Easiest-First Search: Towards Comprehension-based Web Search

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

Although Web search engines have become information gateways to the Internet, for queries containing technical terms, search results often contain pages that are difficult to be understood by non-expert users. Therefore, re-ranking search results in a descending order of their comprehensibility should be effective for non-expert users. In our approach, the comprehensibility of Web pages is estimated considering both the document readability and the difficulty of technical terms in the domain of search queries. To extract technical terms, we exploit the domain knowledge extracted from Wikipedia. Our proposed method can be applied to general Web search engines as Wikipedia includes nearly every field of human knowledge. We demonstrate 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, comprehensibility, Wikipedia mining, readability

1. INTRODUCTION

Web search engines have become frequently used for acquiring information over the Internet. According to the survey conducted by Pew Internet & American Life Project it has been found that about 87% of online users have at one time used the Internet to carry out research on a scientific topic or concept[1]. In many cases, users require Web pages including comprehensible information about their search queries. Yet, conventional search results often contain pages that are difficult to be understood by non-expert users, especially for queries containing technical terms, for example, in the medical, financial and astronomy areas. In such cases, users must manually find out comprehensible Web pages from search results. Consider the following passages contained in search results acquired by issuing a 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. Thus, users may have difficult time finding comprehensible pages on the Web especially for queries containing technical terms.

In this paper, we introduce the concept of comprehension-based Web search. According to the Oxford Advanced Learner’s Dictionary, comprehension means “the ability to understand”. We define the comprehension-based Web search as the Web search that outputs search results considering user’s comprehension level about search topics. In particular, in this paper, we focus on non-expert users and aim at so-called “Easiest-First Search”, which re-ranks the results to find the most comprehensible ones upon user request. The problem is crucial because it is often difficult for non-expert users to modify their queries for acquiring more comprehensible Web pages, while, in contrast, expert users can add some keywords to their queries for obtaining detailed information.

In this work, we propose the method of re-ranking search results by combining two text comprehensibility measures. One is based on the readability index that is the traditional way to estimate how easily documents can be read. Readability index has the advantage of being easily calculable as it is computed using the surface features of texts such as the average number of syllables or the average length of sentences. However, the readability index alone is not sufficient for evaluating the comprehensibility of Web pages that contain technical terms as it is designed for general purpose texts and does not consider the word difficulty in a domain of a search query. The other approach is using technical terms related to a search query. Web pages containing a lot of technical terms are assumed to be little comprehensible because users are required to have prior knowledge related to a search query for understanding them. We define the rate of contained technical terms as *document speciality* [3]. In this approach, the difficulty of technical terms in the domain of search queries is considered.

[1] http://www.pewinternet.org/

CIKM’09, November 2–6,2009,Hong Kong,China.

Our approach is similar to the concept-based document readability in domain specific information retrieval using a technical thesaurus [5]. 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[2]. Also, some researchers have attempted to evaluate the quality of Web pages from various viewpoints [2, 6]. To the best of our knowledge, this is the first attempt to directly approach the problem of comprehensibility of Web search results in a domain-independent fashion.

2. METHODOLOGY

In this section, we describe the method of re-ranking search results by their comprehensibility scores based on readability index and technical terms related to a search query. Note that our proposed method focuses on Web pages written in Japanese. We also assume that search queries are not ambiguous and that Web pages in search results describe the same topics.

2.1 Document Readability

There are formula-based approaches and statistical language model based approaches for predicting document readability. 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 a document presentation style.

For Japanese texts, few readability measures have been proposed. In this work, we utilize *Obi* [4], 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 measurement method determines the school 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. 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:

$$DRS(d) = \frac{13 - Obi(d)}{12} \tag{1}$$

where $Obi(d)$ is an integer value acquired by inputting a document d into the *Obi* program.

2.2 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. We utilize the category and link structure of Wikipedia for extracting technical terms. 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 search result. 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 each step.

2.2.1 Building Semantic Interpreter

Here we provide the way to build semantic interpreter for detecting the domain of a search query. Gabrilovich et al. [1] 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[3]. Every Wikipedia category is then represented as a vector of terms, and entries of these vectors are weighted using the *TFIDF* scheme. We build an inverted index to speed up semantic interpretation.

In this paper, we have used the Japanese Wikipedia database downloaded in July 2008 using Wikipedia's downloading facility[4]. Note that we neglected some Wikipedia categories that do not represent any domain or contain very diverse concepts. Finally, 29786 categories were used for building a semantic interpreter.

