# Context Search Enhanced by Readability Index

Pavol Navrat, Tomas Taraba, Anna Bou Ezzeddine, and Daniela Chuda[1]

**Abstract** Context search is based on gathering information about user's sphere of interest before the search process. This information defines context and augments search query in subsequent phases of search to attain better search results. There are several basic methods for context enhanced searching. The main idea of them is to extract keywords of the found document and compare them with those from the context. The keyword recognition process is difficult to describe in a formally complete way. The context search based on it may, but also may not attain better search results. We propose a modification of the context search by broadening the scope of kinds of attributes, i.e. to consider also implicit attributes rather than only keywords (i.e., explicit ones). Our hypothesis is that it will enable the context search method to fetch more relevant results. This work analyzes the relation between readability index of a document and its content. Improvement idea is based on the kind of knowledge which is difficult to express by keywords, e.g., the fact that user is looking for fairy tales rather than science articles.

## 1 Introduction

A person formulating search query on the web is doing this with some context in mind. Let us call him or her the interested person (IP), since he or she is interested in some specific information at one moment.

In many cases IP works with other documents, files or web pages. Information gathered from these documents can be used to find out the IP's actual scope of interest. This information can be stored in some specific structure and then used to receive better search results.

Many ways to augment search query by context can be explored, but the common base of them seems to be using document keywords. In several related

[1] Slovak University of Technology, 842 16 Bratislava, Slovak Republic email: navrat@fiit.stuba.sk, tomastaraba@wms.sk, ezzeddine@fiit.stuba.sk, chuda@fiit.stuba.sk

*Please use the following format when citing this chapter:*

Navrat, P., Taraba, T., Ezzeddine, A.B. and Chuda, D., 2008, in IFIP International Federation for Information Processing, Volume 276; *Artificial Intelligence and Practice II*; Max Bramer; (Boston: Springer), pp. 373382.

works [1, 2] keywords are used to rate the relation between search result and context using ontology of keywords [3], or simply by using keywords in context vector [1]. Some experiment with semantic query expansion [4]. Keywords from context can be inserted into search query string and sent to standard search engine. Another approach is to submit the original query first, and then reorder the resulting documents according to the rate how they fit the context. There can be made various combinations of query and keywords from context, send them to multiple search engines and then aggregate several sets of results.

Each kind of method may have advantages and disadvantages, but in some sense they are similar. We propose, however, to broaden the concept of context. We modified the keyword-based approach hypothesizing that considering also other attributes rather than only keywords can result in better search precision [5]. We made series of experiments with on-line database and verified that it tends to attain better search results when using the multi-attribute context.

While the experiments have shown that our idea works quite fine with an on-line database where many attributes are present [6], we were not sure how this can be used in the web. There are not so many attributes identifiable in the web like in some on-line database. For example, in the online database we tracked the author of each document in context. We assumed that the name of an author is an important information. It can be quite safely assumed that one author writes on a small set of themes, and one theme is written about by a not so big set of authors. But consider a web page. It is hard to define the exact process to extract the name of the author of each webpage; in most cases we can tell that extraction is impossible.

The rest of the paper is structured as follows. In Section 2 - Motivation we outlined the motivation why we have some concerns about classic context search based on keywords. We formulated a sample problem and described it. The proposed approach to solve the problem, which is based on our improvement idea is described in Section 3 - The Proposed Modification. In section 4 we formulated a hypothesis and attempted to verify it by series of experiments. The results and their consequences have inspired us to proceed in the research, concentrating mainly on readability index values. Their possible interpretation is described. Having gathered sufficient research results, we were able to draw a conclusion and we suggest some future work in section 5 - Conclusion.

## 2 Motivation

As a motivating example for our work, let us consider this problem: Suppose IP looking for fairytales on the web. IP read two documents. The first one was Little Red Riding Hood, and the second one The Wandering Egg.

There are some concerns about the relevance of the classical context search. Ensuring relevance in this case means to get results of documents containing fairy

tales for children under 5 years instead of getting results of Cookery-book for hunters, which can contain more keywords “egg, roe, hunter, food” than a common fairy tale.

Troubles With Keywords. When one compares two different fairy tales, it can be quite hard to find at least a few common keywords. For example, let us consider Little Red Riding Hood in the context of having keywords: little, grandmother, wolf, wood, door, hunter, etc. In our context search with the query Cinderella, the approach is to prefer search results containing fairy tale with keywords: prince, girl, time, sisters, shoe, three nuts, etc. rather than documents about Cinderella band containing keywords: band, tour, metal, rock, album, etc.

