An Unsupervised Technical Readability Ranking Model by Building a Conceptual Terrain in LSI

Shoaib Jameel and Xiaojun Qian

Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong.
{msjameel,xjqian}@se.cuhk.edu.hk

***Abstract*— Searching for domain-specific related information has gained a high popularity in recent years. Naturally, everyone is not at par with each other when it comes to knowledge about the concepts of a domain. A doctor may be well versed in her field of specialization and probably would search for advanced medical documents on the Internet. But she may look for a much simpler material related to Computer Programming. However, current information retrieval (IR) systems just return a mixed set of results based on similarity and popularity of the web pages. Existing methods which have tried to address the issue of matching readers with texts in domain-specific IR either use an ontology or some seed concepts thereby limiting their application in certain domains only. Moreover, readability methods cannot address the issue in domain-specific IR ranking because they fail to give precise prediction when applied on web pages. We address this problem in domain-specific search using a conceptual model where the sequence of the terms in a document is modeled as a connected conceptual terrain. Our model has achieved significant improvement in ranking documents by technical readability.**

I. Introduction

Technical readability relates to the problem of reading difficulty in domain-specific documents. In a study conducted by Pew Internet[1], it has been concluded that fifty two million American adults have used the Internet to get health related information. This clearly portrays the popularity of domain-specific search. People having diverse background query web search engines or domain-specific vertical search engines to find a document which is both relevant and can fit the level of understanding. A student might search for *"VLSI"* and wants *conceptually* very basic content which is easily comprehensible. In contrast, a researcher specializing in *"VLSI"* will expect completely different set of results from an Information Retrieval (IR) system. It is indeed difficult to cater for such needs without building a user model [1] for every user and reflect results accordingly. But building a user model for every user requires a huge amount of individual user session data [2]. Search engines need to record every search session and clicks of the user in order to accomplish personalization of the search results. Many users might not want their sessions to be recorded due to privacy concerns [3].

Present general web search engines cover a diverse range of topics. They mainly use similarity and popularity based methods to find a web page which is a close match to the users' query [4]. Latent Semantic Indexing (LSI) [5], [6] based retrieval has been applied to match the readers with texts using user studies [7]. But LSI for large scale retrieval of documents incurs huge computational cost as queries have to be folded-in the latent space every time a new query is issued [5].

Rank	URL	Title	Category
1	en.wikipedia.org/wiki/Biopsy	Biopsy	Educational
2	webmd.com/cancer/what-is-a-biopsy	Biopsy	Technical
3	nhs.uk/conditions/biopsy	Biopsy	Technical
4	nlm.nih.gov/medlineplus/ency/article/003416.htm	Biopsy	Technical
5	nlm.nih.gov/medlineplus/ency/article/003920.htm	Breast biopsy	Technical
6	netdoctor.co.uk/health_advice/examinations/biopsy.htm	Biopsy	Educational
7	breastcancer.org/symptoms/testing/types/biopsy.jsp	Biopsy	Technical
8	medterms.com/script/main/art.asp?articlekey=2466	Biopsy definition...	Technical
9	breastbiopsy.com/	Breast Care: Mammograms…	Educational
10	cancer.gov/dictionary?cdrid=45164	Definition of biopsy	Educational

Fig. 1. Ranked list of web pages obtained from Google for the query: *"biopsy"*.

Consider a domain-specific query *"biopsy"*. We asked a student who was not well versed in Medical Science to query this term on Google. We show the top ten results that was returned by Google in Figure 1. After clicking on the top result, namely Wikipedia and reading the text, he found that the page was too technical for him as it contained lot of domain-specific terms which required thorough understanding about other medical terminologies. He then sifted through several documents down the ranked list and eventually found a relatively simple web page at the tenth position, which fairly matched his technical comprehension level as the web page defined the jargon in simple terms. So, simply re-ordering the results automatically based on the choice of technical readability of the user will aid the user in getting relevant results for this user. In this case, a relevant document is one which not only follows query-document similarity but is also a good match for technical readability. Although it requires the search engine such as Google to classify queries as domain-specific ones, but we do not address query classification task in this paper. Similar approach to domain-specific ranking on web search engines has also been proposed in [4], where the authors did not address query classification task.

If results are presented to the user in mixed order of technical readability, then the user needs to spend a lot of time to find the web page which could suit his understanding level. It is indeed challenging, though not impossible, for a search engine to predict the intent or the technical readability level from the query itself. But determining the technical readability level of a person from query is also very challenging without user data for personalization of search results. In addition,

[1]http://www.pewinternet.org/Reports/2000/The-Online-Health-Care-Revolution/Summary.aspx

978-0-7695-4794-7/12 $26.00 © 2012 IEEE
DOI 10.1109/SKG.2012.20

most of the queries are short and ambiguous [8]. An alternative solution is to provide an interface to let a user select her own technical readability level using a slider on the search result page as described in [9] or a feature recently introduced by Google (under More Search Tools section) for ordering search results based on "reading level" where a user adjusts the reading level based on her requirements. This generalized feature can be an option in case user data is not available for personalization.