2.2.2 Query Domain Detection

At this step, we detect a query domain by using the semantic interpreter. In our approach, a query domain is represented as a set of Wikipedia categories related to the issued query. First, we acquire the top K search results returned for a search query by a conventional search engine and download their content. Every Web page is then represented as a feature vector by using *TFIDF* weighting method and is mapped into a weighted vector of Wikipedia categories. In this work, the *IDF* weight of a term t is calculated by the following equation: $IDF(t) = \log \frac{N_C}{CF(t)}$, where N_C is the total number of Wikipedia categories and $CF(t)$ is the number of Wikipedia categories containing t. The weight of each Wikipedia category is measured by using cosine similarity between *TFIDF* vectors of a Web page and itself. In result, we obtain a list of weighted vectors $\{\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_K\}$. Finally, we calculate a centroid vector $\bar{\mathbf{v}} = \sum_i \mathbf{v}_i / K$. In this paper, we do not discuss how to determine the threshold and simply regard top 10 Wikipedia categories in a centroid vector as the query domain.

2.2.3 Technical Terms Extraction

After we have detected a query domain we proceed to extract technical terms from search results. We base our method on the the following hypothesis. Wikipedia concepts that are rarely linked from Wikipedia concepts outside the domain of search query are assumed to be technical terms. Below we describe the detailed algorithm of technical terms detection. First, the candidates of technical terms are extracted from the documents of search results. Here each candidate term is represented as one Wikipedia concept, that is the title of a Wikipedia article. Next, we determine which ones are actually real technical terms of the query domain. 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. D_q denotes Wikipedia articles included in domain categories, and $D_{\overline{q}}$ denotes the remaining Wikipedia articles. The degree of bias of the link occurrence distribution between D_q and $D_{\overline{q}}$ is measured by

[2] http://en.wikipedia.org/

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

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

the χ^2-measure. However, note that, when the links to a Wikipedia concept appear rarely in D_q but mostly in $D_{\overline{q}}$, the χ^2-measure can also have a high value. Therefore, we remove candidate terms that do not satisfy the following condition: $\frac{LF(t,D_q)}{|D_q|} > \frac{LF(t,D_{\overline{q}})}{|D_{\overline{q}}|}$, where $LF(t, D)$ denotes the number of articles that link to the article of a candidate term t and that are contained in the article set D. In this way we received a set of actual technical terms related to the domain of the query.

2.2.4 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:

$$DSS(d, q) = \exp\left(-\frac{1}{\log |d|} \sum_{t \in TT(d,q)} \frac{1}{LF(t, D_q)}\right) \quad (2)$$

where $TT(d, q)$ 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. $LF(t, D_q)$ is the same one as defined in the above section.

2.3 Re-ranking Method

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 with a varying parameter α.

$$DCS(d, q) = (1-\alpha) \cdot DRS(d) + \alpha \cdot DSS(d, q), \quad \alpha \in [0, 1] \quad (3)$$

We regard Web pages with higher DCS value as more comprehensible. In Equation 3, both the query independent factor and the query dependent factor of comprehensibility are taken into consideration. The simplest approach for easiest-first Web search is to re-rank search results by their DCS values.

3. EXPERIMENT

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. Moreover, the comprehensibility of Web pages strongly depends on users' prior knowledge about the domain of the search queries. In order to check the correlation between our proposed measures and the comprehensibility of Web pages, we prepared some technical term queries and a test corpus automatically extracted from Web pages. Below we describe the details of the experiment.

3.1 Query Set and Test Corpus

Table 1: Test Queries for user evaluation.

Domain	Query
Astronomy	Black hole, Dark matter, Pluto, Halley's comet
Medicine	Parkinson's disease, Atherosclerosis, Aortic dissection, Meniere's disease
Physics	Theory of relativity, Superstring theory, Superconductivity, Lorentz force
Biology	RNA, Chromosome, Kinase, Mitochondrion
Economics	Subprime lending, Marginal utility, Stagflation, Diminishing returns

The advantage of our proposed method is that it can be applied regardless of the domain of a search query. Therefore, we manually prepared 20 search queries from different domains as shown in Table 1. Our query set is composed of terms from five different domains (astronomy, medicine, physics, biology and economics) that the evaluators are not familiar with.

Then, for each query, we extracted short passages from each of 10 Web pages acquired by a Web search engine. The details of a passage extraction process for each query are described as follows:

1. Acquire top search results by using Yahoo! Web search API service.

2. Download Web pages from the search results.

3. From each page, extract a head paragraph of 200-400 characters that contains the query term. In this work, we regarded a text surrounded by a <*p*> tag as a paragraph.

4. Repeat the step 3 until the top 10 pages are processed.

Thus, we formed a data set of the total number of 200 distinct passages using Web search results returned for the 20 test queries.

3.2 Evaluation Metrics

Eight users participated in our experiment. All of them are Japanese graduate and undergraduate students in informatics, and are non-expert users about any test queries. The difficulty of each page was rated on a scale of 1 (difficult) to 5 (easy) by the evaluators. In this experiment, we regarded the average of scores given by evaluators as the correct answers of the comprehensibility score for each page. 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 1).

- ***DSS***: Only the document speciality is used to compute the comprehensibility score (Equation 2).

