Measuring Comprehensibility of Web Pages Based on Link Analysis

Kouichi Akamatsu, Nimit Pattanasri, Adam Jatowt, Katsumi Tanaka
Department of Social Informatics
Graduate School of Informatics, Kyoto University
Kyoto, Japan
{akamatsu, nimit, adam, tanaka}@dl.kuis.kyoto-u.ac.jp

Abstract— **We put forward a hypothesis that if there is a link from one page to another, it is likely that comprehensibility of the two pages is similar. To investigate whether this hypothesis is true or not, we conduct experiments using existing readability measures. We investigate the relationship between links and readability of text extracted from web pages for two datasets, set of English and Japanese pages. We could find that links and readability of text extracted from web pages are correlated. Based on the hypothesis, we propose a link analysis algorithm to measure comprehensibility of web pages. Our method is based on the TrustRank algorithm which is originally used for combating web spam. We use link structure to propagate readability scores from source pages selected based on their comprehensibility. The results of experimental evaluation demonstrate that our method could improve estimation of comprehensibility of pages.**

Keywords-comprehensibility; link analysis; readability

I. INTRODUCTION

With the development of the web people obtain information on topics about which they do not know much. Web search engines are useful for users to collect information efficiently through the web. To decide which web pages to return to users, search engines consider several factors such as relevance to a query and the importance of pages computed based on link analysis. However, in general, the comprehensibility of web pages does not seem to be considered in ranking search results. Therefore, especially for difficult topics, the top returned pages may not be easy-to-understand for general users. If search engines can return pages whose comprehensibility is suitable for users' request, web search will be more effective. To do this, we need a method to automatically measure comprehensibility of pages on the web.

The widely accepted definition of readability can be found in [2]: "The sum total (including all the interactions) of all those elements within a given piece of printed material that affect the success a group of readers have with it. The success is the extent to which they understand it, read it at an optimal speed, and find it interesting." This definition is general enough to encompass several factors of readability such as writing styles and ease of vocabulary. In addition, the definition also blurs the distinction between readability and comprehension. For simplicity, in this paper, we use the term readability and comprehension interchangeably.

Several methods to measure readability of text have been proposed so far and it might be possible to use one of them. Such methods usually consider various features of textual content such as sentence length, word length or average number of syllables. Although, they have been successfully used for many document genres such as legal documents and school textbooks, they are not effective enough to be directly applied for web pages. This is because web pages consist of not only text but also of tables, images, videos, sound and of other elements that affect their comprehensibility. In addition, it is possible to make the content of pages more understandable for users just by improving design or layout. Therefore, to measure comprehensibility of web pages, analyzing only extracted text is insufficient. However, on the other hand, there are no ready methods to be used for exhaustive analysis of the whole content of web pages including images, videos, layout and so on from the comprehensibility viewpoint.

In this research, we try to establish a method to measure comprehensibility of web pages not by analyzing content but by analyzing link structure. With this approach, we aim to correctly measure the part of comprehensibility of pages which cannot be measured only by estimating text-based readability. We make two contributions in this paper. First, we show the results of the analysis of readability of web pages to show the correlation between link structure and readability on the web. We succeed in confirming that linkage on the web is positively correlated with page readability levels. Second, we propose a novel link-based method to evaluate page comprehensibility levels. The method is similar to TrustRank algorithm [5] that has been originally used for web spam elimination. We verify the effectiveness of our approach by experiments on a real world dataset.

The remainder of the paper is structured as follows. We discuss the related work in Section II. We explain how datasets were created by using a search engine in Section III. We show the relationship, through some statistics, between comprehensibility of web pages and link structure in Section IV. We describe our method that propagates readability scores on the web in Section V. In Section VI we demonstrate our method to measure comprehensibility of web pages and evaluate its effectiveness. Section VII contains general discussion of comprehension on the web. Lastly, we make our conclusions in Section VIII.

