Adaptive Ranking of Search Results by Considering User's Comprehension

Makoto Nakatani Adam Jatowt Katsumi Tanaka
Graduate School of Informatics, Kyoto University
Yoshida-honmachi, Sakyo, Kyoto, 606-8501 Japan
{nakatani, adam, tanaka}@dl.kuis.kyoto-u.ac.jp

ABSTRACT

Given a search query, conventional Web search engines provide users with the same ranking although users' comprehension levels can be different. It is often difficult especially for non-expert users to find comprehensible Web pages from the list of search results. In this paper, we propose the method of adaptively ranking search results by considering user's comprehension level. The main issues are (a) estimating the comprehensibility of Web pages and (b) estimating the user's comprehension level. In our method, the comprehensibility of each search result is computed by using the readability index and technical terms extracted from Wikipedia. User's comprehension level is estimated by the users' feedback about the difficulty of search results that they have viewed. We implement a prototype system and evaluate the usefulness of our approach by user experiments.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Retrieval models

General Terms

Algorithm

Keywords

Web search, adaptive ranking, comprehensibility, Wikipedia mining, user interaction

1. INTRODUCTION

Web search engines have become frequently used for acquiring information over the Internet. According to the online survey that we have conducted on 1000 respondents in Japan [18], it was found that users often search the Web because they require the explanation of unknown or unfamiliar keywords[1]. In that situation, users usually desire Web pages including comprehensible information about their search queries. However, it is difficult for users to efficiently find comprehensible Web pages with conventional search engines because of two problems.

The first problem is that the ranking in conventional Web search engines does not reflect the comprehensibility of pages. In consequence, search results often contain pages that are difficult to be understood by non-expert users, especially for queries containing technical terms, for example, the ones in the medical, financial and astronomy areas. Consider the following passages that is contained in search results acquired by issuing query "black hole" to a conventional Web search engine.

1. According to Einstein's *theory of general relativity*, a black hole is a region of space in which the *gravitational field* is so powerful that nothing, including *electromagnetic radiation* (e.g. *visible light*), can escape its pull after having fallen past its *event horizon*.

2. A black hole is a region of space whose *gravitational force* is so strong that nothing can escape from it. A black hole is invisible because it even traps light.

Passage 2 is intuitively easier to be understood for non-expert users than Passage 1. However, the page containing Passage 1 is actually ranked higher in search results than the one with Passage 2. In such cases, users must manually find out comprehensible Web pages from search results. This task is often tiresome as search results consist of limited information such as titles and snippets. Also, comprehensible pages may be within lower search results, thus, finding them may be difficult. The second problem is that users do not have any effective ways to convey their comprehension level to the search engine. In many cases, search results are less accurate for queries containing keywords that explicitly represent the comprehensibility of Web pages (e.g. "easy" and "introduction"). This is because search engines based on the query relevancy just provide users with Web pages containing query keywords and thus may not return content appropriate to users' intent. Both of above two passages about black hole cannot be actually acquired by search queries such as "black hole easy".

To solve these problems, in this paper we propose the method of re-ranking search results by considering the users' comprehension level about the search topics. In our previous work, we introduced the concept of comprehension-based Web search [19]. We defined it as the Web search that outputs search results considering user's comprehension level about search topics.

Whether a user can comprehend the content of a Web page is the result of the interaction between the user and the page. Therefore, both the comprehensibility of Web pages themselves and the user's comprehension level should be taken into consideration in order to realize the comprehension-based Web search. That is, the

[1] 46.0% of respondents selected it as a first reason.

ICUIMC'10, January 14–15,2010, Suwon, Korea.

main issues for comprehension-based Web search are (a) to estimate the comprehensibility of Web pages and (b) to estimate the user's comprehension level. Both of them are essentially challenging tasks because they are multi-dimensional concepts that involve many different aspects. In our approach, the comprehensibility score of each search result is first computed by combining two text measurements, document readability and document speciality [19]. Then the user's comprehension level is interactively presumed by the user feedback about the comprehensibility of Web pages that he/she has viewed, and the whole search results are re-ranked based on the estimated user's comprehension level and the comprehensibility score of each search result. In this paper, we apply this adaptive ranking method to the definition search and introduce a prototype system of comprehension-based Web search. Note that although the target language is Japanese, our proposed method can be easily applied to other languages.

The contributions of this paper are the followings:

- We introduce a general model of the comprehension-based ranking of search results.
- We propose the method of estimating the comprehensibility of search results by combining two text measurements, readability and speciality. Our method can be applied independently of the domain of the search query.
- We implement the system that searches for Web pages containing the explanation of the query keyword and that re-ranks search results according to the user's feedback about the difficulty of Web pages that the user has viewed.

The remainder of this paper is organized as follows: In Section 2, we propose the comprehension-based ranking model. In Section 3, we describe the method of calculating the comprehensibility score of Web pages in search results. In Section 4, we introduce the prototype system of comprehension-based Web search and present the re-ranking method by using the feedback information from users. Section 5 reports experimental settings and results. In the next section, we discuss the further potential and problems for comprehension-based applications. In Section 7, we describe related works. Finally, Section 8 provides conclusion and outlines our future work.

2. COMPREHENSION-BASED RANKING MODEL

In this section, we propose the comprehension-based ranking model for Web search. The purpose of comprehension-based ranking model is to rank Web pages based on the correspondence with the users' comprehension level and is different from traditional relevance-based ranking models. The input of comprehension-based ranking is actually a set of Web pages that are relevant to the search query. In this meanings, our comprehension-based ranking model is configured on the query-relevancy model. Whether a user can comprehend a Web page depends on both the user and the Web page. Although the user's factor and the Web page's factor are closely related to each other, these two factors are divided in our proposed model in order to simplify the problem.

