2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology

Towards Solving Comprehensibility-Relevance Trade-off in Information Retrieval

Kouichi Akamatsu
Dept. of Social Informatics
Kyoto University
Yoshida-honmachi, Sakyo-ku, Kyoto
akamatsu@dl.kuis.kyoto-u.ac.jp

Adam Jatowt
Dept. of Social Informatics
Kyoto University
Yoshida-honmachi, Sakyo-ku, Kyoto
adam@dl.kuis.kyoto-u.ac.jp

Katsumi Tanaka
Dept. of Social Informatics
Kyoto University
Yoshida-honmachi, Sakyo-ku, Kyoto
tanaka@dl.kuis.kyoto-u.ac.jp

***Abstract*—Comprehensibility is an important quality aspect of documents. Incomprehensible documents are of little utility to readers even if they are relevant. However, for many difficult queries such as technical ones, the topically relevant documents tend to be characterized by poor comprehensibility. This makes it difficult for users to satisfy their information needs when searching for documents about difficult topics. In this paper, we propose a novel approach to search for documents that explain query topics and are easy to understand for average users. In particular, we measure the comprehensibility and the relevance of documents based on the concept of Query Domain Graph constructed from Wikipedia articles related to the query. For estimating document comprehensibility we use the frequency and density of difficult terms within documents as well as we utilize graph-based document representation. We then propose retrieval techniques that balance the relevance and comprehensibility based on the concept of difficult word substitution, in which difficult words are replaced by the sets of easy and related words.**

I. Introduction

Documents on the Web are characterized by widely varying readability levels, and, so are also the reading skills, topic familiarity and education levels of Web users. This situation effectively limits information access leaving many web pages poorly comprehensible for significant population of users. The comprehensibility mismatch between authors and readers is also exemplified by initiatives such as Simple Wikipedia[1] project which offers simplified versions of Wikipedia pages written using no more than 2,000 most common English words. Note that Simple Wikipedia has however considerably smaller coverage than the English Wikipedia (respectively, 114k vs. 4,959k articles as of Sep 4, 2015).

Although, state-of-the-art web search engines return relevant documents, they may not always satisfy users' needs when it comes to the reading ease. For users who wish to learn about a difficult concept (e.g., technical or legal ones), it is sometimes difficult to find easy-to-understand pages using state-of-the-art search engines. A simple suggestion could be to extend original queries with comprehensibility indicators such as "easy", "beginner" or "tutorial" to indicate the need for easy to understand search results. However, this approach performs poorly since few documents contain such terms, and, even if they do so, it does not mean the returned documents are necessarily understandable [6]. In addition, suitable terms to be added usually depend on a query topic and might not be easily conceptualized by inexperienced users. For example, the queries "C programming language beginner" and "C programming language tutorial" might be effective for outputting easy pages, while the query "Alzheimer's disease beginner" will likely not be much useful.

Since the straightforward addition of special keywords to queries does not always work for difficult queries, a more refined approach has to be proposed. Prior works on text analysis developed a range of metrics for estimating text reading ease. Usually, such methods measure comprehensibility by extracting several surface features which influence comprehensibility. The most popular metrics use signals such as sentence length, word length or number of syllables in text [2], [3]. Others consider the popularity levels of content words in English language as measured by corpus statistics following the hypothesis that common words tend to be simple [1]. Note, however, that the opposite is not true. Rare words are not necessarily difficult. For example, many named entities (e.g., locations, person names) appear rarely in language corpora but they are not necessarily as difficult as are words denoting technical or scientific concepts (e.g., "Parkinson's disease" or "theory of relativity"). Neverteless, search results should be then improved when incorporating comprehensibility estimation in ranking.

In this work, we measure the intrinsic difficulty of words. In particular, we employ an approach which analyzes the link distribution in Wikipedia according to the intuition that Wikipedia articles on domain-specific terms tend to be linked mainly from the same topical domain. Having found the difficulty levels of words we estimate comprehensibility of web pages. They can be then ranked according to both their relevance and understandability as suggested above. However, such approach, although intuitive, still does not work well. The problem is that comprehensible documents tend to be about easy topics, while difficult documents tend to be about difficult topics. In other words, there are few easy-to-understand documents about difficult topics[2]. We call this phenomenon *comprehensibility-relevance trade-off* in Information Retrieval (IR). In practice, when a search engine does not consider topical relevance and comprehensibility in an appropriate way, search results returned for difficult queries will likely contain documents either comprehensible but without information that

[1] http://simple.wikipedia.org

[2] The opposite seems also to hold: there are few difficult-to-understand documents about easy topics.

978-1-4673-9618-9/15 $31.00 © 2015 IEEE
DOI 10.1109/WI-IAT.2015.209

a user needs (i.e., query words may be just a passing mention) or ones related to the query, yet, difficult for non-experts to comprehend or even to read. This hypothesis is reasonable considering that, usually, one needs skills and experience to be able to explain difficult concepts in an understandable and accessible way.