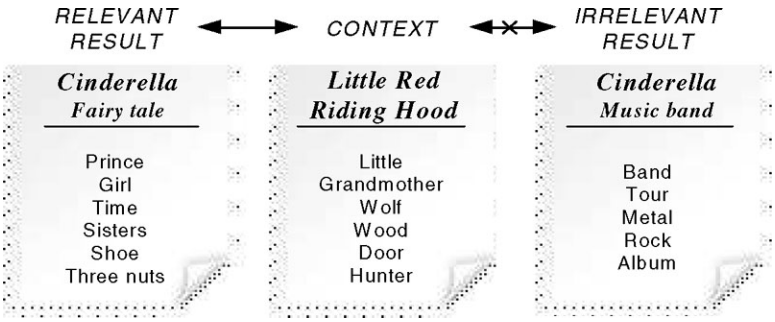



**Fig. 1** Comparison of relevant and irrelevant result keywords

As one can see in Fig. 1, there is no relation between the two fairy tales with regards to keywords, because each one tells a different story. Keywords do not tell us that Cinderella story is relevant to Little Red Riding Hood as both are fairy tales and Cinderella band presentation isn’t relevant because it is not a fairy tale.

Improvement Idea. Going out from previous work described in introduction, the improvement idea is to use some combination of explicit and implicit attributes to rank the search results. We introduce implicit attributes which can help us to guess whether the found document is fairy tale, romantic novel, presentation of the band, advertisement, scholarly article, etc.

If we were in an on-line database, and we had the name of the author, we could use it. Since only a relatively small number of authors write fairy tales, the author attribute can be effective in search. But on the web, there is usually no such explicit attribute available. Unfortunately, we cannot guarantee that every document has the name of the author included in it somewhere. Besides that, there is also not an explicit attribute in every text that would inform us how the text has been written.

Flesch Readability Index. The improvement idea is based on a formula defined by Flesch [7]. Let us consider user’s age and level of ability to read and comprehend the text. There is a method to categorize readability of a given text by

the Flesch readability index (FRI). This method is used for estimating the reading comprehension level necessary to understand a written document. For a given document, the Flesch readability index is an integer (0-100) indicating how difficult the document is to understand, with lower numbers indicating greater difficulty.

$$FRI = 206.835 - (1.015 x \frac{number of\ words}{number of\ sentences}) - (84.6 x \frac{number of\ syllables}{number of\ words}) \quad (1)$$

Flesch categorized readability indexes into 7 educational levels and describe Flesch Reading easy scale. In 1948, Flesch published [7] the results of his study of the editorial content of several magazines and he found that about 45% of the population can read The Saturday Evening Post, nearly 50% of the population can read McCall's, Ladies Home Journal, and Woman's Home Companion, slightly over 50% can read American Magazine and 80% of the population can read Modern Screen, Photoplay, and three confession magazines. For example comics have readability index 95, New York Times 39, Auto Insurance 10.

What Improvement Do We Suggest*?* Let us consider that IP has never read a document classified as more difficult than a document comprehensible by a high school student. Is there a reason to return him also documents understandable solely by a law school graduate? In our example, why to return e.g. academic analytical studies on the Cinderella fairy tale, or why to return lyrics of the Cinderella music band songs, or why to return any other documents for which the readability index indicates that they are not fairy tales?

## 3 The Proposed Modification

To develop our improvement idea, it should be incorporated in some existing context search method. The improvement means essentially adding more attributes into context and considering more attributes in the process of result selection. A suitable way is to change the ranking function of the Rank-Biasing method [1]. In general, this method uses context to change the score of every search result and then sorts the results by the new score. Finally, the new set of results will contain all the results from original set (as if the context was not used) but the results are sorted in a new order with more relevant results on the top. The relevancy of result is determined by how much the result fits the context – how many keywords from the context are contained in the search result.

Our modification proposes to change the rank function. We use a context of keywords and a context of readability index values. The relevancy of a result depends on how much the result fits the keywords in the context, but also on how much it fits the readability indexes in the context.