The comprehensibility of Web pages is multi-dimensional concept and can be evaluated from various viewpoints such as content, presentation style and media richness. In our model, the comprehensibility of Web pages is formulated by simply combining multiple comprehensibility metrics $\{f_{c_1}, f_{c_2}, ..., f_{c_n}\}$ as follows:

$$f_c(d) = \sum_{i=1}^{n} \alpha_i f_{c_i}(d) \quad (1)$$

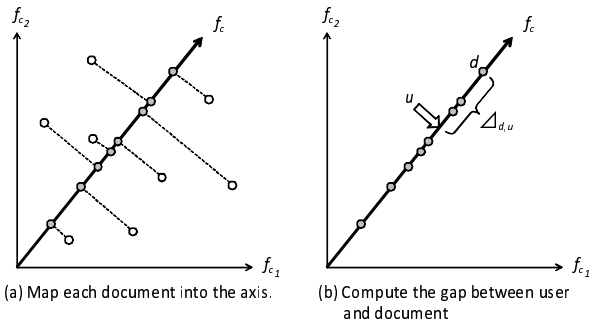

Figure 1: Concept of comprehension-based ranking model.

where α_i is the weighted parameter for each comprehensibility metric. Web pages can be mapped into one axis by using this evaluation function f_c and can be sorted as shown in Figure 1 (a). If the user comprehension level is also mapped into the axis by some methods and is represented as u, the gap between the user level and the comprehensibility of a Web page d can be calculated by the following equation:

$$\Delta_{d,u} = |f_c(d) - u| \quad (2)$$

We assume that the smaller $\Delta_{d,u}$ is, the more appropriate the comprehensibility of the Web page is for the user. To perform comprehension-based Web search, all search results can be re-ranked in an ascending order of $\Delta_{d,u}$.

3. DOCUMENT COMPREHENSIBILITY

In this section, we describe the method of estimating the comprehensibility of Web pages. We focus on two text comprehensibility measurements: *document readability* and *document speciality*. Document readability is an indicator for predicting how easily the document can be read and is calculated by using the surface features of documents. On the other hand, document speciality is our proposed indicator and measures how many technical terms related to a search query are contained in the document. Given a set of Web documents containing query keywords, our method provides each document with a comprehensibility score calculated by combining two text comprehensibility metrics.

Our approach is similar to the concept-based document readability proposed by Xin et al [30]. They focused on medical documents and evaluated document readability by using a medical thesaurus, Medical Subject Headings (MeSH)[2]. However, for arbitrary queries, domain knowledge is necessary for evaluating comprehensibility of Web pages because Web search engines unlike vertical search engines must output search results for queries in many different domains. Using different thesauruses for every query is infeasible because their structures are not standardized, and because appropriate thesauruses may not necessarily be available. To satisfy this demand, we exploit the world's largest human knowledge base, the Wikipedia[3]. In our approach, we acquire technical terms related to a search query by analyzing the link and category structure of Wikipedia.

3.1 Document Readability

Readability is one of the significant factors of the comprehensibility of documents. Readability was defined as "the ease of under-

[2]http://www.nlm.nih.gov/pubs/factsheets/mesh.html
[3]http://en.wikipedia.org/

standing or comprehension due to the style of writing" in [13].

Many readability indexes have been proposed and have been commonly used in a field of education. There are formula-based approaches and statistical language model approaches for predicting the document readability. Traditional readability indexes such as Gunning-Fog Index, ARI [28] and Dale-Chall Readability Index [6] are formula-based approaches and provide numeric scores which represent the requisite reading level for understanding documents. Most of formula-based readability indexes are calculated by the syntactic measures such as the count of syllables, words and sentences. Recently some researchers have proposed statistical language modeling approaches for estimating document readability [27, 5, 25]. In these approaches, readability is estimated within the classification framework. That is, a class which corresponds to a grade level of readability is defined as a sample corpus, and the classifier determines the class to which a given text is the most similar. Formula-based readability indexes have the advantage of being easily calculable by using only syntactic measures. However, readability measures that do not require sentence analysis are preferable as Web pages have many incomplete sentences and non-regular text fragments, such as titles, itemized lists, inline figures, and URLs. Therefore, we use a statistical language model approach that is less affected by the document presentation style.

For Japanese texts, few readability measures have been proposed. In this work, we utilize *Obi* [4] [24], a readability analyzer of Japanese texts based on a statistical language model, for measuring the readability of Web pages. For a given text passage, the readability analyzer determines the grade level to which the passage is most similar by using character-unigram models, which are constructed from the educational textbook corpus. The *Obi* program outputs an integer between 1 and 13, which indicates a Japanese school grade as follows:

$$\begin{aligned} 1-6 &: \text{elementary school (6 years)} \\ 7-9 &: \text{junior high school (3 years)} \\ 10-12 &: \text{high school (3 years)} \\ 13 &: \text{over high school} \end{aligned}$$

A Web page indicating a low readability score by *Obi* is deemed to be comprehensible. Thus we define a comprehensibility metric DRS based on document readability as the following equation:

$$DRS(d) = \frac{14 - Obi(d)}{13} \tag{3}$$

where $Obi(d)$ is an integer value acquired by inputting a document d into the Obi program. The value of DRS decreases in proportion to the output of Obi program. Note that DRS is a query-independent measure of the comprehensibility of Web pages.

3.2 Document Speciality

Document speciality is another feature of document comprehensibility. Intuitively, it measures how many technical terms related to a search query are contained in the document. Here, technical terms related to a search query are the terms that occur mostly in the domain of the search query and rarely outside of it. For example, when a search query is "black hole", frequent terms appearing only in the domain of astronomy or astrophysics such as "theory of relativity" and "dark matter" are extracted as technical terms.

In this work, we utilize the category and link structure of Wikipedia for extracting technical terms regardless of the query domain. In Wikipedia, an article is linked to other articles with the aim of helping the understanding of the concepts that emerge in the article. Our basic idea is that Wikipedia concepts rarely linked from Wikipedia concepts outside the domain of search query are assumed to be technical terms. On the other hand, Wikipedia concepts that are frequently linked from both of inside and outside the query domain are considered to be general terms thus not technical terms related to the query. In Figure 2, "theory of relativity" is regarded as a technical term and "Japan" is not.

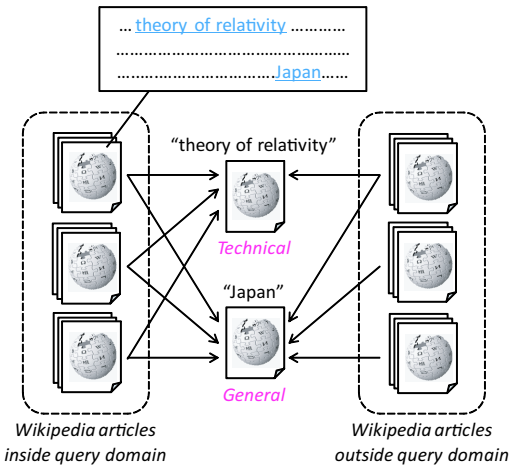

Figure 2: Overview of the link distribution analysis for technical terms extraction. Each arrow represents a link between Wikipedia articles.