II. Related Work

Measuring readability of texts is subjective and depends on background knowledge of readers. The gap between a text and a reader measured, for example, in school grade levels, often determines whether the text is readable or not. This interpretation, therefore, casts the original problem to simply classifying texts according to reading levels. This line of research enjoys the success of text classification which exploits various features ranging from surface text features (e.g., word length) to discourse-level features (e.g., the number of entities involved in a text) and from a manually compiled list of vocabulary to statistical language modeling. Broadly speaking, these features address the problem at different levels of granularity: grammatical complexity, vocabulary complexity, and story complexity.

Average sentence length is often a good indicator for grammatical difficulty. Flesch Reading Ease [4], one of the earliest standard measures, defines readability as a function of word length and sentence length. The approach is simple to implement but sensitive to outliers especially in web pages which offer rich presentation styles. To work well, this measure requires a strict rule of punctuation to form complete sentences. However, a long list of phrases or incomplete sentences would form an unexpectedly long sentence, finally leading to an incorrect low readability score [15, 8]. Predicting readability of text summaries returned by search engines also renders this approach impractical [7]. In the extreme case, most snippets contain only fragments of sentences which cause similar measurement error. Deeper syntactic features that rely on the assumption of complete sentences such as pronoun count per sentence also suffer the same problem.

A second feature category focuses on difficulty levels of words themselves while ignoring grammatical difficulty to achieve a more robust measure. Word length or syllable count is a good approximation for word complexity [4]. However, this simple heuristic needs manual parameter tuning. A more intuitive approach is to use a predefined list of common or easy words to identify unfamiliar words [1]. However, due to a dynamic characteristic of language, the static list requires an update once new common words become available. Using Wikipedia could be helpful to deal with the dynamic change of language. Nakatani et al. [10, 11] extracts technical or difficult terms related to user queries from Wikipedia and uses them for measuring difficulty levels of web search results. The results are then re-ranked to provide “easiest-first” search. Statistical language modeling offers a more systematic approach to building a dynamic vocabulary list. In either case, these lexical features were claimed to yield a better result than using syntactic features alone [14].

The relationship between readability and cognitive load has been investigated for a special group of readers who have limited working memory [3]. The key idea is that story complexity grows with the number of entities mentioned in the text. Effectiveness of this approach relies on named entity recognition software that plays an essential role for automatic feature extraction. Unlike this approach which is reader-dependent, our study focuses on general readers.

To combine the strength in each feature type, most researches, unsurprisingly, employ as many features as they deem relevant (e.g., [3, 7, 12, 14]). Our study differs from the previous work in that we address this text classification problem not by those content-based features but by surrounding context of web pages.

Closely related work is found in [8] which studies distribution of readability levels of all local web pages within a web site. Their study focuses on estimating site-level readability and improving website usability while our work proposes to exploit hyperlink structure of the whole web to improve readability measurement.

III. Datasets

In this section we describe the way in which we have created datasets for the comprehensibility analysis. Since the web structure indexed by search engines is not accessible to us, we seek an alternative approach. Issuing several queries to collect relevant pages from search engines is a common way to populate a digital library [6]. We exploit a similar approach to create a web graph in a topic-specific manner; each query generates a graph of topically related web pages illustrated in Fig. 1.

We carefully selected queries in an attempt to obtain web pages that contain expository texts, most likely a good genre to be comprehended by most readers.

Equally divided for two target languages: English and Japanese, 20 queries were issued to Yahoo! search API[1] (Table I and Table II). For each query, we collected the top 30 and top 100 search results for English and Japanese queries, respectively. Next, we collected up to 50 inlink pages and 50 outlink pages for each of the search result pages issued by English queries. For Japanese, we took up to 30 inlink pages and 30 outlink pages for each of the search result page. In total there were 23,612 and 54,418 pages in English and Japanese datasets, respectively. For each query, its top search result, inlink pages and outlink pages formed a web graph of topically related pages. We also downloaded page contents, making it possible to apply standard readability measures on those pages.

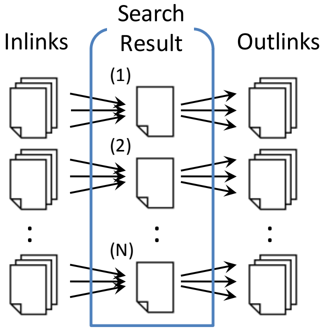

Figure 1. Web pages that constitute a dataset by issuing a query to a search engine. Note that the graph obtained in this way is usually well connected. Some actual links are omitted here for the sake of simplicity.