We then propose a set of methods for measuring comprehensibility and relevance. All of them utilize Wikipedia[3] in order to capture the levels to which documents explain topical domains of queries as well as to estimate their topic-based difficulty degrees. As the combination of comprehensibility and relevance evaluation is not trivial, considering the lack of the orthogonality between both the concepts for difficult queries, we propose a substitution-based approach. Our idea is to substitute difficult terms by the set of easy terms which retain the core meaning of difficult terms, yet, altogether, are more comprehensible. In the experimental section of the paper we then demonstrate the effect obtained by such approach using various combinations of methods for returning both relevant and comprehensible documents.

To sum up, we make the following contributions. (a) We introduce a novel research problem of solving the *comprehensibility-relevance trade-off* within IR. (b) To solve it we propose various methods for computing comprehensibility of documents based on Wikipedia and for combining them with relevance estimators. (c) We demonstrate in the experiments the results of different combinations of the proposed methods for alleviating the effect of the comprehensibility-relevance trade-off in web search.

II. Related Work

Measuring text readability can be done on different levels. On a syntactic level, the average sentence length is often considered as an indicator of grammatical difficulty of texts. Flesch Reading Ease [3], one of the earliest standard measures, defines the readability as a function of word length and sentence length. The approach is simple to implement but may not perform well on web pages with rich presentation styles. Moreover, this measure requires correct usage of punctuations and complete sentences. Thus long lists of phrases or words may lead to incorrect readability scores [7], [13].

Another category of works focuses on difficulty levels of words themselves. Word length or syllable count can be simple approximations for word complexity [3]. Another approach is to use a predefined list of common or easy words to identify unfamiliar words [1]. However, due to dynamic characteristics of languages, such static list requires an update every time new common words appear, or, when previously difficult words become widely understood.

The third category of works treats readability estimation as a classification problem. This line of research utilizes various features ranging from surface text features (e.g., word length) to discourse-level features (e.g., entities involved in a text) and from a manually compiled list of vocabulary to statistical language models (e.g., [5], [10], [12]). In general, these features address the problem at different levels of complexity: grammatical complexity, vocabulary complexity, and story complexity.

[3]In particular, we use the Japanese Wikipedia.

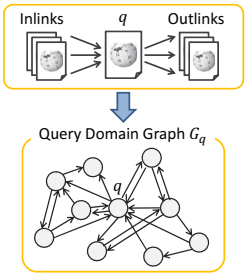

Fig. 1. Construction of the Query Domain Graph. Each node corresponds to a Wikipedia article and represents a concept expressed by the article title.

While the need for outputting comprehensible search results has been already noticed before [4], [5], [8], to the best of our knowledge, no prior work explicitly focused on the trade-off between comprehensibility and relevance in IR. The lack of the orthogonality between these two concepts in the search for difficult topics makes ineffective any straightforward approaches that simply re-rank outputted documents based on their comprehensibility (due to risk of outputting irrelevant results). In this work we focus more deeply on the interplay between the relevance and the comprehensibility of terms with regards to queried concepts.

Lastly, search engines for children can be considered related to this work. For example, AgeRank algorithm [4] ranks web pages according to their appropriateness for children. Our methods target however general users and standard search engines with the focus on the relation between the content difficulty and relevance.

III. Estimating Term Difficulty

Traditional readability indexes often exploit word difficulty as one of the features to measure text readability based on the assumption that average term difficulty is a strong indicator of document comprehensibility. In our approach, we utilize so-called *Query Domain Graph*, denoted as G_q, constructed by collecting Wikipedia articles related to the query in order to measure difficulty levels of words related to the query.

The construction of Query Domain Graph is illustrated in Fig. 1. Given a query q, we obtain a Wikipedia article which corresponds to q. This can be done by finding an article whose title has the highest cosine similarity with the query vector or by another retrieval technique. Next, we fetch all the inlinking and outlinking articles of that article[4]. We then incorporate all the collected articles including the one corresponding to the query into the Query Domain Graph as nodes. Edges in the graph represent links between the articles. An edge from a node A to node B within G_q means there is a hypertext link from the Wikipedia article corresponding to A to the one corresponding to B.

The nodes in the graph represent articles considered to be related to the input query. We then regard the titles of the Wikipedia articles represented in the Query Domain Graph G_q as *topic terms* related to the query, q. However, to measure the topic term difficulty we need to perform additional analysis as

[4]Note that the size of Query Domain Graph is arbitrary. Articles more than one hop away from the one representing the query can also be included.

TABLE I. EXAMPLES OF EXTRACTED TOPIC TERMS WITH THEIR DIFFICULTY SCORES FOR QUERY "DATA MINING".

Topic Terms	Difficulty Scores
Data mining	1.000
Organic intelligence	0.571
Support vector machine	0.458
Web mining	0.400
Decision tree	0.375
Statistical classification	0.348
Machine learning	0.250
Information retrieval	0.125
JavaScript	0.016
Google	0.008
Digital	0.006
English language	0.001

TABLE II. EXAMPLES OF EXTRACTED TOPIC TERMS WITH THEIR DIFFICULTY SCORES FOR QUERY IS "PARKINSON'S DISEASE".