- ***DCS***: Both the document readability and the document speciality are used to compute the comprehensibility score (Equation 3). A parameter α can be varied from 0 to 1.

For each query, we then calculated Spearman's rank correlation coefficient between the above output rankings and the user ranking.

3.3 Experimental Results

Figure 1 shows the Spearman's rank correlation coefficients between the user ranking and the system ranking that were calculated according to the parameter α with 0.1 increments. The horizontal axis represents the value of parameter α, while the vertical axis shows the rank correlation between two rankings computed according to the value of α. For each parameter value and for each query domain, we computed the average of rank correlation coefficients

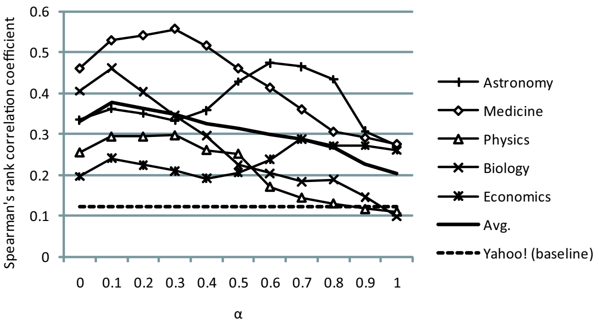

Figure 1: Rank correlations between user evaluation and our proposed method and the original ranking. α is a parameter for computing *DCS* values.

Table 2: Comparison of the Spearman's rank correlation coefficients between user evaluation and our methods in five different domains.

Domain	Yahoo!	DRS	DCS
	(baseline1)	(baseline2)	($\alpha = 0.1$)
Astronomy	0.1279	0.3340	**0.3618**
Medicine	0.0706	0.4596	**0.5282**
Physics	0.2661	0.2551	**0.2938**
Biology	-0.14	0.405	**0.4606**
Economics	**0.2818**	0.1961	0.2407
Avg.	0.1213	0.3300	**0.3770**

acquired by 20 queries. Moreover, we calculated the rank correlation between the user ranking and the original ranking of Yahoo! Web search and regarded it as a simple baseline. As shown in Figure 1, the document readability, the document speciality and the combination method have a positive correlation with users' evaluations and exceed on average the baseline.

In Table 2, we report the comparison of performance of our methods for five different query domains. When the document readability measure, i.e. *DRS*, is considered as another baseline method, we can see that our proposed combination measure, i.e. *DCS*, is superior to the baseline in every domain. This result indicates that our proposed method using the domain knowledge derived from Wikipedia can be successfully applied to the comprehensibility-based ranking system for general search engines using queries from various domains.

However, for some queries our proposed comprehensibility have a negative correlation with users' evaluations. We consider that the limitation of Wikipedia knowledge may cause the lower performance. Although Wikipedia is the world's largest human knowledge base, all technical terms are not necessarily contained in Wikipedia. If a Web page contains many technical concepts that are not contained in Wikipedia, our method cannot fully estimate whether the page is of low-comprehensibility or not.

From the results of our experiments, the following conclusions can be made:

- Document readability and document speciality have a positive correlation with the users' evaluation.

- The performance can be improved by combining the document readability measure with the document speciality measure regardless of query domains.

4. CONCLUSIONS AND FUTURE WORK

In this paper, we have introduced the concept of comprehension based Web search and proposed the method of estimating the comprehensibility of Web pages by combining the document readability and the document speciality that is calculated by using Wikipedia-based domain knowledge. We prepared test queries derived from five distinct domains and performed user experiments in order to demonstrate the effectiveness of our proposed method. The experimental results suggest that the performance of our combination measure is on average superior to two baseline measures, which indicates that it can be successfully applied to the comprehensibility-based ranking system for general search engines using queries from various domains.

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 implement a prototype system of re-ranking search results based on their comprehensibility scores.

In this paper, we focused on non-expert users and proposed the concept of "easiest-first". In future work, we plan to work on other advanced interaction models such as "level-up" and "related-but-easier".

- *Level-up* is an interaction model that provides users with more difficult materials for their next learning steps.

- *Related-but-easier* is an interaction model that presents users with easier materials that are at the same time related to the currently viewed document. This model will be able to enhance the Google's "similar" pages search.

We believe that these interaction models are a new feedback paradigm on Web search. Also, the comprehension-based approach can be applied in several other areas such as question answering or document summarization.

5. 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, a Kyoto University GCOE Program entitled "Informatics Education and Research for Knowledge-Circulating Society," and the National Institute of Information and Communications Technology, Japan.

6. REFERENCES

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

[2] T. Mandl. Implementation and evaluation of a quality-based search engine. In *17th HYPERTEXT*, pages 73–84. ACM, 2006.

[3] M. Nakatani, A. Jatowt, H. Ohshima, and K. Tanaka. Quality evaluation of search results by typicality and speciality of terms extracted from wikipedia. In *14th DASFAA*, pages 570–584. Springer-Verlag, 2009.

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

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

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