[1] http://developer.yahoo.com/search/boss/

TABLE I. TOPICS AS QUERIES TARGETED FOR ENGLISH WEB PAGES.

alzheimer’s disease, parkinson’s disease, bipolar transistor, quantum computer, comparative advantage, derivative, complex number, mitochondrion, black hole, halley’s comet

TABLE II. TOPICS AS QUERIES TARGETED FOR JAPANESE WEB PAGES (TRANSLATED FROM JAPANESE).

complex plane, fatty acid, alzheimer’s disease, muscular dystrophy, pagerank, black hole, doppler effect, stock option, synapse, gaia hypothesis

IV. RELATION BETWEEN LINKS AND READABILITY SCORES ON THE WEB

We first introduce two standard readability measures, each applied in a target language, and then present statistical results of readability scores distribution with relation to linking pages.

A. Selected Readability Measures

To measure readability of English pages we use the well-known *Flesch Reading Ease* which is defined by:

$$206.876 - 1.015ASL - 84.6ASW. \qquad (1)$$

where ASL is the average number of syllables per word and ASW is the average number of words per sentence. Flesch Reading Ease scores can be interpreted according to reading school levels shown in Table III.

For Japanese, we use *Obi* [13] which basically follows a vocabulary-based approach. A textbook corpus is used as training data for building character-based bigram models. Each model corresponds to one of the 13 school grade levels: 1–6 for elementary school (6 years), 7–9 for junior high school (3 years), 10–12 for high school (3 years), and 13 for above high school. Obi returns a reading level for a text which is most likely to be generated from the language model counterpart.

To get a general idea of how effective this readability measure is, we show the scores distribution over two datasets, generated by two different queries varied in their difficulty (Fig. 2). As expected, the proportion of difficult pages generated by a relatively more difficult query “complex plane” outnumbers the corresponding group of an easier query “Pokémon”. Similarly, the proportion of easy pages of “Pokémon” outnumbers that of “complex plane”.

B. Statistical Results

Our research objective is to develop a method to measure comprehensibility of web pages based on link analysis. To do so, we first need to verify how the comprehensibility of web pages and links are related.

TABLE III. INTERPRETATION OF FLESCH READING EASE SCORE[2].

Score	Reading Level
90.0-100.0	easily understandable by an average 11-year-old student
60.0-70.0	easily understandable by 13- to 15-year-old students
0.0-30.0	best understood by university graduates

Our hypothesis is that:

- There are many links from easy pages to easy ones.
- There are few links from easy pages to difficult ones.
- There are few links from difficult pages to easy ones.
- There are many links from difficult pages to difficult ones.

If this hypothesis holds, we could estimate the comprehensibility of a page based on the comprehensibility of inlinks of the page or, in other words, based on the page context. If a page is linked to by many easy pages, the page should have high probability to be easy too. On the other hand, if a page is linked to by many difficult pages, it should have high probability to be difficult.

In this section, we validate our hypothesis on English and Japanese datasets separately. At a glance, the results (Table IV and Table V) follow a trend that pages link to others with generally same level of difficulty.

It is shown in Table IV that if there is a link from a page A to another page B, the readability scores of A and B are correlated to a certain degree. If there is a link from an easy page A to a page B, the chance is that B is most likely to be easy as well (i.e., 53.2%). Similarly, if there is a link from a medium-difficulty page A to a page B, B is most likely to be of medium-difficulty (i.e., 56.8%). However, if there is a link from a difficult page A to a page B, B is most likely to be of medium-difficulty (47.8%), although the chance of it being difficult is also high (33.4%). Similar interpretation can be obtained for Japanese datasets from Table V. However, we get a relatively worse result for the easy-to-easy case but a significantly better result for the difficult-to-difficult case.

Of course, readability of web pages depends not only on texts but also on other multimedia contents embedded on those pages (e.g., images and videos), and their layout and design. However, most search engines still rely on text indexing for web search, indicating that texts represent the main content in most web pages. We speculate that the results we present in this section still hold in general even if we exclude non-textual factors from analysis.