Topic Terms	Difficulty Scores
Parkinson's disease	1.000
Myerson's sign	1.000
Tic disorder	0.887
Physical dependence	0.780
Lewy body	0.600
Psychosis	0.323
Asperger Syndrome	0.087
Perspiration	0.049
Michael J. Fox	0.037
Vitamin A	0.025
Coffee	0.007
Japan	0.001

topic terms may sometimes be easy even though queries are difficult. For example, when the query is "data mining" and the Japanese version of the Wikipedia is used, we can find that the article "Data mining" links both to the article "Support Vector Machine" and to the article "English language". The latter is a common and general term, while the former is more semantically constrained and more technical. To accurately capture the term difficulty, we utilize a method similar to the one proposed in [8]. Given the graph we calculate the difficulty scores for terms based on the idea that:

For a difficult query, if a given Wikipedia article is linked frequently from the inside of the query domain, yet, rarely from the outside of the query domain, then the topic term corresponding to that article is considered to be a difficult term (e.g., technical term).

Let G_A be the graph constructed from all the Wikipedia articles regarded as nodes and their inter-links represented as edges. Then the difficulty score of a term t is calculated as:

$$S(t) = \frac{LF(W_t, G_q)}{LF(W_t, G_A)} \tag{1}$$

where W_t is the Wikipedia article whose title contains t, and $LF(W, G)$ is the in-degree of W in graph G. Note that since G_q is a subset of G_A ($G_q \subseteq G_A$), then $LF(t, G_q) \leq LF(t, G_A)$ always holds, and, therefore, $S(t) \leq 1$. The above equation binds the term difficulty to the number of in-links of the Wikipedia article associated with the term. In addition, this number is normalized by the total indegree of the Wikipedia article to prevent very common terms (e.g., "English language") from receiving high difficulty scores.

Table I and II show the examples of extracted topic terms and their difficulty scores for the query "data mining" and "parkinson's disease", respectively.

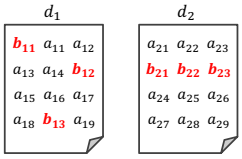

Fig. 2. Documents which have the same difficult term frequency but different difficult term density. a_{ij} and b_{kl} denote easy and difficult terms, respectively.

IV. DOCUMENT COMPREHENSIBILITY ESTIMATION

In this section, we describe different approaches to measure document comprehensibility by utilizing the difficulty scores of topic terms related to query.

A. Direct Measures

Term Difficulty. The first approach takes into account only the frequency of difficult terms based on a simple idea that the more difficult topic terms are in a document, the more difficult the document is. It also considers document length since long documents have higher probability to contain difficult topic terms. The comprehensibility of a document d is calculated by the following equation:

$$Comp(d) = -\frac{1}{\log |d|} \sum_{t_i \in TT(d)} S(t_i) \tag{2}$$

where $TT(d)$ is the set of the topic terms which appear in d, $|d|$ is the document length and $S(t_i)$ is the difficulty score of term t_i.

Term Difficulty and Density. The next method takes into account the density of difficult topic terms in a document in addition to their frequency based on the following hypothesis:

Document where difficult terms appear densely is more difficult than a document where difficult terms occur sparsely.

Fig. 2 illustrates two example documents which have the same length but different difficult term density. In d_1, the difficult words are sparsely distributed, while in d_2, they are densely arranged in a small section of the document. The difficult terms occurring in d_1 may be explainable by surrounding them easy terms. Or, if their meanings, associations and roles are not directly given in text, readers might still be able to infer them from the nearby context. On the other hand, the difficult terms in d_2 are mixed with only few easy terms. As it may be hard for an average user to comprehend such difficult terms without resorting to online search or other external resources, d_1 is considered more comprehensible than d_2. We make an assumption here that readers wish to fully understand the document they read. We thus neglect cases when readers need to understand only a part of a document or to just find some specific information (e.g., a single fact) in the document.

Following, in the next proposed method, we calculate the distance between every pair of topic terms in a document. The distance between two terms is calculated based on the number of characters separating the terms. If the same topic terms appear more than once in the document, we select the pair of terms such that the number of characters between them will be smallest, and then we represent the distance based on

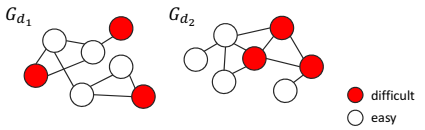

Fig. 3. Examples of Document Graphs containing easy and difficult words.

this number. The comprehensibility score of a document d is calculated by the following equation:

$$Comp(d) = -\sum_{t_i \in TT(d)} \sum_{t_j \in TT(d)} \frac{S(t_i) + S(t_j)}{\log(c+2)} \qquad (i < j) \tag{3}$$

where c is the number of characters between t_i and t_j.

B. Measures Based on Graph Representation of Documents

We next propose to measure the comprehensibility based on *Document Graph* constructed from a target document. Document Graph is a graph whose nodes are words[5] extracted from the document and edges represent their co-occurrence relations. Two nodes are connected if their corresponding words co-occur within a window of maximum N words (N=2 by default). We denote the Document Graph constructed from document d by G_d.

