numbers) are referred to as CRF and CNC respectively. The sum of CRF and CNC is called the Total Click Count (TCC).

When the subjects in the RF group need to perform less actions in order to achieve their goal (lower TCC) than the subjects in the control group, the RF interface provides a better means for children to search images. Accordingly the following null hypothesis was defined:

- H_0: There is no difference in mean TCC score between the control group and RF group

The mean TCC scores of the RF group and control group are given in table 1. The score of the RF group is lower than the score of the control group. However, an independent samples t-test shows this result is not significant: $t = -0,586$, $p > 0,05$. H_0 could not be rejected.

It is interesting to examine the mean TCC per exercise as well. Some exercises could have benefited more from RF

	N	Mean	Std. Dev
RF Group	26	4,88	3,93
Control group	31	5,68	6,19

Table 1: Mean TCC score

Exercise:	1	2	3	4	5	6
No RF	0	9	3	2	7	4
	4	10	2	2	14	4
	0	9	3	5	9	0
	4	7	4	1	13	0
	4	7				
RF	1	17	3	1	5	3
	6	9	1	3	2	4
	0	12	5	0	9	19
	0	22	0	3	12	6
	4	0	3	3	5	18

Table 2: TCC scores per exercise

than others, like the butterfly exercise (number 5). The TCC scores per exercise can be found in table 2.

The mean TCC scores per group for the butterfly exercise are given in table 3. The mean score of the RF group is lower than the mean score of the control group. However, a t-test shows this result is not significant either: $t = 1.689$, $p > 0.05$. A t-test was performed for all the other exercises. Neither of them was significant; no difference was found between the TCC means of the two groups for any of the exercises.

> Do children value an image retrieval system with relevance feedback higher than a more traditional image retrieval system?

The children were asked for their opinion: which interface they liked best, and which interface they found most easy. The result is shown in table 4. The children indicated that they prefered the RF interface, but they did not find it easier to use.

It might be interesting to see whether children preferred searching using RF or preferred navigating using page numbers. To test whether there is a difference the following null hypothesis is defined.

- H_0: There is no difference between the mean CRF score and CNC score in the RF group.

The mean CRF and CNC score of the RF group is shown in table 5. In the calculation of the mean scores, some outliers were omitted. If the children found the image using only queries, the CRF and the CNC score are both 0. This means that the child did not have a choice to click on a button, so we left those results out of the calculation of the

	N	Mean	Std. Dev
Butterfly RF Group	4	10,75	3,30
Butterfly Control group	5	5,60	3,91

Table 3: Mean TCC score for the butterfly exercise

	No RF	Equal/no opinion	RF
Easier	7	3	9
Nicer	5	5	9

Table 4: Children's preference

	N	Mean	Std. Dev
CRF	24	1,92	2,83
CNC	24	4,71	4,77

Table 5: Mean CRF and CNC score of RF group

means. We also left out an outlier where a child started clicking on many negative feedback smileys and worked herself further away from a good result set. In our opinion, leaving out the outlier is justified, as it could be seen as an error in the interface.

The mean score of the CRF is lower than the CNC score. If outliers are ignored, a paired samples t-test [10] shows the difference is significant: $t = -2.66$, $p < 0.05$. Therefore, H_0 is rejected. Children used the RF controls less than the traditional page navigation controls.

6. DISCUSSION & CONCLUSIONS

Unfortunately there was no evidence found that RF provides a better means for children to search images. However, because the quality of the user experience is important as well, RF is not rendered useless, as the children did value the RF system higher than the traditional system. A similar result of insignificant difference but more appreciation has been reported in a comparable study with adults [2].

While conducting the experiment, we quickly discovered children did not quite search the way we anticipated. For example, they came up with keywords we had not thought of, which sometimes led to the result in just a single query. Then they did not have to search using RF or navigate through the results at all. Also, looking at the results of TCC per exercise, it appears that not all exercises had the same level of difficulty; exercise 2 (finding the image of the race flag) turned out to be the most difficult. Although the level of difficulty of the exercises does not necessarily have to be equal if they are all tested in large numbers, it may be wise to first test the exercises with a small group of children to confirm that they do not produce unexpected results.

Another issue was the fact that some children started clicking on the negative feedback control immediately. This does not work very well. For example, when a negative click on a flag is given, the system can come up with anything dissimilar to a flag; it might even return a bike. In these cases the children continued clicking on the negative RF controls resulting in even worse behaviour (TCC>17 in some cases). This issue could be solved by removing the negative RF control, or by limiting the number of negative feedback images to a maximum of the number of selected positive images. The results will then not only be more reliable, but the experience for the children may improve as well.

The children did not use the RF controls as much as the standard navigational controls, which might have been caused by the children's lack of understanding of the RF concept. Although the interface was explained to all chil-

dren using an example exercise, it was not clear whether the children understood the concept of RF. Only three children convinced us of their understanding. Most times, when the children were asked "Why did you click on the green smiley of that butterfly?", they answered "Because that is the picture I was looking for" or "Because I like it!", instead of "Because that picture looks like the one I'm looking for!". This might be caused by a lack of experience with the system. Instead of giving the children just one example question and some explanation, the level of understanding may be improved by letting the children play with the interfaces for a couple of minutes before starting with the tests.