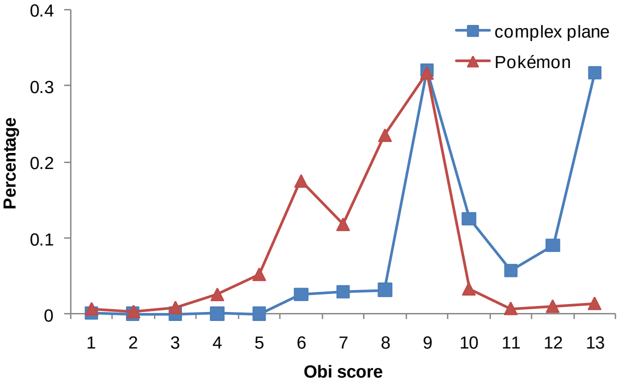

Figure 2. Distribution of readability scores, measured by Obi, over two datasets. The proportion of easy and difficult pages (low and high scores) compared between both datasets intuitively meets our expectation.

[2] http://en.wikipedia.org/wiki/Flesch-Kincaid_Readability_Test

TABLE IV. THE PERCENTAGE OF LINKS CATEGORIZED BY FLESCH READ EASE SCORES OF SOURCE AND DESTINATION PAGES: 60-100 AS EASY, 30-60 AS MEDIUM, AND 0-30 AS DIFFICULT. ALL LINKS ARE OBTAINED FROM DATASETS GENERATED ACROSS 10 ENGLISH QUERIES. THE NUMBER IN PARENTHESIS SHOWS THE ACTUAL NUMBER OF LINKS.

		Readability of Destination Pages (Linked)		
		Easy	***Medium***	***Difficult***
Readability of Source Pages (Linking)	***Easy***	53.2% (3,594)	42.1% (2,840)	4.68% (316)
	Medium	28.8% (3,233)	56.8% (6,381)	14.5% (1,630)
	Difficult	18.8% (355)	47.8% (903)	33.4% (631)

TABLE V. THE PERCENTAGE OF LINKS CATEGORIZED BY OBI SCORES OF SOURCE AND DESTINATION PAGES: 1-6 AS EASY, 7-9 AS MEDIUM, AND 10-13 AS DIFFICULT. ALL LINKS ARE OBTAINED FROM DATASETS GENERATED ACROSS 10 JAPANESE QUERIES. THE NUMBER IN PARENTHESIS SHOWS THE ACTUAL NUMBER OF LINKS.

		Readability of Destination Pages (Linked)		
		Easy	***Medium***	***Difficult***
Readability of Source Pages (Linking)	***Easy***	43.0% (15,601)	49.9% (18,117)	7.13% (2,590)
	Medium	3.37% (33,936)	84.2% (846,576)	12.4% (125,176)
	Difficult	1.46% (6,605)	34.6% (155,869)	64.0% (288,468)

V. PROPAGATING READABILITY SCORES ON THE WEB

Based on our hypothesis, we propose a method to measure comprehensibility of web pages. Our method is similar to TrustRank algorithm [5]. First we explain TrustRank in Section V-A and then we discuss our method in Section V-B.

A. *TrustRank*

TrustRank is a link analysis algorithm used to separate good pages from spam pages. It has been adopted from well-known PageRank algorithm [9]. TrustRank is based on the empirical observation that good pages seldom point to spam pages. According to this observation, a page linked to by a good page is likely to be good. In TrustRank, some non-spam pages have to be manually selected as seeds. By propagating scores from pages which are identified as good ones using biased PageRank algorithm, scores which represent the likelihood that pages are good are then calculated for pages in the entire set of pages. TrustRank is a semi-automatic algorithm which only needs human judgment for selecting seed pages.

Vector $\boldsymbol{r}$ whose entries are biased PageRank scores of pages is defined as:

$$\boldsymbol{r} = \alpha \cdot \boldsymbol{T} \cdot \boldsymbol{r} + (1 - \alpha) \cdot \boldsymbol{d}. \qquad (2)$$

α is a damping factor. $\boldsymbol{T}$ is the transition matrix. $\boldsymbol{d}$ is a vector of non-negative entries summing up to one. In the regular PageRank, all entries of vector $\boldsymbol{d}$ have the same values *1/N* if the total number of pages is *N*.