In the Document Graph, we assume the distance between two words to be equal to the length of the shortest path between the nodes corresponding to these words. In the examples shown in Fig. 3, difficult words appear sparsely in document d_1 while densely in d_2. According to the hypothesis described in Section IV-A, d_1 is then more comprehensible than d_2.

Before introducing the next methods for estimating document comprehensibility we need to calculate the difficulty score of each word in a document based on the difficulty scores of topic terms.

If a given word w in the input document can be exactly matched to some topic term t (i.e. title of a Wikipedia article) in the Query Domain Graph, then the difficulty score of w is equal to the difficulty score of that term:

$$S(w) = S(t) \qquad (w = t, t \in TT_q) \tag{4}$$

where TT_q is the set of topic terms of query q. However, if no topic term exactly corresponds to w, yet, some topic terms contain w (e.g., "mining" in "data mining" or "web mining"), then we calculate the difficulty score of w as the average value of the difficulty scores of all such topic terms. Let $t_1, t_2, ..., t_n$ be the topic terms which contain the word w. The difficulty score of w is then:

$$S(w) = \frac{1}{n} \sum_{j=1}^{n} S(t_j) \tag{5}$$

[5]Note that one can also apply POS-based filter for words. For example, we use only nouns and adjectives as nodes in the experiments.

Lastly, we let the difficulty score of w be 0 if there are no topic terms that correspond to w or that contain w.

Word Difficulty and Document Graph Based Density. This method considers both the difficult term frequency and their density using Document Graph. The equation is similar to Eq. 3 except that now the distance between two words is calculated based on the Document Graph. We calculate the distance between two words as the length of the shortest path between their corresponding nodes. The document comprehensibility is then:

$$Comp(d) = -\sum_{w_i \in TW(d)} \sum_{w_j \in TW(d)} \frac{S(w_i) + S(w_j)}{GD(w_i, w_j)} \qquad (i < j) \tag{6}$$

where $GD(w_i, w_j)$ is the length of the shortest path between the nodes corresponding to w_i and w_j on the Document Graph. $TW(d)$ is the set of words which appear in document d, and satisfy the condition $S(w) > 0$ ($w \in TW(d)$).

Word Difficulty Propagation. In the last method, we modify the word difficulty scores based on word co-occurrence relations. The assumption here is as follows:

If a word co-occurs with many difficult words, it is likely to be a difficult word.

We use here modified *TextRank* algorithm [9]. In particular, we calculate the word difficulty scores by the biased TextRank computation in which the random walk on the Document Graph is biased towards the nodes considered difficult. In other words, under the random surfer model, the surfer at each node has a small probability to teleport to the nodes denoting difficult terms.

$$S'(w_i) = \alpha * \sum_{w_j \in Adj(w_i)} \frac{S'(w_j)}{|Adj(w_j)|} + (1 - \alpha) * S(w_i) \tag{7}$$

where $Adj(w)$ is the set of adjacent nodes of a node w in the Document Graph and α is the damping factor which assumes a value between 0 and 1 (set by default to 0.85).

The document comprehensibility is then calculated using the newly computed word difficulty scores:

$$Comp(d) = -\frac{1}{\log |d|} \sum_{w_i \in W(d)} S'(w_i) \tag{8}$$

V. Topical Relevance Estimation

In this section, we describe how to measure the topical relevance of a document.

Since the topic terms described in Section III are the set of terms related to a query, we can assume a document containing many topic terms to be topically relevant to the query. We then generate a feature vector of query domain, called the *topical feature vector*, by using topic terms and we let the topical relevance of a document d be computed as the cosine similarity of the topical feature vector and d's feature vector.

However, this approach cannot discover documents which are relevant, yet, written using easy vocabulary, especially,

when most of the topic terms extracted from Wikipedia consist of difficult vocabulary. Our idea is then to, first, substitute the difficult terms and, then, to create a new feature vector over such substituted topic terms. Consider, for example, two texts that describe four main symptoms of the Parkinson's disease as shown in Table III. Although the contents are similar, the vocabulary is different. The symptoms "tremor", "rigidity", "bradykinesia" and "postural instability" in document A are expressed as "shaking in hands and legs", "stiff muscles", "slowness of movement" and "difficulty with balance" in document B, respectively. As we can see, while the content is similar in this case, easier vocabulary can be used to convey the same meaning.

To substitute a topic term, we utilize the abstract of its corresponding Wikipedia article based on the assumption that the objective of Wikipedia abstracts is to summarize the concept or entity described in the articles. Since there are many ways to perform topic term substitution, multiple topical feature vectors will be generated. The description of the way in which we create different topical feature vectors is deferred to Section VI-D.

After generating multiple topical feature vectors, we calculate the topical relevance of a document as follows:

$$Rel(d) = \max_i CosineSimilarity(\boldsymbol{d}, \boldsymbol{r_i}) \tag{9}$$

where $\boldsymbol{d}$ is the feature vector of a document d, $\boldsymbol{r_i}$ $(i = 1, 2, ...)$ is the set of the generated topical feature vectors after performing different substitutions.

VI. Experiments

In this section we first describe the experimental settings and then evaluate our methods of document comprehensibility and relevance estimation. Finally, we test two combinations of comprehensibility and relevance. Note that we do not evaluate the time cost in this work. We assume that in practice most of the scores are precomputed.