Our proposed method for technical terms extraction consists of the following three processes: (i) query domain detection; (ii) candidate terms extraction and (iii) link distribution analysis. First, search results are mapped into related Wikipedia categories, namely *query domain*, using a semantic interpreter built in advance. Next, the candidates of technical terms are extracted from each Web page. Finally, we calculate the degree of terms to be technical terms by analyzing the distribution of link frequency in Wikipedia. Below we describe the details of technical terms extraction and the method of measuring the document speciality of each page by using the extracted technical terms.

3.2.1 Query Domain Detection

First of all, we provide the way to build semantic interpreter for detecting the domain of a search query. Gabrilovich et al. [9] proposed Wikipedia-based explicit semantic analysis (*ESA*), which maps a fragment of text into a weighted sequence of Wikipedia concepts ordered by their relevance to the input. Instead of the original *ESA*, we propose the category-based explicit semantic analysis (*C-ESA*) that maps a text fragment onto a weighted sequence of Wikipedia categories. In *C-ESA*, we combine articles included in a single Wikipedia category and regard it as one document. We then extract noun terms from each such connected document using *MeCab*, a morphological analyzer for Japanese language[5]. Every Wikipedia category is then represented as a vector of terms, and entries of these vectors are weighted using the *TFIDF* scheme. To speed up semantic interpretation, we build an inverted index, which maps each word into a weighted sequence of Wikipedia categories in which it appears.

Given a set of document, each document is mapped into a

[4]http://kotoba.nuee.nagoya-u.ac.jp/sc/readability/obi_e.html

[5]http://mecab.sourceforge.net/

Table 1: First ten Wikipedia categories in sample interpretation vectors.

#	input: "*black hole*"	input: "*RNA*"
1	Black holes	RNA
2	Sanseido	Nucleic acids
3	Dark matter	DNA
4	Compact stars	Molecular biology
5	Neutron stars	Chromosomes
6	Cosmology	Dicarboxylic acids
7	Universe	Genes
8	Gravitation	Peptides
9	Radio astronomy	Cell biology
10	Supernovae	Fluorescent dyes

weighted vector of Wikipedia categories by using the *C-ESA* semantic interpreter. Every Web page is first represented as a feature vector by using *TFIDF* weighting method. In this work, the *IDF* weight of a term w is calculated by the following equation:

$$IDF(w) = \log \frac{N_c}{N_c(w)} \tag{4}$$

where N_c is the total number of Wikipedia categories and $N_c(w)$ is the number of Wikipedia categories containing w. Every page is then mapped into a weighted vector of Wikipedia categories by calculating the cosine similarity between the page itself and every Wikipedia category. Then we obtain the list of weighted vectors $\{\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_n\}$. Finally, we calculate a centroid vector $\overline{\mathbf{v}} = \frac{1}{n}\sum_i \mathbf{v}_i$. Here, we simply regard top K Wikipedia categories in a centroid vector as the query domain. Examples of the semantic interpretation are illustrated in Table 1.

3.2.2 Candidate Terms Extraction

The goal of this step is to extract the candidates of technical terms from the documents of search results. Here each candidate term is represented as one Wikipedia concept, that is the title of a Wikipedia article. First, we extract all possible n-grams from the input Web document and search for all matching n-grams among Wikipedia article titles. Note that some n-grams may be mapped into Wikipedia "Redirect Pages". Wikipedia guarantees that there is only one article for each concept by using redirect pages to link equivalent concepts to a preferred one. For example, a Wikipedia article of "dark matter" is redirected from the articles of "darkmatter", "missing mass" and "black matter" and so on. We use the titles of redirect pages as synonyms for each concept in order to correctly analyze the link distribution in the next step. In this work, if a document contains one of the synonyms of a concept, we regard the concept as a candidate term. In result, we can obtain the list of the candidates of technical terms for every search result.

3.2.3 Link Distribution Analysis

After we have extracted candidate technical terms we proceed to determine which ones are actually real technical terms of the query domains. All Wikipedia articles can be classified into two groups by whether they are included at least in one of the categories within the query domain. W^+ denotes Wikipedia articles included in domain categories, and W^- denotes the remaining Wikipedia articles. The degree of bias of the link occurrence distribution between W^+ and W^- is measured by the χ^2-measure that is how more frequently a given candidate technical term is linked from articles in W^+ as opposed to the ones in W^-. However, note that, when the links to a Wikipedia concept appear rarely in W^+ but mostly in W^-, the χ^2-measure can also have a high value. Therefore, we remove candidate terms that do not satisfy the following condition:

$$\frac{LF(t, W^+)}{|W^+|} > \frac{LF\left(t, W^-\right)}{|W^-|} \tag{5}$$

where $LF(t, W)$ denotes the number of articles that link to the article of a candidate term t and that are contained in the article set W. In this way we receive a set of actual technical terms related to the domain of the query.

3.2.4 Measuring Document Speciality

We provide the method of measuring speciality of search results by using technical terms extracted from Wikipedia. In the above section, terms rarely appearing outside the query domain were regarded as technical terms related to a search query. However, the distribution of actual term occurrence does not necessarily correlate with the term difficulty. For example, "Schwarzschild radius" is deemed to be more difficult than "black hole" despite both of them are technical terms in the domain of astronomy. Thus the document speciality should be concerned with the domain-dependent difficulty of each technical term. We use the link frequency of technical terms in the query domain as a proxy of term difficulty. Moreover, the document length should be considered when computing the document speciality since long documents likely contain many technical terms. We define the document speciality by the following equation:

$$DSS(d) = \frac{1}{\log |d|} \sum_{t \in TT(d)} \frac{1}{\log LF(t, W^+)} \tag{6}$$

where $TT(d)$ represents a set of technical terms that are related to the search query q and that are contained in the document d, and $|d|$ is the length of the document d. $LF(t, W^+)$ is the same one as defined in the above section. Web pages with lower DSS value are regarded as more comprehensible.