By assigning positive values for entries corresponding to good pages and zero for other entries of vector $\boldsymbol{d}$, TrustRank propagates scores from good pages.

B. *EasyRank and DiffRank*

TrustRank is based on the empirical observation that good pages seldom link to spam pages. However, the converse is not assumed according to the empirical observations. Table VI compares TrustRank's hypothesis with our hypothesis.

TrustRank propagates scores from good pages through links. According to the relationship shown in Table VI, scores propagated from good pages will be propagated to good pages with high probability. However, propagating scores from spam pages is not helpful for the purpose of separating good pages from spam pages. This is because scores propagated from spam pages would be transferred both to good and spam pages with high probability. We thus need to take different approach than the one in TrustRank in the case of measuring page comprehensibility.

We have found in Section IV that there are many links from easy pages to easy pages and there are few links from easy pages to difficult pages. If we propagate scores from easy pages by using biased PageRank, scores will tend to go to easy pages. On the other hand, there are many links from difficult pages to difficult pages and there are few links from difficult pages to easy pages. If we propagate scores from difficult pages, scores will tend to go to difficult pages. By propagating scores from either easy or difficult pages instead of only from good pages, we calculate scores which represent easiness or difficulty. We call our methods that propagate readability scores from easy and from difficult source pages EasyRank and DiffRank, respectively.

VI. EXPERIMENTS

A. *Test Collections*

Due to limited resources, we use only Japanese datasets (see Section III) for evaluating effectiveness of our proposed method. As each query in Table II generates a topic-specific web graph, a total of 10 test collections are subject to evaluation. Table VII shows some statistics of our test collections.

Ground truths for the test collections were constructed by a human judge and a standard readability measure separately. While human subjects provide direct assessment on readability of web pages, the process is time consuming, preventing large-scale evaluation. Despite ignoring non-textual contents which affect readability, a standard readability measure has an advantage over human subjects to produce indirect assessment "at scale".

Because it is hard to obtain ground truths manually for all pages, we avoid the bias in page selection by using a search engine. Top-twenty search results from Yahoo! constitute a pool for making manual judgment in each dataset (which corresponds to a query). These top-*k* results of a pool are then reranked by each method (see Section VI-B) for evaluating performance by using the precision metric. Each page was judged carefully based on both textual and non-textual contents. A potential bias toward our method was avoided because the link structure was not available at the time of making judgment.

TABLE VI. COMPARISON BETWEEN TRUSTRANK'S HYPOTHESIS AND OURS.

Link Type	Proportion
good→good	many
good→bad	few
bad→good	many
bad→bad	many

Link Type	Proportion
easy→easy	many
easy→difficult	few
difficult→easy	few
difficult→difficult	many

In addition, we use Obi as an external judge to obtain "approximate" ground truths for the whole datasets. Of course, constructing ground truths in this way prevents us to make performance comparison against Obi, but it allows us to understand readability on a large scale and from a different perspective. In particular, we can compare effectiveness of our method against the one that considers page popularity which is a good indicator for page quality.

B. Methods and Experimental Setup

The goal of our proposed method is to find easy-to-understand pages by link analysis. We compare effectiveness among these methods:

Yahoo!. As each test collection corresponds to a query, this method simply ranks web pages according to relevance to the query (i.e., we simply use top-k search results from Yahoo! API). This baseline reflects the current situation in web search.

PageRank. Page popularity has been one of the major factors for measuring quality of web pages. PageRank, a query-independent ranking algorithm, computes popularity scores based on the following formula:

$$\boldsymbol{r} = \alpha\cdot\boldsymbol{T}\cdot\boldsymbol{r} + (1-\alpha)\cdot\frac{1}{N}\cdot\mathbf{1}_N. \qquad (3)$$

PageRank calculates page popularity scores based on link structure. We believe that it is unlikely that many people support difficult pages. Documents that are hard to be understood should not acquire many links as they are not useful for majority of web users. Therefore, intuitively, difficult pages should not be popular. On the other hand, popular pages might be relatively easy. Of course this is only an assumption that should be tested by extensive experiments. We assume in this paper that popularity calculated by PageRank is related to page comprehensibility in this way. In our experiment, we set a damping factor to 0.85.