A. Test Collection

To collect data we have used 10 queries from different domains as shown in Table IV. We have then issued each query to the Bing Search API[6] and downloaded the top 30 results, creating thus a collection of 300 documents. Each web document was then scored by a human judge in terms of the comprehensibility and topical relevance. The comprehensibility was judged on a 1 to 5 scale where the meanings of scores are as follows: (1) very difficult, (2) difficult, (3) neutral, (4) easy and (5) very easy. On the other hand, the topical relevance scores range from 0 to 2: (0) irrelevant, (1) neutral and (2) relevant. The topical relevance of a page was judged based on whether the page explains the query concept or not. The question whether the page explains the query concept is an important factor for a user who wants to learn that concept. Note that this definition of topic relevance is narrower than the typical notion of relevance used in IR (documents explaining a query vs. documents about the query).

[6]http://datamarket.azure.com/dataset/bing/search

TABLE IV. The numbers of extracted topic terms for each query.

Query	Num. of Topic Terms
parkinson's disease	502
data mining	148
exchange traded fund	60
ribonucleic acid	296
complex number	424
ips cells	208
asbestos	347
philosophical realism	180
comparative advantage	121
neutron star	232

For each document, we also computed the combined score of the comprehensibility and topical relevance by multiplying the two kinds of ground truth scores.

B. Topic Term Extraction

For experiments, we downloaded the Japanese Wikipedia database which was dumped on May 21, 2014 by the Wikimedia Foundation[7]. We then used it to extract topic terms as described in Section III. The numbers of extracted topic terms for each query are shown in Table IV.

C. Evaluation of Comprehensibility Estimation

We evaluate in this section our approach for comprehensibility estimation.

In addition to the methods described in Section IV, we also compare the performance of baselines as shown below:

Bing. This baseline represents the original ranking returned from the Bing Search API. Its performance reflects how much a state-of-the-art web search engine (Bing, in this case) considers comprehensibility when ranking search results.

Obi-2[8]. Obi-2 is a readability measure tool developed for texts written in Japanese [11]. The output scores range from 1 to 13 and correspond to the Japanese primary school grades. Obi-2 uses a corpus extracted from textbooks used in Japan to calculate the likelihood of each grade based on the occurrence probability of the sequence of bigrams (two characters). The output is the score corresponding to the grade whose likelihood value is the highest for an input text.

For each query, we compare the agreement between the document ranking based on the ground truth comprehensibility scores of documents (the one that reflects human evaluation) and that generated by each method by using the Spearman's Rank Correlation Coefficient and nDCG. We show the results by the Spearman's Rank Correlation Coefficient in Table V and the results by nDCG in Table VI.

The results indicate that the proposed methods usually perform better than baselines with the exception of graph based approaches measured by the Spearman's Rank Correlation Coefficient Also, compared to measuring only the frequency of difficult terms we can still improve the performance by considering both the frequency and density of difficult terms as measured by the Spearman's Rank Correlation Coefficient. Moreover, for different queries different methods perform best.

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

[8]http://kotoba.nuee.nagoya-u.ac.jp/sc/obi2/obi.html

TABLE III. TWO EXAMPLE TEXTS WHICH DESCRIBE SYMPTOMS OF THE PARKINSON'S DISEASE.

A)	Four symptoms are considered cardinal in Parkinson's disease: tremor, rigidity, bradykinesia and postural instability.
B)	The four main symptoms of Parkinson's disease are shaking in hands and legs, stiff muscles, slowness of movement and difficulty with balance.

TABLE V. PERFORMANCE OF COMPREHENSIBILITY MEASURES BY SPEARMAN'S RANK CORRELATION COEFFICIENT.

	Bing	Obi-2	Term Difficulty	Term Difficulty and Density	Word Difficulty & Document Graph Based Density	Word Difficulty Propagation
parkinson's disease	0.268	0.204	**0.490**	0.442	0.415	0.451
data mining	-0.074	0.121	0.329	0.361	**0.438**	0.390
exchange traded fund	0.115	0.214	0.010	0.291	**0.390**	0.293
ribonucleic acid	-0.362	0.618	**0.757**	0.745	0.578	0.680
complex number	-0.021	0.128	**0.481**	0.393	0.310	0.365
ips cells	-0.140	**0.373**	0.236	0.273	0.205	0.228
asbestos	-0.392	0.259	0.259	0.271	0.269	**0.279**
philosophical realism	-0.310	**0.452**	0.269	0.273	0.057	-0.019
comparative advantage	-0.240	**0.537**	0.349	0.076	0.078	0.231
neutron star	-0.217	**0.273**	-0.029	0.141	0.231	0.089
Avg.	-0.137	0.318	0.315	**0.327**	0.297	0.299

TABLE VI. PERFORMANCE OF COMPREHENSIBILITY MEASURES BY NDCG.