3.3 Combined Measure

Lastly, we propose the following equation to compute final document comprehensibility score (DCS) of Web pages, which is a linear combination of document readability and document speciality.

$$DCS(d) = \alpha_1 \cdot DRS(d) + \alpha_2 \cdot DSS(d) \tag{7}$$

Note that this equation is included in the general formula for estimating the comprehensibility of Web pages (Equation 1). As *DRS* is a positive feature of comprehensibility, α_1 should be a positive value. On the other hand, as *DSS* is a negative feature, α_2 should be a negative value. We regard Web pages with higher DCS value as more comprehensible. The document readability and the document speciality are respectively the surface-level metric and the concept-level metric for estimating the comprehensibility of documents, and are expected to supplement each other.

4. COMPREHENSION-BASED SEARCH

4.1 Overview

In this section, we describe the overview of our proposed system *ComprehensionSearch* that is a Web search engine that specializes in the term explanation retrieval. The system outputs the list of Web pages including the explanation of the query keyword such as encyclopedias. A notable function of our system is to re-rank search results by the users' feedback about the difficulty of Web pages that

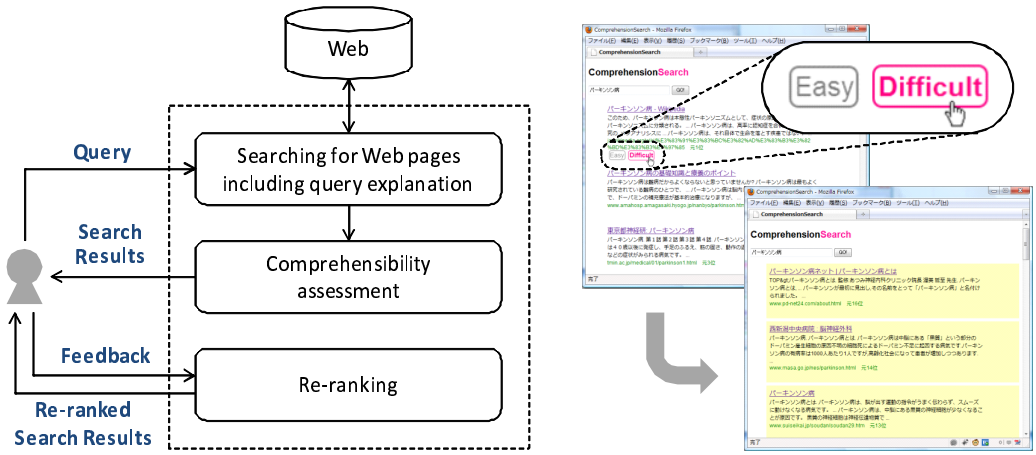

Figure 3: System flow of ComprehensionSearch and its screenshots.

they have viewed. Users can efficiently find comprehensible Web pages only by clicking two kinds of button: "easy" and "difficult". The system flow and its screenshots are illustrated in Figure 3. The system flow is as follows:

1. The user inputs a query to the system.
2. The system receives the search results by using a Web search engine. Then the system extracts only Web pages including the explanation of the query keyword.
3. The system computes the comprehensibility score for each extracted Web page.
4. The system shows the search results to the user.
5. The user selects a Web page among the search results.
6. If the selected search result is not appropriate for the user's comprehension level, the user clicks either the "easy" button or the "difficult" button attached to the search result.
7. The system detects the user's operation and guesses the user's comprehension level. Then the system re-ranks the search results according to this guess and shows the re-ranked search results to the user. Go to 5.

There are three technical issues for ComprehensionSearch: (i) acquisition of Web pages explaining the query keyword, (ii) comprehensibility assessment of Web pages and (iii) re-ranking of search results by the user's feedback. Below we describe the details of each step.

4.2 Acquisition of Web Pages Explaining Query Keyword

We describe here the method of acquiring Web pages explaining the query keyword by using a Web search engine. In this work, Web pages including the definition sentence of the query keyword are regarded as ones explaining the query keyword.

Many methods for extracting term definitions have been proposed in text mining [8, 15, 29]. In these researches, term definitions are extracted by using several typical syntactic patterns such as "X is a $*$" or "X is one of $*$", where X is a target term and '$*$' denotes a word string containing one or more words. In [8], some syntactic patterns for Japanese texts have been proposed, e.g. "X *toha* $*$ *dearu*. (X is a $*$.)" or "$*$ *wo* X *to-iu*. (X is defined as $*$.)". We utilize these syntactic patterns to test whether a Web page contains the explanation of the query keyword. Below we describe the detailed method of acquiring Web pages explaining the query keyword.

1. Given a query q, issue the search query "q OR q *toha* OR q *ha*" (in Japanese) to a Web search engine and receive the list of search results.
2. Download Web pages from the search results and discard HTML tags such as `<strong>` and `<font>` used for decoration.
3. For each page, test whether the page contains the typical syntactic patterns for the term definition, such as "q *toha* $*$ *dearu*".
4. Filter out Web pages not containing the query's definition.

In result of this step, we obtain the list of Web pages containing the explanation of the query keyword.

4.3 Comprehensibility Assessment

Next, the system gives the comprehensibility score to each Web page as follows:

1. Extract the body content from each Web page by several heuristic methods as a pre-processing. One method is to extract the HTML elements tagged with the same tags as the element that contains the definition sentence.
2. Compute the comprehensibility score of each Web page by our proposed method described in Section 3. Here the score is normalized to $[0, 1]$.

After the comprehensibility assessment, the search results are shown to the users with the original ranking kept.