Context Acquisition. The context is acquired while the user is browsing [2]. For example, we can acquire it by tracking every click (i.e, loading of a document

specified by URL) which the IP does and store the information from the tracked documents in the context. The context is represented by a vector, which contains two kinds of dimensions: dimensions of keywords and dimensions of readabilities. A dimension is represented by a vector of attribute values (attributes being either of keyword or readability index kind). Each value has a score, which determines how many times the keyword occurred in all documents of context, or how many documents had the given readability index:

$$C = \begin{Bmatrix} D_1 \\ D_2 \\ \dots \\ D_N \end{Bmatrix}, \qquad D_i = \{(v_1 \to s_1),(v_2 \to s_2),\dots,(v_j \to s_j),\dots,(v_M \to s_M)\}, \tag{2}$$

where $C$ is context vector, $D_i$ is a dimension of the context vector, $v_j$ is a value of an attribute represented by the given dimension, $s_j$ is score of value $v_j$.

Search Process. The search process is initiated by a search query sent by IP. First, the search query is sent to a standard search engine to receive a set of results. Next, the ranking score is modified for each result from the result set. The ranking score given by the standard search engine is a number indicating how much the result fits the query. We modify it to indicate how much the document fits not only the query, but also the context. Having modified the score of each result, the result set will be sorted.

Ranking Function. We propose in our modification to change the ranking function. While the original ranking function calculates the rank score only using context of keywords, the modified function calculates it by considering also the context of readabilities. There are several possibilities how to combine the two rankings. In [3], an additive formula is used, but we found it more useful to use a multiplicative one. The final score is calculated by the formula $R = R' * R_K * R_R$. In this formula, $R'$ is the original score assigned by the standard search engine, $R_K$ is the rank factor calculated when ranking by keywords and $R_R$ is rank factor calculated when ranking by readability index. For every rank factor we require it to have values from $\langle 1,2 \rangle$ interval, so every value has to be mapped into this interval.

Ranking by Attribute of Keywords. The rank factor for an attribute of keywords says how much the keywords in document fit the keywords in context. In detail we have two sets: set of document keywords and set of context keywords. We require the rank factor to have value of 2 if every context keyword is found in the given document and have value of 1 if no context keyword is found in the given document.

$$R_K = 1 + \frac{sum(score\_of\_every\_context\_keyword\_found\_in\_document)}{sum(score\_of\_every\_context\_keyword)} \tag{3}$$

Ranking by Readability Index Attribute. A readability index attribute has a numeric value. In the process, the relevant document is the document, which has the readability index value within the specific interval. This interval is determined by readability index values in context. For example - if the readability index vector in the context is: {70, 72, 74, 76}, then the interval would be <70; 76>.

The rank factor for the readability index attribute has the value of 2 when the readability index of given document is in the centre of the interval and value of 1 when the value is outside far from the borders of the interval.

## 4 Experimental Testing

We performed series of experiments to test the improved method. We wanted to see how ranking by readability influences the precision of the search. We devised three cases and performed simulations in these three areas of user's interest:

Fairy Tales. Relevant results are pages containing text of a searched fairy tale. Among irrelevant results, we also marked pages containing information on movies, bookstore catalogues, and other.

Population Diseases. We searched for documents describing a given population disease. As relevant we marked pages containing statistical studies related to the given population disease, scholarly articles, articles on research in this area, popular articles explaining the terms related to given disease. As irrelevant we marked pages presenting, propagating, or selling medicaments, presentations of health organizations, centers, and founds.

Predators. As relevant we marked pages containing some information about the given predator, group of predators or presentations of Zoo's. As irrelevant we marked pages using the name of predator in meanings other than that of an animal (e.g., a car, computer, etc.), presentation of companies using the name of the predator as the brand, pages of conservation organizations.

For each area we collected hundreds of results by series of queries and marked them as relevant or irrelevant. Then we added each relevant document into context in successive steps simulating user's clicks on relevant documents. After each addition we recalculated score of each result, sorted a result set and calculated the precision of the search for every configuration of context vector.

Results. As we tested the method in three different areas, we have three different experimental results. We tracked how the precision has changed considering the change of context size.

In all the figures, the KW curve represents the precision of method using only keywords in ranking, the curve marked as RI represents precision of method ranking only by readability index attribute. The curve marked as KWRI represents the combination of both kinds of attributes aggregated together. SS represents the precision of standard search engine. Precision of the standard search engine does not change when the context size grows.

In the first case – searching fairy tales – there is a very significant improvement. The modified method produces more precise results than the original method. As we can see in Fig. 2, major improvement is caused by using the readability index attribute in ranking.