	Bing	Obi-2	Term Difficulty	Term Difficulty and Density	Word Difficulty & Document Graph Based Density	Word Difficulty Propagation
parkinson's disease	0.874	0.847	0.952	0.951	0.954	**0.958**
data mining	0.893	0.881	0.927	0.921	0.926	**0.929**
exchange traded fund	0.921	0.935	0.934	0.942	**0.946**	0.942
ribonucleic acid	0.839	0.931	**0.996**	0.994	0.954	0.961
complex number	0.858	0.896	**0.941**	0.918	0.914	0.912
ips cells	0.914	0.930	**0.953**	0.941	0.939	0.935
asbestos	0.912	0.933	**0.966**	**0.966**	0.959	0.948
philosophical realism	0.805	0.914	**0.918**	0.917	0.848	0.836
comparative advantage	0.886	0.931	**0.939**	0.889	0.876	0.895
neutron star	0.850	0.887	0.883	0.896	**0.905**	0.901
Avg.	0.875	0.909	**0.941**	0.934	0.922	0.922

D. Evaluation of Topical Relevance Estimation

For the topical relevance estimation task, we compare the results of the methods listed below:

Bing. The original ranking by Bing Search API.

Topic Term Count. This method assumes the document length-normalized number of topic terms as the topical relevance score of the document.

Single Topical Feature Vector. This method generates a single topical feature vector based on the topic terms. Before generating the topical feature vector, difficult topic terms are selected and are substituted using the abstracts of their corresponding Wikipedia articles. As there are many ways to select topic terms for substitution, we evaluate the following cases:

- **No Substitution.** No topic terms substituted.
- **Threshold-based Substitution.** Topic terms with difficulty scores higher than pre-determined threshold are substituted. We set the threshold to 0.25, 0.5 and 0.75.
- **Complete Substitution.** All topic terms substituted.

To compute the topical feature vectors we divide each of the topic terms or abstract into words[9] and remove stop words. To determine words' weights, we use the term frequency (tf) weighting. In the same way, we generate a tf-based feature vector of each document. Finally, the topical relevance score of the document is calculated as the cosine similarity between the topical feature vector and the feature vector of the document.

Multiple Topical Feature Vectors. In this method, we generate multiple topical feature vectors and calculate the topical relevance score as the maximum of the cosine similarity scores between each topical feature vector and the feature vector generated from a document (see Eq. 9). We utilize here all of the five topical feature vectors described in the *Single Topical Feature Vector* method.

We evaluate each method in the same way as the evaluation of comprehensibility. That is, we compare the ground truth ranking given by the document relevance scores and the ranking given by each proposed method. We show the results by the Spearman's Rank Correlation Coefficient in Table VII and the results by nDCG in Table VIII.

As we can observe, the methods based on the vector space model perform better than the simple count of topic terms. We also notice that the substitution of topic terms is effective to improve the performance of the topical relevance estimation. Moreover, using the multiple topical feature vectors is superior than selecting only one of topical feature vectors.

E. Evaluation of the Combined Approach

Lastly, we combine both the comprehensibility and topical relevance measures and we evaluate the combined method. The ground truth data is obtained by multiplying the manually judged scores of the comprehensibility and topical relevance.

In addition to the Bing's original ranking used as a baseline method, we evaluate the selected combination methods as below:

[9]To extract words, we utilize MeCab, a Japanese morphological analyzer. http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html

TABLE VII. PERFORMANCE OF TOPICAL RELEVANCE MEASURES BY SPEARMAN'S RANK CORRELATION COEFFICIENT.

	Bing	Topic Term Count	Single Topical Feature Vector					Multiple Topical Feature Vectors
			No substitution	Threshold			Substitute all	
				0.75	0.5	0.25		
parkinson's disease	-0.065	0.466	0.319	0.588	**0.629**	0.588	0.580	0.621
data mining	-0.113	**0.544**	0.476	0.340	0.226	0.226	0.204	0.226
exchange traded fund	0.026	0.003	0.565	0.582	0.556	0.561	0.579	**0.588**
ribonucleic acid	0.156	0.455	0.533	**0.558**	**0.558**	0.571	0.533	**0.558**
complex number	-0.355	-0.009	0.082	0.191	0.173	0.191	**0.228**	0.209
ips cells	0.017	0.460	0.456	0.622	0.663	0.680	**0.688**	0.646
asbestos	0.115	0.571	0.201	0.497	0.623	0.647	**0.679**	0.623
philosophical realism	**0.610**	0.344	0.175	0.144	0.105	0.105	0.096	0.144
comparative advantage	0.193	-0.338	0.015	0.142	0.099	0.142	**0.199**	0.185
neutron star	0.373	0.383	0.351	**0.469**	0.419	0.419	0.411	**0.469**
Avg.	0.096	0.288	0.317	0.413	0.405	0.413	0.420	**0.427**

TABLE VIII. PERFORMANCE OF TOPICAL RELEVANCE MEASURES BY NDCG.