Obi. This method relies on a bigram language model to classify Japanese texts according to reading levels. Before applying the Obi measure, we remove HTML tags in all pages, and, ignore non-textual contents from analysis. We use Obi software which is available on the web[3].

TABLE VII. THE NUMBER OF PAGES AND LINKS IN EACH TEST COLLECTION.

Test collection for corresponding query	Num. of pages	Num. of links
complex plane	2,079	30,346
fatty acid	3,146	36,106
alzheimer's disease	16,925	593,777
muscular dystrophy	2,519	31,015
pagerank	11,262	344,483
black hole	3,125	33,961
doppler effect	2,147	30,892
stock option	7,354	394,108
synapse	3,235	38,167
gaia hypothesis	2,626	46,225

EasyRank. Our method uses TrustRank algorithm, originally applied for suppressing spam pages, to propagate readability scores on a web graph. We perform seed selection semi-automatically by exploiting a standard readability measure. That is, all pages whose Obi score is between 1 and 6 are selected as seed pages in each dataset. The results are ranked in a descending order of readability.

DiffRank. This method is similar to EasyRank, except that the seed set consists of pages whose Obi score is 13 (low readability). The returned results are ranked in an ascending order of page difficulty.

DiffRank+Obi. We also try a linear combination between our method and Obi to see whether it helps improve the performance further. Because DiffRank significantly performs better than EasyRank, we choose DiffRank. A combined approach is calculated by the formula:

$$\beta * \mathit{DiffRank}' + (1-\beta) * \mathit{Obi}'. \qquad (4)$$

DiffRank' and Obi' are DiffRank and Obi scores normalized by their maximum values, respectively. As for the value of β, we chose the one with the best performance from three values: 0.25, 0.5 and 0.75. The β parameter is set to 0.25 in our experiment.

C. Results

The results in Table VIII give a general idea of what happens at the top of ranking. Table VIII shows that compared to DiffRank, EasyRank is not so much good to measure comprehensibility. This phenomenon could be explained when we look at the results of statistical experiment on Japanese datasets that was described before (Table V). For Japanese datasets difficult pages tend to refer to difficult pages relatively more frequently that easy pages link to easy pages. This may be due to the fact that pages with an outline or easy introduction of a concept may link to pages with more detailed or elaborated information. Since there are few difficult pages that link to easy pages there is comparatively little "leakage" to easy pages when propagating scores from difficult pages in DiffRank. The propagation is however less efficient in the case of EasyRank due to the relatively higher percentage of links from easy to difficult pages.

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

The overall performance comparison in Table VIII shows that exploiting the link structure together with a standard readability measure improves readability measurement. DiffRank outperforms significantly at the top rank while Obi performs better than others at lower ranks. This explains why a linear combination between both methods yields a better result than each method in isolation. In Table IX we also show performance comparison of linear combination of DiffRank and Obi for different values of *β*.

Note that the top-ranked precision of Yahoo! is unexpectedly low in Table VIII. Closer inspection reveals that the top-ranked search results of all queries except one are Wikipedia pages which were manually judged to be difficult, but are the most relevant from a search engine viewpoint. In addition, Fig. 3 and Fig. 4 show that the effectiveness of our method depends on topics and sometimes is better than Obi for certain topics.

Large-scale evaluation is explored in light of Obi scores as approximated ground truths of readability (Fig. 5). PageRank could be considered somewhat as a weak indicator of readability as its descending ordering performs slightly better than its ascending ordering. There is however a huge gap between DiffRank and PageRank (both descending and ascending ordering) indicating that page popularity measured by link structure is not very effective for estimating readability.

VII. Discussion

Experiments. The main weakness of our studies is the limited experimental evaluation. The evaluation should be done on a larger number of queries and pages with multiple human judgments. We plan to perform such extensive evaluation in our future work.