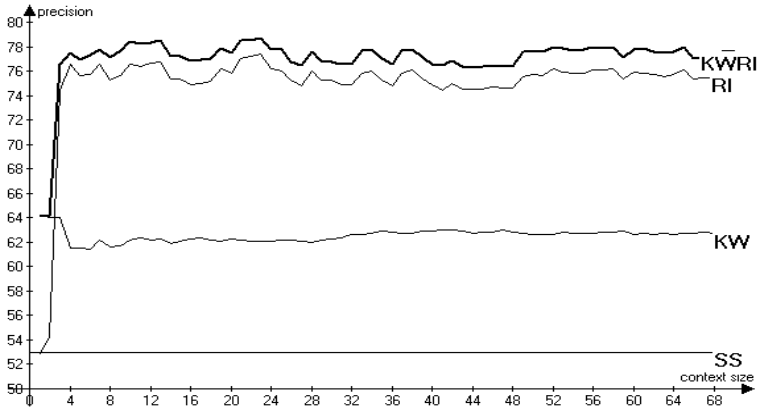



**Fig. 2** Precision in the case of fairy tales

In the second case – population diseases – the modified method produces more precise results, but it does not yield a significant improvement (Fig.3). The average difference in precision between the original and modified method is around 3 - 4%.

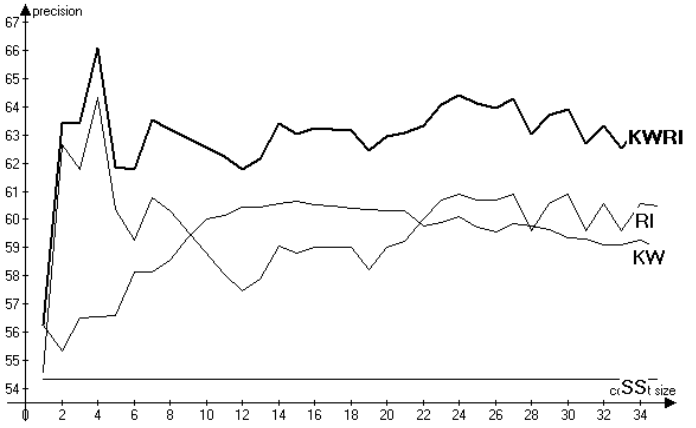



**Fig. 3** Precision in the case of diseases

In the third case – predators in the nature – the modified method produces more precise results most of the time, but sometimes the precision could be worse than that of the original method (Fig. 4). It was caused by adding documents very different in readability index attribute into context at context size of 30. On the other side, method using only keywords in rank function keeps on the same precision during the context growth. It may mean that keywords are nearly the same and adding each new document into context does not strongly influence the vector of keywords. In combination of both methods, the precision is sometimes better than the precision of the original method (KW), but sometimes the search produced worse results.

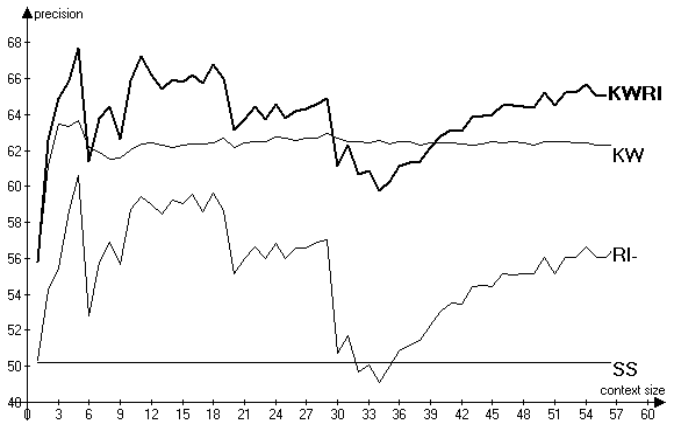



**Fig. 4** Precision in the case of predators

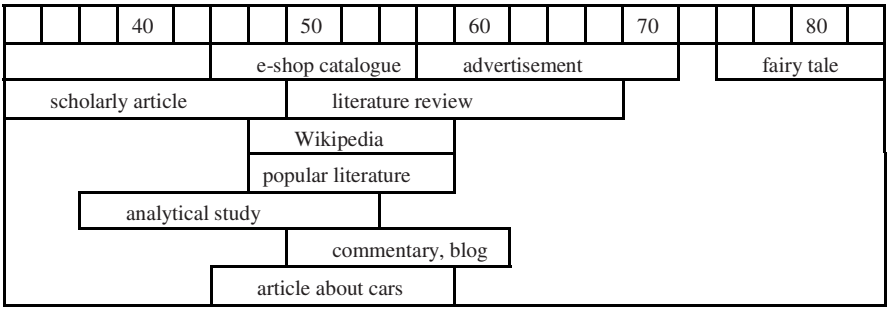
In all three figures we can see the cold-start problem. The RI method does not yield good results when the size of the context is small. While the KW method works fine with 1 or more documents in the context, the RI method requires having a minimum of 2 documents in the context. This is caused by readability index value interpolation with requirement minimum of 2 values to find the interval border values.