	Bing	Topic Term Count	Single Topical Feature Vector					Multiple Topical Feature Vectors
			No substitution	Threshold			Substitute all	
				0.75	0.5	0.25		
parkinson's disease	0.879	0.924	0.904	0.971	**0.974**	0.965	0.971	0.973
data mining	0.872	**0.978**	0.971	0.944	0.926	0.928	0.925	0.927
exchange traded fund	0.846	0.862	0.977	**0.979**	0.968	0.969	0.973	0.974
ribonucleic acid	0.967	0.981	0.987	0.992	0.992	**0.993**	0.991	0.992
complex number	0.932	0.971	0.977	**0.986**	0.979	0.979	0.983	0.982
ips cells	0.821	0.908	0.886	0.959	0.964	**0.972**	0.966	0.962
asbestos	0.855	0.960	0.735	0.929	0.959	0.956	**0.963**	0.959
philosophical realism	**0.992**	0.977	0.955	0.920	0.919	0.917	0.916	0.924
comparative advantage	**0.975**	0.921	0.923	0.930	0.924	0.958	0.968	0.964
neutron star	0.979	**0.980**	0.884	**0.980**	0.967	0.967	0.959	**0.980**
Avg.	0.912	0.946	0.920	0.959	0.957	0.960	0.962	**0.964**

1) Obi-2 - Topic Term Count
2) Term Difficulty - Single Topical Feature Vector (No substitution)
3) Term Difficulty - Multiple Topical Feature Vectors
4) Term Difficulty and Density - Single Topical Feature Vector (No substitution)
5) Term Difficulty and Density - Multiple Topical Feature Vectors
6) Word Difficulty and Document Graph Based Density - Single Topical Feature Vector (No substitution)
7) Word Difficulty and Document Graph Based Density - Multiple Topical Feature Vectors
8) Word Difficulty Propagation - Single Topical Feature Vector (No substitution)
9) Word Difficulty Propagation - Multiple Topical Feature Vectors

To calculate the final scores for each of these methods, we try two different combination ways. The first one is the multiplication of both the comprehensibility and relevance scores as given by the methods that estimate the comprehensibility and the topical relevance, respectively:

$$Score(d) = Comp'(d)^{\alpha} \cdot Rel'(d)^{(1-\alpha)} \tag{10}$$

where $Comp'(d)$ and $Rel'(d)$ denote the normalized scores of comprehensibility and topical relevance, respectively.

The results of this combination are illustrated in Fig. 4 (Spearman's Rank Correlation Coefficient) and Fig. 5 (nDCG). Note that the vertical axis represents the average Spearman's Rank Correlation Coefficient or nDCG score over all the queries.

The second combination method we test is the threshold-based approach:

$$Score(d) = \begin{cases} Comp'(d) & (Rel'(d) > \theta) \\ 0 & (Rel'(d) \leq \theta) \end{cases} \tag{11}$$

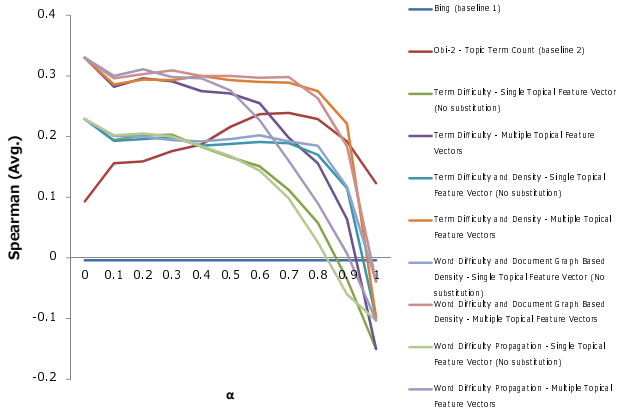

Fig. 4. Performance comparison of the combined methods (combination by product) by Spearman's rank correlation coefficient.

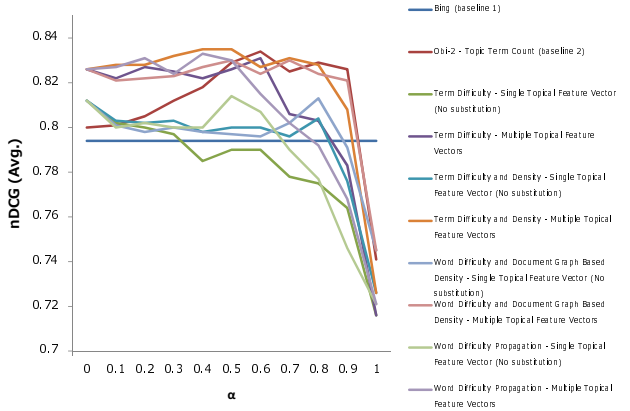

Fig. 5. Performance comparison of the combined methods (combination by product) by nDCG.

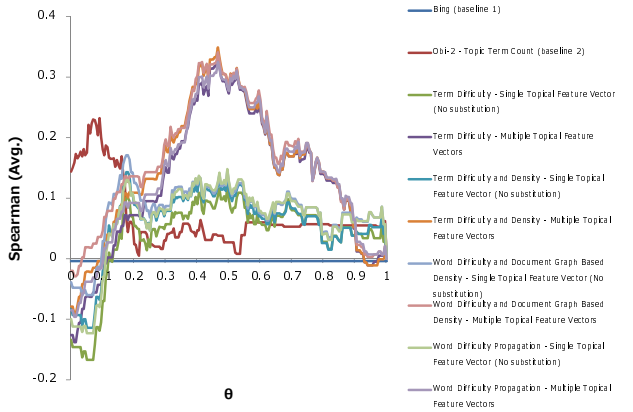

Fig. 6. Performance of the combined methods (threshold-based combination) by Spearman's Rank Correlation Coefficient.