Page Relevance. We have focused on a single metric for improving effectiveness and satisfaction of web search. Obviously, relevance and importance or popularity of pages are other major metrics. Search engine designed to return easy-to-understand search results for difficult user queries should naturally output only relevant results. Thus a comprehensive search approach should combine the measures of relevance and comprehensibility. One could imagine a "perfect" search engine from the comprehensibility viewpoint that would always output simple yet unrelated pages. Obviously such a service would be completely useless for users.

TABLE VIII. The overall effectiveness comparison by precision at top-*k* across 10 test collections. Ground truths were obtained from a human judge.

k	Yahoo!	Obi	EasyRank	DiffRank	DiffRank+Obi (β=0.25)
1	0.00	0.56	0.22	**0.67**	0.56
3	0.22	0.56	0.37	0.48	**0.63**
5	0.29	0.47	0.36	0.47	**0.53**
10	0.31	**0.44**	0.31	0.40	0.43
15	0.30	**0.37**	0.32	0.35	0.36

TABLE IX. The comparison of the DiffRank+Obi performance over different *β*.

k	β=0	β=0.25	β=0.5	β=0.75	β=1
1	0.56	0.56	0.56	0.56	**0.67**
3	0.56	**0.63**	0.59	0.59	0.48
5	0.47	**0.53**	0.51	**0.53**	0.47
10	**0.44**	0.43	0.43	0.42	0.40
15	**0.37**	0.36	036	0.35	0.35

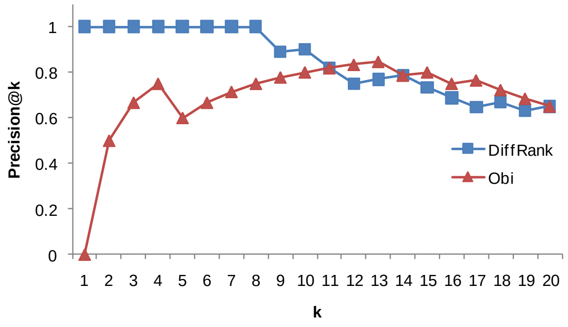

Figure 3. Precision at top-*k* for a test collection generated by the query "synapse".

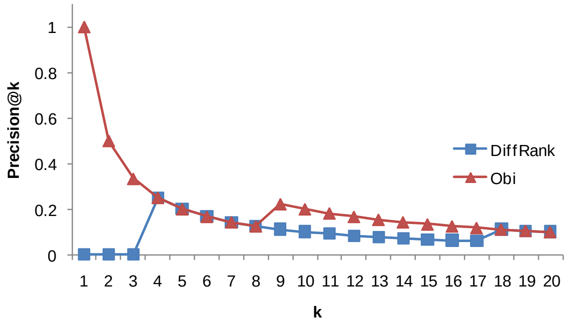

Figure 4. Precision at top-*k* for a test collection generated by the query "muscular dystrophy".

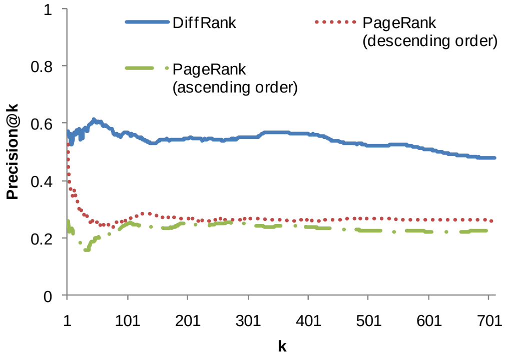

Figure 5. The overall effectiveness comparison by precision at top-*k* across 10 test collections. Ground truths were obtained automatically from Obi scores.

User Dependence. Comprehensibility of pages is naturally dependent on user level of knowledge, experience and cognitive skills. Thus certain pages may obviously be easier for certain groups of people. In this work we assume user-independent approach and we consider general users following educational system in Japan. Nevertheless, we are aware that the exhaustive approach for comprehensibility estimation should also include user factors. Nakatani et al. [10] demonstrated an interactive web search system for allowing users provide *comprehensibility feedback* to more effectively rerank search results. As users are usually unwilling to provide explicit feedback, approaches that can automatically estimate user knowledge and match it with difficulty levels of web pages are necessary.