Additional Research: What Was Wrong? As we can see in the graph representing experiment results, the method works fine for searching fairy tales, but does not work so well for the other areas. We tried to find out where the problem is and why it occurs.

After a small analysis of acquired pages, we found out the problem. We manually classified the type of the content of every ranked page provided as search result. We classified 1,000 pages into 10 groups of content type and then statistically determined the interval of readability index values for each group. Results of this experiment are represented in Table 1.

The conclusion is - Fairy tales are very special documents. Their readability index is very high due to the simplicity of the text. In Table 1 we can see that fairy tales are sufficiently well separated from other documents. The case with the popular literature on predators is not similar, which is overlapped by several other different genres (articles about cars, advertisements, e-shop catalogue, blogs). We conjecture this is the fundamental reason why fairy tales are easily identifiable by the readability index but e.g. popular articles on predators are not.

**Table 1** Overlapping of the readability index values for different types of text




A popular article on a predator is too common in comparison to a fairy tale, or a scholarly article, which are quite specific kinds of texts. In the table above we can see that very common and ordinary documents have readability index values around the value of 50. Experiments show the average value of the readability index is 48. Based on this research we assume that major “common” documents have readability index around 50. Other documents with much higher or much lower values are more “specific”.

## 5 Conclusion

This work is focused on investigation in the relation between readability index and the character (related also to elements of style, genre) of the web page. We tried to find out how the readability index can be used to attain better search results in context search.

What Did We Find Out? In general, our improvement idea works fine, but the degree of improvement depends on the type or character of documents. Readability index is related to the character of the text, but generally it may not be sufficiently restrictive to allow the desired identification of the sphere of interest of the query in all cases. Relying on the readability index in the context search was not effective when the user is interested in very common things. It is particularly effective when the IP is interested in not so common types of texts, e.g. fairy tales

or scholarly articles. In those cases we can quite safely tell that documents have similar content and the user has some specialized sphere of interest.

Future Research Work. There may be other ways to overcome the problem of ordinary texts. When dealing with our motivating problem, there is no need to consider the interval in which the readability index belongs. A better indication may be the distance of readability index of current document from the "general centre". In other view, in the context the readability index measures how different the documents of the context are from the „ordinary ones". As it can be seen from the table of readability index of different types of web pages, ordinary documents are pages with the readability index around the value of 50. The more specific the document is, the bigger is the distance between its readability index and the value of 50.

## ACKNOWLEDGMENTS

This work was partially supported by the Slovak State Programme of Research and Development "Establishing of Information Society" under the contract No. 1025/04 and the Scientific Grant Agency of Republic of Slovakia, grant No. VG1/3102/06.

## References

1. Kraft R., Chang C.C., Maghoul F., Kumar R.: Searching with Context. In Proceedings of the 15th International Conference on World Wide Web WWW '06. Edinburgh, pp. 477- 486, ACM Press, (2006).
2. Bharat K.: SearchPad: Explicit Capture of Search Context to Support Web Search. In Proceedings of the 9th International World Wide Web Conference, pp. 493-501, Elsevier, Amsterdam (2000).
3. Challam V., Gauch S., Chandramouli A.: Contextual Search Using Ontology-Based User Profiles. In Proceedings of the 8th Large-Scale Semantic Access to Content Conference (RIAO'2007), Pittsburgh (2007).
4. Malecka J., Rozinajova V.: An Approach to Semantic Query Expansion. In: Tools for Acquisition, Organisation and Presenting of Information and Knowledge. Research Project Workshop Proceedings, pp. 148-153, STU, Bratislava (2006).
5. Clarke S., Willett P.: Estimating the Recall Performance of Search Engines. ASLIB Proceedings, 49 (7), pp. 184-189 (1997).
6. Navrat P., Taraba T.: Context Search. In: Y. Li, V.V. Raghavan, A. Broder, H. Ho: 2007 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology (Workshops), Silicon Valley, USA, pp. 99-102, IEEE Computer Society, (2007).
7. Flesch R.: A New Readability Yardstick. Journal of Applied Psychology, Vol. 32, pp. 221-233, (1948).