4.4 Re-ranking Method

Here we provide the method for re-ranking search results by using the user feedback about the difficulty of Web pages that he/she has viewed. Our objective is to highly rank Web pages corresponding to the user's comprehension level. That is, the goal is to efficiently predict the user's comprehension level (u in Equation 2)

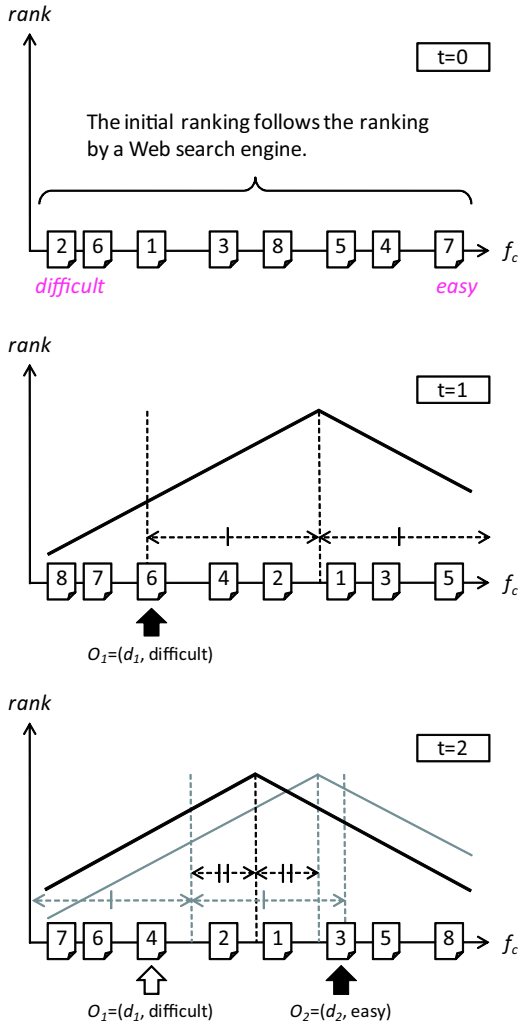

Figure 4: Overview of re-ranking method. The number attached to Web page represents its rank in each step.

from a few operations. In our proposed method, the expected value of the user's comprehension level is computed after every operation by using an approach based on binary search, and search results are re-ranked in an ascending order of the gap between the expected value and the comprehensibility score of each search result according to the comprehension-based ranking model described in Section 2.

The overview of the re-ranking method is illustrated in Figure 4. For the initial ranking of search results ($t = 0$), we utilize the original ranking by a Web search engine as mentioned above. Therefore, at this step the ranking is independent of the comprehensibility of each Web page. When the user views a Web page d_1 and clicks the "difficult" button ($t = 1$), he/she requires Web pages that are easier than d_1. Then, the mean value within the range of $[f_c(d_1), 1]$ is regarded as the expected value. Web pages with the comprehensibility score closer to the expected value are ranked more highly as shown in Figure 4. Suppose that at the next step a Web page d_2 is viewed and the "easy" button is clicked ($t = 2$), which means that the user desires Web pages that are more difficult than d_2 and that are still easier than d_1. Then, the expected value of the user's comprehension level is calculated by considering the previous step, and the middle value of the range of $[f_c(d_1), f_c(d_2)]$ is regarded as the expected value. Note that our method for predicting the users' comprehension level seems to be a binary search but is different from it in that Web pages outside the selected range are not removed from the list of search results. This is because our proposed method for estimating the comprehensibility of Web pages cannot necessarily output the correct rankings and because the user's judgments may sometimes cause a contradiction. If Web pages outside the selected range were removed, sometimes all search results would disappear.

Below we formalize the re-ranking method by user's feedback. The operation in our system is regarded as the tuple of a Web page d and a user judgment J as follows:

$$O = (d, J), \quad J \in \{0, 1\} \tag{8}$$

Here, $(d, 0)$ is the operation when the user finds a Web page d too easy and requires more difficult Web pages whereas $(d, 1)$ is the operation when the user finds a Web page d difficult and desires easier Web pages. The expected value of user's comprehension level after n times operations , which is defined as u_n, is calculated by the following equations:

$$u_n = \frac{1}{n} \sum_{i=1}^{n} E(O_i) \tag{9}$$

$$E(O_i) = \frac{f_c(d_i) + J_i}{2} \tag{10}$$

$E(O_i)$ is a function for calculating the expected value derived of each operation. According to our proposed model, the gap between the comprehensibility score of Web page and the expected value of user's comprehension level after n times operations is calculated by the following equation:

$$\Delta_{d,u_n} = |f_c(d) - u_n| \tag{11}$$

After each operation, all search results are re-ranked in an ascending order of Δ_{d,u_n}.

5. EXPERIMENT

5.1 Query Set and Test Corpus

We performed a user-based evaluation to demonstrate the effectiveness of our method. In the experiment, we tested whether the comprehensibility scores calculated by our method correspond to users' evaluations. As far as we know, there is no test set for evaluating the comprehensibility assessment of Web pages. We manually prepared 20 technical term queries from five different domains as shown in Table 2. Then, for each query, we extracted 5 short passages containing the explanation about the query from Web pages acquired by a Web search engine[6]. In result, we formed a data set of the total number of 100 distinct passages for the 20 test queries.

5.2 Wikipedia Data

In the experiment, we have used the Japanese Wikipedia database downloaded in July 2008 using Wikipedia's downloading facility[7], which contains 511,165 articles and 58,690 categories.

When building an inverted index for *C-ESA*, we neglected some Wikipedia categories such as "*2008*" and "*articles with unsourced statements*" because they either do not represent any domain or they

[6]We used Yahoo! Web search API service.
http://developer.yahoo.co.jp/webapi/search/websearch/v1/websearch.html

[7]http://download.wikimedia.org

Table 2: Test Queries for user evaluation.

Domain	Query
Electromagnetism	Electromagnetic induction, Superconductivity, Operational amplifier, Light-emitting diode
Medicine	Parkinson's disease, Leprosy, Graves' disease, Spinal disc herniation
Biology	Mitochondrion, Chromosome, Golgi apparatus, Telomere
Economics	Subprime lending, Stock option, Capital gain, Derivative
Astronomy	Black hole, Supernova, Neutron star, Milky Way

Table 3: Precision (P), Recall (R) and F-measure (F) for the method of extracting technical terms. R' and F' respectively represent recall and F-measure after removing non-Wikipedia concepts from the answer set. K is the number of Wikipedia categories contained in the query domain.

	(P)	(R)	(R')	(F)	(F')
$K = 5$	**61.34**	54.38	74.92	56.78	66.59
$K = 20$	59.59	60.3	83.06	**59.08**	**68.69**
$K = 50$	56.66	**60.7**	**83.56**	57.92	66.92

contain very diverse concepts. In addition, we eliminated all the sub-categories of "*Category: Timelines*" and "*Cateogry: Fundamental*". Finally, 29786 categories were used in the experiment.