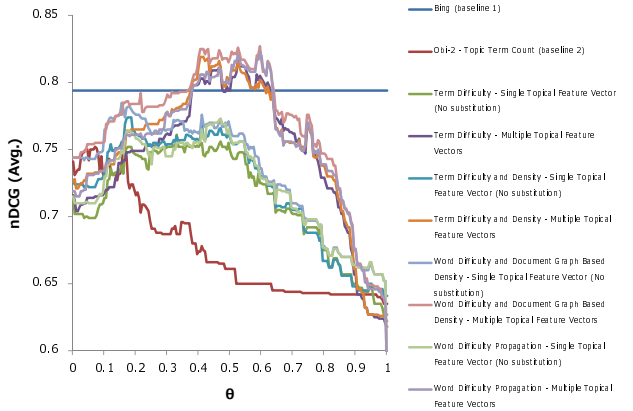

Fig. 7. Performance of the combined methods (threshold-based combination by nDCG.

The results of this combination strategy are illustrated in Fig. 6 and 7. As we can observe, our methods in general perform better than the baselines for both combination ways. The best performance is achieved when combining the *Term Difficulty and Density* or *Word Difficulty and Graph Based Density* methods as a comprehensibility measure and *Multiple Topical Feature Vectors* as a topical relevance measure using the threshold-based combination (Spearman's Rank Correlation Coefficient value of 0.349). *Word Difficulty Propagation* when combined with the same topical relevance method tends to output similarly good results.

We consider the combination using the threshold to be more effective for discovering both comprehensible and topically relevant pages. When calculating the score using Eq. 10, one cannot know whether the score is small/high because of the comprehensibility or rather due to the topical relevance. On the other hand, when using Eq. 11, if topical relevance is low (lower than the threshold), no matter how high comprehensibility is, the page score would be zero. If the comprehensibility is low and the topical relevance is high, page score would be small because the score is calculated as the comprehensibility score itself.

Finally, we explain the sudden drop in the performance for α close to 1. It is because comprehensible but irrelevant documents are returned for which the combined ground truth score becomes 0 due to the relevance score equal to 0 (relevance scale from 0 to 2).

VII. Conclusions

In this paper, we demonstrate our approach to realize effective search based on document comprehensibility. In particular, we focus on the comprehensibility-relevance trade-off in IR which means that easy documents returned for difficult queries are often noisy ones that are not related to the topic of the query and, hence, ones that do not let readers understand the queried concepts. We then assume that comprehensibility has to be considered in strict combination with the document relevance to construct an efficient search engine that would output appropriately explanatory search results. To this end, we propose several measures for estimating the comprehensibility and topical relevance of documents using Wikipedia as knowledge source. The combination of these measures is then used for assigning utility scores to documents (i.e., combined relevance-comprehensibility scores).

In future, we plan to conduct more extensive experimentation with queries ranging over different topical domains.

VIII. Acknowledgments

This work has been partially supported by Grant-in-Aid for Scientific Research (No. 15H01718) from MEXT of 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] R. Flesch, "A new readability yardstick.", Journal of Applied Psychology, 32(3), 221-233, 1948.

[4] K. Gyllstrom and M.-F. Moens, "Wisdom of the ages: toward delivering the children's web with the link-based agerank algorithm," in *Proc. of WSDM 2009*, 2009, pp. 202–211.

[5] T. Kaungo and D. Orr, "Predicting the readability of short web summaries," in *Proc. of CIKM 2010*, 2010, pp. 159–168.

[6] M. P. Kato, T. Yamamoto, H. Ohshima, K. Tanaka, "Investigating users' query formulations for cognitive search intents.", SIGIR, 577-586, 2014.

[7] T. P. Lau and I. King, "Bilingual web page and site readability assessment," in *Proc. of WWW 2006*, 2006, pp. 993–994.

[8] M. Nakatani, A. Jatowt, and K. Tanaka, "Easiest-first search: Towards comprehension-based web search," in *CIKM 2009*, pp. 2057–2060.

[9] R. Mihalcea and P. Tarau, "Textrank: Bringing order into texts," in *Proc. of EMNLP 2004*, 2004, pp. 404–411.

[10] E. Pitler and A. Nenkova, "Revisiting readability: A unified framework for predicting text quality," in *Proc. of EMNLP 2008*, pp. 186–195.

[11] S. Sato, S. Matsuyoshi, and Y. Kondoh, "Automatic assessment of japanese text readability based on a textbook corpus," in *Proc. of LREC 2008*, 2008, pp. 654–660.

[12] L. Si and J. Callan, "A statistical model for scientific readability," in *Proc. of CIKM 2001*, 2008, pp. 574–576.

[13] A. L. Uitdenbogerd, "Web readability and computer-assisted language learning," in *Proc. of ALTW 2006*, 2006, pp. 99–106.