In this paper we focused on finding easy pages as our prime objective, assuming that majority of users would require easy-to-understand documents especially for difficult and unknown topics. However, certain users such as professional ones may actually have opposite expectation wishing to retrieve scientific, professional or more detailed content. In such a case our proposed method may also be useful for ordering pages in descending order of their difficulty levels.

Other Comprehensibility Indicators. Except for text and multimedia, there are 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 type documents or than the documents that lack any such definitions or examples. We are aware however that the detection of this kind of indicators is not trivial.

Document structure and content presentation are other important aspects that influence the levels of content comprehensibility. 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.

VIII. CONCLUSIONS

In this paper we have proposed a novel method for comprehensibility measure of web documents. Instead of relying only on content analysis we utilized linkage between pages. Based on the experimentally verified hypothesis that readability of linked pages is similar to the one of the linking pages, we have used TrustRank algorithm to detect easy web pages. The experimental results proved the effectiveness of our approach.

ACKNOWLEDGMENT

This work was supported in part by the following projects and institutions: Japanese Ministry of Education, Culture, Sports, Science and Technology Grant-in-Aid for Young Scientists B: (Grant#: 22700096), and a Kyoto University GCOE Program entitled “Informatics Education and Research for Knowledge-Circulating Society,” and the National Institute of Information and Communications Technology, Japan.

REFERENCES

[1] J.S. Chall, and E. Dale, "Readability revisited: the new Dale-Chall readability formula", Brookline Books, Cambridge, Mass., 1995.

[2] E. Dale and J.S. Chall, "The concept of readability", Elementary English, 26(23), 1949.

[3] L. Feng, N. Elhadad, and M. Huenerfauth, "Cognitively Motivated Features for Readability Assessment", Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL 09), 2009, pp. 229-237.

[4] R. Flesch, “A new readability yardstick”, Journal of Applied Psychology, 1948, 32(3), pp. 221-233.

[5] Z. Gyöngyi, H. Garcia-Molina, and J. Pedersen, ”Combating Web Spam with TrustRank”, Proceedings of the 30th International Conference on Very large data bases, VLDB Endowment, 2004, pp. 576-587.

[6] M.-Y. Kan, "SlideSeer: A Digital Library of Aligned Document and Presentation Pairs", Proceedings of the Joint Conference on Digital Libraries (JCDL 2007), ACM, 2007, pp. 81-90.

[7] T. Kanungo, and D. Orr, "Predicting the Readability of Short Web Summaries", Proceedings of the 2nd International Conference on Web Search and Data Mining (WSDM 2009), ACM, 2009, pp. 202-211.

[8] T. P. Lau, and I. King, "Bilingual Web Page and Site Readability Assessment", Proceedings of the 15th International Conference on World Wide Web (WWW 2006), ACM, 2006, pp. 993–994.

[9] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank Citation Ranking: Bringing Order to the Web", Technical report, Stanford InfoLab, 1998.

[10] M. Nakatani, A. Jatowt, and K. Tanaka, “Adaptive Ranking of Search Results by Considering User’s Comprehension”, Proceedings of the 4th International Conference on Ubiquitous Information Management and Communication (ICUIMC 2010), ACM, 2010, pp. 182-192.

[11] M. Nakatani, A. Jatowt, and K. Tanaka, “Easiest-First Search: Towards Comprehension-based Web Search”, Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM 2009), ACM, 2009, pp. 2057-2060.

[12] E. Pitler and A. Nenkova, "Revisiting Readability: A Unified Framework for Predicting Text Quality", Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP 2008), Association for Computational Linguistics, 2008, pp. 186—195.

[13] S. Sato, S. Matsuyoshi, and Y. Kondoh, “Automatic Assessment of Japanese Text Readability Based on a Textbook Corpus”, Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008), ELRA, 2008.

[14] L. Si, and J. Callan. "A Statistical Model for Scientific Readability", Proceedings of the 10th International Conference on Information and Knowledge Management (CIKM 2001), ACM, 2001, pp. 574-576.