5.3 Performance of Technical Terms Extraction

We evaluated the technical terms extraction methods by comparing the technical terms automatically extracted with those manually selected. Table 3 shows the evaluation results for the parameter $K = 5, 20, 50$, which means the number of Wikipedia categories selected as the query domain. We measure the performance by using precision (P), recall (R) and F-measure (F), where precision is defined as the number of correctly extracted technical terms from a page divided by the total number of technical terms provided by our method; recall is defined as the number of correctly extracted technical terms divided by the total number of manually extracted technical terms from a page; and F-measure is the harmonic mean of the precision and recall. In our proposed method, the candidates of technical terms are limited to the concepts whose articles exist in Wikipedia. Hence, we also compute recall (R') and F-measure (F') by using the answer set limited to the Wikipedia concepts. Note that the precision does not change even if terms contained in the answer set are limited to the Wikipedia concepts because of its definition.

As shown in Table 3, the smaller K is, the higher the precision is and the lower the recall is. This result indicates that if the size of the query domain is small, we can obtain the small number of relevant Wikipedia categories but that coverage is not enough. The recall did not change too much even if the value of K was raised from 20 to 50, whereas the precision became lower then. Compared by F-measure, our method actually showed the best performance when $K = 20$. Finally, both the precision and the recall are about 60%, and the recall is about 80% if concepts that are not contained in Wikipedia are removed from the answer set. The following evaluations are performed by setting the parameter K to 20.

5.4 Performance of Document Comprehensibility Estimation

We performed a user-based evaluation to test whether our method of estimating the document comprehensibility corresponds to users' evaluations. Five users participated in this experiment. All of them are Japanese graduate and undergraduate students in informatics, and are non-expert users about any test queries. We showed evaluators 5 short passages containing the explanation of the test query, and then let them rank those passages in an order of ease-of-understanding. The rankings based on users' evaluations are defined as the mean reciprocal rank (MRR) of 5 distinct users' ranking. The MRR of a passage d_i is computed by the following equation:

$$MRR(d_i) = \frac{1}{n} \sum_{j=1}^{n} \frac{1}{R(d_i, u_j)} \tag{12}$$

where n is the number of evaluators ($n = 5$ in this experiment) and $R(d, u)$ means the rank of the passage d in the ranking by u. The rankings based on MRR were regarded as the correct answers of the comprehensibility score for each passage. We compared then the user evaluation with the comprehensibility level described as follows:

DRS Only the document readability is used to compute the comprehensibility score (Equation 3). We regarded this ranking method as the baseline.

DSS Only the reverse of the document speciality is used to compute the comprehensibility score (Equation 6).

DCS Both the document readability and the document speciality are used to compute the comprehensibility score (Equation 7), where α_1 and α_2 are respectively set to 0.5 and -0.5.

For each query, we then calculated Spearman's rank correlation coefficient between the above output rankings and the user ranking. Spearman's rank correlation coefficient ranges from -1 to 1 where -1 indicates that two rankings are completely reverse while 1 means that the rankings are completely the same.

The experimental results are shown in Table 4. We can see that our proposed measures, *DSS* or *DCS*, are superior to the baseline method *DRS* in every domain, and that on average the combination measure, i.e. *DCS*, shows the best performance. This results indicate that our proposed method using the domain knowledge derived from Wikipedia can be successfully applied to the comprehensibility-based ranking system independently of the query domain.

For some queries our proposed comprehensibility has a negative correlation with the users' evaluations. We consider that the limitation of Wikipedia knowledge is one factor that may cause lower performance. Although Wikipedia is the world's largest human knowledge base, all technical terms are not necessarily contained in Wikipedia. If a document contains many technical concepts that

Table 4: Comparison of the Spearman's rank correlation coefficients between user evaluation and our method in five different domains.

	DRS(baseline)	*DSS*	*DCS*
Electromagnetism	0.3436	**0.4911**	**0.4911**
Medicine	0.7472	0.6013	**0.8795**
Biology	0.3225	**0.375**	**0.375**
Economics	-0.1046	0.175	**0.225**
Astronomy	0.3699	**0.4417**	0.418
Avg.	0.3350	0.4168	**0.4777**

are not contained in Wikipedia, our method cannot fully estimate whether the document is of low comprehensibility or not.

From the results of this experiment, the following conclusions can be made:

- Document readability and document speciality have a positive correlation with users' evaluation.
- The performance can be improved by combining the document readability measure with document speciality measure.

5.5 Evaluation of User Interaction

We performed another user experiment to clarify the effectiveness of the re-ranking based on the user interaction in our system. In this experiment, we used only four medical queries that achieved the highest performance among five different domains in the above evaluation. Six evaluators participated in this experiment. We let evaluators search for comprehensible Web pages for each query by using our system. In order to examine the influences that user interactions bring, participants were divided into two groups; one group could utilize the re-ranking interface provided by our system and the other group could not (they could just only use the system without the interaction facility). We measured the length of the session times, the number of viewed pages and the frequency of interactions in each search session. Here, a search session is defined as the time period starting from the time when the user inputed a search query to the time when he/she found an appropriate Web page.

Table 5 shows the comparison of average length of session times and average number of viewed pages between the sessions that users could use the re-ranking interface and the sessions that they could not. It was revealed that both average length of session times and the average number of viewed pages increased when the re-ranking interface was available, which means that our re-ranking method did not work well from the viewpoint of efficiency. We consider that one of the causes is that the evaluators could not adjust themselves to the novel search interface during the experiment while they had already been familiar with the normal search engine without the re-ranking interface. In Table 6, we present some examples of search sessions in which our re-ranking method did not work effectively. In these examples, though finally selected pages were highly ranked and easy to find in the initial rankings made by search engine, evaluators first selected other pages and re-ranked search results based on the pages. Consequently, the length of session times, the number of viewed pages and the frequency of interaction all increased. On the other hand, in the search sessions by other users in which the same queries were issued, our system provided them with pages that were initially ranked low but that corresponded to users' comprehension level as shown in Table 7. The experimental results indicate that whether our system can provide users with appropriate search results strongly depends on users' judgment.

Table 5: Comparison of average length of session times and average number of viewed pages between sessions with user interaction and sessions without user interaction.

	Session time (sec)	**# of viewed pages**
without interaction	62.0	2.75
with interaction	117.3 (+189%)	4.1 (+149%)

6. DISCUSSION

In this section, we discuss the following topics: (i) other comprehensible metrics and (ii) other potential comprehension-based applications.