The observant reader notices that the CRF, CNC or TCC scores do not contain the number of times that the children (re)started the search with a query by pushing the "Zoek!" button ('zoeken' is Dutch for 'to search'). We opted to leave this score out, as clicking on this button is required for both means of searching and we did not expect a significant difference in number of clicks on that button; restarting a query is a mere chance of a child posing the right query terms. The small difference was confirmed by our results; there appeared to be no significant difference between the number of (re)starts and the type of interface. 79% of the questions was answered using only one query. The other 12 questions were answered with an average of 2.5 queries, with no large deviation in favor of a particular interface.

All children seemed to like working with the system and were having fun searching for the pictures. When they got back in their classroom they spoke enthusiastic about the exercises to their classmates. Of course, this was very new for most children. Only seven of them had searched images before on a computer. According to the result the children liked the interface that uses RF best. This can be explained in several ways though. The children find it hard to express their opinions. When they were asked: "Why did you like that system best?" they could not explain. Perhaps they just liked the smileys.

7. FUTURE RESEARCH

Although RF seems to be a promising technique for aiding children in their search, more research is required to validate our suspicions. Except for redoing the experiment with the change suggestions mentioned in the previous section, an interesting follow-up study would be to compare the results of children of different grades. Older children have worked with computers more often and they may be able to understand RF more quickly than the children of grade four. Also interesting is to measure the level of understanding of the concept RF. Our measurements were taken by asking children whether they found one interface 'nicer' or 'easier to work with' than the other, but perhaps better ways of evaluation can be used [13, 1].

8. ACKNOWLEDGMENTS

We would like to thank the 2007 fourth graders and their teachers of the RC Paulus primary school in Enschede, the Netherlands, for participating in our research.

9. REFERENCES

[1] W. Barendregt, M. Bekker, and M. Speerstra. Empirical evaluation of usability and fun in computer games for children. *Proceedings of the IFIP 8th International Conference on Human-Computer Interaction INTERACT-03*, 3:705–708, 2003.

[2] M. Ben Moussa, M. Pasch, D. Hiemstra, P. van der Vet, and T. Huibers. The potential of user feedback through the iterative refining of queries in an image retrieval system. In *Proceedings of the fourth International Workshop on Adaptive Multimedia Retrieval (AMR), Lecture Notes in Computer Science 4398*. Springer, 2007.

[3] T. Deselaers, D. Keysers, and H. Ney. Fire - flexible image retrieval engine: Imageclef 2004 evaluation. In *Proceedings of the CLEF 2004 Workshop*, pages 535–544, Bath, UK, September 2004.

[4] T. Deselaers, T. Weyand, D. Keysers, W. Macherey, and H. Ney. Fire in imageclef 2005: Combining content-based image retrieval with textual information retrieval. In *Proc. Working Notes of the CLEF 2005 Workshop*, 2005.

[5] L. Earlbaum. Analysis of variance. *Journal of Consumer Psychology*, 10(1&2):5–35, 2001.

[6] A. Greenwald. Within-subjects designs: to use or not to use? *Psychological Bulletin*, 83(2):314–320, 1976.

[7] C. J. Hamelink. *Communicating in the Information Society*, chapter Human Rights for the Information Society. Geneva: UNRISD, 2003.

[8] L. Hanna, K. Risden, and K. Alexander. Guidelines for usability testing with children. *interactions*, 4(5):9–14, 1997.

[9] E. Hatcher and O. Gospodnetic. *Lucene in Action*. Manning Publications, 2004.

[10] D. A. Hull. Using statistical testing in the evaluation of retrieval experiments. In *Research and Development in Information Retrieval*, pages 329–338, 1993.

[11] O. of the United Nations High Commissioner for Human Rights. Convention on the rights of the child. Adopted and opened for signature, ratification and accession by General Assembly resolution 44/25 of 20 November 1989, September 1990. `http://www.unhchr.ch/html/menu3/b/k2crc.htm`.

[12] A. Prasad and D. Patel. Lucene search engine: An overview. In *DRTC-HP International Workshop on Building Digital Libraries using DSpace*, page paper I, DRTC, Bangalore, India, 7th - 11th March 2005.

[13] J. Read, S. MacFarlane, and C. Casey. Measuring the Usability of Text Input Methods for Children. *HCI2001, Lille, France*, 1:559–572, 2001.

[14] S. Robertson. *Evaluation in information retrieval*, pages 81–92. Springer-Verlag New York, Inc., New York, NY, USA, 2001.

[15] R. C. Veltkamp and M. Tanase. Content-based image retrieval systems: A survey. Technical Report UU-CS-2000-34, Department of Computer Science, Utrecht University, October 2002.

[16] I. H. Witten, A. Moffat, and T. C. Bell. *Managing Gigabytes: Compressing and Indexing Documents and Images*, chapter Indexing, pages 72–115. Morgan Kaufmann Publishers, San Francisco, CA, 1999.