6.1 Other Comprehensible Metrics

In this paper, we have based our approach on the readability index and the domain knowledge extracted from Wikipedia. There are however other potential comprehensibility indicators in Web documents that could be utilized. For example, documents containing definitions of difficult concepts or their concrete, intuitive examples should be more understandable for users than abstract-level documents or the documents that lack any such definitions or examples. Detection of this kind of indicators is however not trivial. Potential solutions here could be based on applying fixed language patterns such as the ones that we have utilized for acquiring Web documents containing the explanations of a query keyword.

Document structure and content presentation are other important aspects that influence the levels of content comprehensibility. Web pages containing well structured and thematically organized content, clear section names, lists or other content arrangement techniques should be on average more readable than the ones without this kind of content presentation. In the future research we plan to extend our approach by incorporating techniques for the analysis of different content presentations.

We need to also emphasize that multimedia content plays an important role in increasing document understandability. Thus, a complete comprehensibility-focused Web page analysis should also investigate the explanatory role of multimedia elements in Web pages. However, such a complete analysis does not seem to be immediately feasible with the state-of-the-art technologies. Nevertheless, certain processing steps could be already applied here such as the ones based on OCR or on measuring the semantic connection between images and their surrounding textual content.

6.2 Other Potential Application

In this paper, we proposed a system for retrieving comprehensible Web documents including the explanation of the query keyword. There are however many other potential comprehension-based applications. One potential application could be a browser enhancement in the form of automatic search for definitions on technical terms in browsed Web pages. A somewhat similar idea has been proposed by Mihalcea et al. [17] who introduced a system for automatically linking Wikipedia pages to concepts extracted from browsed pages. However, Wikipedia articles are not necessarily comprehensible and we believe that our proposed approach should be more useful for non-expert users.

A comprehensibility-aware navigation support could become also useful for guiding Web surfers to the parts of Web sites that

Table 6: Examples of search session in which the re-ranking method did not work effectively.

Query	# of interactions	# of viewed pages	Session time (sec)	Title & initial rank of a finally selected page
Parkinson's disease	4	6	195.7	Tokyo Metropolitan Institute for Neuroscience: Parkinson's disease [5th]
Leprosy	4	7	271.1	What is Leprosy? [3rd]
Graves' disease	3	6	133.1	Graves' disease [5th]

Table 7: Examples of search session in which the re-ranking method worked effectively.

Query	# of interactions	# of viewed pages	Session time (sec)	Title & initial rank of a finally selected page
Leprosy	1	2	36.6	Correct understanding about leprosy [93rd]
Graves' disease	2	4	89.6	About graves' disease - KURE CLINIC [110th]
Spinal disc herniation	1	3	92.4	Spinal disc herniation and body distortion [65th]

contain relatively easy-to-understand content. Coiro and Dobler[8] noticed that "*...compared to print-based text, hypertexts often provide link labels with less semantic clarity and fewer surrounding context cues to guide the reader's anticipation about where a certain hyperlink may lead...*". The potential support could be made through annotating links on pages with estimated comprehensibility scores of their linked pages that could be automatically fetched by a browser during browsing. Similar-style interaction mechanisms, although not concerned with the issue of comprehensibility, were proposed earlier for query-focused [26] as well as freshness-focused navigation supports [11].

7. RELATED WORK

7.1 Web Search and Web Page Quality

The quality of Web pages has been evaluated so far from various viewpoints. Some theoretical studies regard comprehensiveness of Web pages as a key feature of their quality [31, 32].

Relevancy ranking method by vector space model is one of the classical search technologies and implies that relevant documents to a search query are of high-quality. Vector space model represents documents as vectors of index terms and is simply called bag-of-words model [23]. Relevancy between a search query and documents is calculated by the similarity between respectively their corresponding vectors. Latent semantic indexing [7] improves the vector space model by analyzing the implicit structure in the association of terms with documents.

Link analysis has been probably the most frequently exploited approach for the quality evaluation in information retrieval. PageRank [3] and HITS [14] are well-known algorithms in which the number of in-links of a Web page are used as a rough measure for the popularity and, indirectly, the quality of the page. The possibility of evaluating the quality of Web pages by using the information extracted from social bookmarking sites such as Del.icio.us[9] is described in [2].

Some researchers also proposed machine learning approaches for evaluating the quality of Web pages [1, 10, 16]. In these approaches, HTML structure, the number of links and language features such as the number of unique words and so on are used as parameters for machine learning. Mandl [16] proposed AQUAINT, a quality-based search engine, using a machine learning method. Zhou et al. [31] described a document quality language model approach for Web ad hoc retrieval. Our method is different from these works in that we explicitly focus on the comprehensiveness of Web pages. To the best of our knowledge, this is the first attempt to directly approach the problem of comprehensiveness of Web search results in a domain-independent fashion. We also use Wikipedia as a underlying knowledge base that is constructed by the collaborative effort of multiple users.

[8]http://ctell1.uconn.edu/coiro/CoiroDobler2006.doc

[9]http://del.icio.us

7.2 Relevance Feedback

Relevance feedback [22] is the most popular information retrieval system that uses the user feedback so as to improve the retrieval performance. There are three types of relevance feedback model: explicit feedback [21], implicit feedback [12] and pseudo feedback [4]. Explicit feedback is the most related to our work among them, and its basic cycle is as follows: the retrieval system shows users the initial set of results and then asks the user to judge whether some documents are relevant or not; after that, the system reformulates the query based on the user's judgments. Relevance feedback is not useful for information retrieval in which users require a few relevant documents such as our system because they will be satisfied if they find a relevant document on the relevance judgments. Also, the feedback about the comprehensibility that users have viewed is always negative in our system, which means that the operations are not executed when users find one Web page that is appropriate to their comprehension level. Non-relevance feedback that uses only non-relevant documents to find the target documents from a large data set of documents has been proposed in [20]. However, our work is different from these works in that we focus on the comprehensibility of Web pages, and in that we have to deal with two types of negative feedback, i.e. "easy" and "difficult".

8. CONCLUSIONS AND FUTURE WORK

In this paper, we have proposed the ranking model for comprehension-based Web search at first. For estimating the comprehensibility of Web pages themselves, we have proposed two indicators, document readability and document speciality that is calculated by using Wikipedia-based domain knowledge. The document speciality approach is based on analyzing the link and category structure of Wikipedia. Moreover, we have implemented a prototype system of comprehension-based Web search engine specialized in the term explanation. In our system, search results are adaptively re-ranked by presuming the users' comprehension level based on whether Web pages that they have viewed are easy or difficult for them.

We conducted experimental evaluations in order to clarify the effectiveness of our proposed methods. We prepared test queries derived from five distinct domains and performed user experiments for revealing the performance of estimating the comprehensibil-

ity of Web pages including the explanation about the query keyword. The experimental results suggest that the performance of our combination method is on average superior to baseline measures, which indicates that it can be successfully applied to the comprehensibility-based ranking system regardless of the query domain.

In future work, we are going to perform further evaluations with a range of queries from different domains or with varying difficulty levels. Also, we are going to improve the user interface of our prototype system and evaluate it from various aspects.

9. ACKNOWLEDGMENTS

This work was supported in part by the following projects and institutions: Grants-in-Aid for Scientific Research (No. 18049041) from MEXT of Japan and a Kyoto University GCOE Program entitled "Informatics Education and Research for Knowledge-Circulating Society,"

10. REFERENCES

[1] B. Amento, L. Terveen, and W. Hill. Does "authority" mean quality? predicting expert quality ratings of web documents. In *Proceedings of the 23rd SIGIR*, pages 296–303. ACM, 2000.

[2] S. Bao, G. Xue, X. Wu, Y. Yu, B. Fei, and Z. Su. Optimizing web search using social annotations. In *Proceedings of the 16th WWW*, pages 501–510, 2007.

[3] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30:107–117, 1998.

[4] C. Buckley, G. Salton, J. Allan, and A. Singhal. Automatic query expansion using smart: Trec 3. In *TREC*, pages 69–80, 1994.

[5] K. Collins-Thompson and J. P. Callan. A language modeling approach to predicting reading difficulty. In D. M. Susan Dumais and S. Roukos, editors, *HLT-NAACL 2004: Main Proceedings*, pages 193–200. Association for Computational Linguistics, 2004.

[6] E. Dale and J. Chall. *Readability Revisited: The New Dale-Chall Readability Formula*. Brookline Books/Lumen Editions, 1995.

[7] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41:391–407, 1990.

[8] A. Fujii and T. Ishikawa. Utilizing the world wide web as an encyclopedia: extracting term descriptions from semi-structured texts. In *Proceedings of the 38th ACL*, pages 488–495. Association for Computational Linguistics, 2000.

[9] E. Gabrilovich and S. Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of the 20th IJCAI*, pages 1606–1611, 2007.

[10] M. Y. Ivory and M. A. Hearst. Statistical profiles of highly-rated web sites. In *Proceedings of CHI 2002*, pages 367–374. ACM, 2002.

[11] A. Jatowt, Y. Kawai, and K. Tanaka. Personalized detection of fresh content and temporal annotation for improved page revisiting. In *Proceedings of the 17th DEXA*, pages 832–841, 2006.

[12] D. Kelly and J. Teevan. Implicit feedback for inferring user preference: a bibliography. *SIGIR Forum*, 37(2):18–28, 2003.

[13] G. Klare. The measurement of readability. *Journal of Business Communication*, 1(2):56–58, 1964.

[14] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of ACM*, 46(5):604–632, 1999.

[15] B. Liu, C. W. Chin, and H. T. Ng. Mining topic-specific concepts and definitions on the web. In *Proceedings of the 12th WWW*, pages 251–260. ACM, 2003.

[16] T. Mandl. Implementation and evaluation of a quality-based search engine. In *Proceedings of the 17th Conference on Hypertext and Hypermedia (HT 2006)*, pages 73–84. ACM, 2006.

[17] R. Mihalcea and A. Csomai. Wikify!: linking documents to encyclopedic knowledge. In *Proceedings of the 16th CIKM*, pages 233–242. ACM, 2007.

[18] S. Nakamura, S. Konishi, A. Jatowt, H. Ohshima, H. Kondo, T. Tezuka, S. Oyama, and K. Tanaka. Trustworthiness analysis of web search results. In *Proceedings of the 11th ECDL*, pages 38–49. Springer, 2007.

[19] M. Nakatani, A. Jatowt, and K. Tanaka. Easiest-first search: Towards comprehension-based web search. In *Proceedings of the 18th CIKM*, pages 2057–2060. ACM, 2009.

[20] T. Onoda, H. Murata, and S. Yamada. Non-relevance feedback document retrieval based on one class svm and svdd. In *Proceedings of the IJCNN*, pages 1212–1219, 2006.

[21] J. Rocchio. *Relevance Feedback in Information Retrieval*, pages 313–323. 1971.

[22] G. Salton. *The SMART Retrieval System—Experiments in Automatic Document Processing*. Prentice-Hall, Inc., 1971.

[23] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620, 1975.

[24] S. Sato, S. Matsuyoshi, and Y. Kondoh. Automatic assessment of japanese text readability based on a textbook corpus. In *Proceedings of the 6th LREC*. ELRA, 2008.

[25] S. E. Schwarm and M. Ostendorf. Reading level assessment using support vector machines and statistical language models. In *Proceedings of the 43rd ACL*, pages 523–530. Association for Computational Linguistics, 2005.

[26] S. P. Shashank. Navigation-aided retrieval. In *Proceedings of the 16th WWW*, pages 391–400, 2007.

[27] L. Si and J. Callan. A statistical model for scientific readability. In *Proceedings of the 10th CIKM*, pages 574–576. ACM, 2001.

[28] E. A. Smith and R. J. Senter. *Automated readability index*. AMRL-TR-66-22, 1967.

[29] J. Xu, Y. Cao, H. Li, and M. Zhao. Ranking definitions with supervised learning methods. In *Proceedings of the 14th WWW*, pages 811–819. ACM, 2005.

[30] X. Yan, D. Song, and X. Li. Concept-based document readability in domain specific information retrieval. In *Proceedings of the 15th CIKM*, pages 540–549. ACM, 2006.

[31] Y. Zhou and W. B. Croft. Document quality models for web ad hoc retrieval. In *Proceedings of the 14th CIKM*, pages 331–332. ACM, 2005.

[32] X. Zhu and S. Gauch. Incorporating quality metrics in centralized/distributed information retrieval on the world wide web. In *Proceedings of the 23rd SIGIR*, pages 288–295. ACM, 2000.