We present an unsupervised method for re-ranking domain-specific documents given a query based on the technical readability of documents. Technical readability relates to reading difficulty of technical documents. We first obtain search results for a query from an IR system which mainly aids in retrieving relevant documents but in a mixed technical readability order. Subsequently, documents retrieved by an IR system are automatically re-ranked based on the readability of the documents using our proposed model. The contrasting feature of our approach from the previous readability based approaches is that apart from the obvious domain-specific terms with high syllable counts (they can also be easily captured by the readability formulae), our model can also capture domain-specific terms which have low-syllable counts and are central to a particular technical context such as "shock" (in the context of earthquakes) which has only one syllable, "star" (in computer networking) again one syllable. To accomplish this goal, we employ a latent semantic method known as LSI. LSI exploits the co-occurrence matrix to bring out new structural relationships between the terms and documents in the latent space. A common observation in the LSI latent space is that terms which are semantically linked with the contents of the documents come close to the document vector where as common terms move far from the document vector. In the low-dimensional latent space, we compute cohesion between the sequences of terms. We compute the aggregated technical readability via a terrain traversal cost. This cost will be used to re-rank the search results obtained from a general purpose information retrieval system.

Our research contributes in the following ways: Our approach can differentiate the technical centrality based on contextual information (semantics of the text) and does not depend on text's surface level features. Past works mainly unsupervised readability methods cannot capture such feature and hence underperform in domain-specific documents. The main innovation of our approach is in using a conceptual model to tackle technical readability problem. Domain-specific readability methods proposed in the past have made use of an ontology or some seed concepts in their method. This limits the scalability of their method in other domains. We conduct experiments in two popular domains and compare the ranking effectiveness of our model with the current state-of-the-art methods.

II. Literature Review

Much research has been done in measuring the text's readability, for example [10], [11], [12]. But readability tests do not perform well on technical texts and web pages [13], [14]. Readability metrics mainly work on syllables and number of characters, sentence difficulty, and common terms etc, which are surface level features of a text. Hence, they are not effective when finding the technical readability of text [13] and [15]. Readability formulae measure the difficulty of a particular discourse using two main components - syntactic and semantic. The syntactic component measures the individual sentence length and so on whereas the semantic component measures individual term's syllable counts. More information about these components can be found in [13], [14]. Nakatani et al. [16] describe a way to re-rank the search results of a web search engine (Easiest-First Search) in descending order of their comprehensibility using Japanese Wikipedia. They have also used readability based method in their approach.

Readability methods do not give reliable results when applied on web pages [14] and very large text collections as a whole because of high computational costs [13]. Consider short snippets of texts taken from Wikipedia about *"biopsy"*.
1. A biopsy is a medical test involving sampling of cells or tissues for examination.
2. A biopsy of the temporal arteries is often performed for suspected vasculitis.

Sentence 1 has a Flesch reading ease of 41.55 (marginally readable) and Sentence 2 has 18.40 (difficult text). But examining these two texts closely, one can notice that both are difficult for a general reader. In the first sentence, terms used although appear common to a readability formula, but require some technical knowledge to comprehend. Terms such as *"sampling"*, *"cells"* and *"tissues"* carry domain-specific technical meaning. Consider another example *"Similar and popular web pages."* For a readability formula, terms such as *"similar"* and *"popular"* are common. But these two terms carry underlying technical meaning in the domain of IR. One objective of our method is to address this issue.

Some supervised machine learning approaches have also been studied. Language modeling approach has been applied to readability [14] where the authors describe a smoothed unigram model for computing the readability of non-traditional documents like web pages. Their method computes probabilities of every token in the corpus based on the usage across documents classified into various American grade levels. Another work [17] uses Support Vector Machines (SVM) and the authors have used an automated method for recognizing the reading levels of texts from user queries. They have used syntactic and vocabulary based features to train the classifier. Topic familiarity is different from traditional readability [9] where the authors study the re-ranking of search engine results based on familiarity. The authors suggest that traditional readability methods cannot predict familiarity level of a document. The authors inferred that stop words are the most important feature in their familiarity classifier. Importance of conjunctions has also been studied in [18], [19] where it has been concluded that conjunctions help an average reader in comprehending a discourse. In [17], the authors use SVM to determine the reading level from queries. Supervised

machine learning approaches need huge training data labeled as introductory, intermediate or advanced [20], which is the limitation of the supervised methods. In contrast, our method does not need any annotated data.

A recent study conducted in [21] describes the personalization aspect of ranking the search results (apart from general search engine ranking). The authors study key problems of the estimation of user proficiency, the search result difficulty and re-ranking of the search results based on readability. Other works which have considered personalization include [20] and [22]. Some works have considered semantic content in text by using diverse linguistic features [23], [24] but have mainly focused on classification for text quality rather than ranking and the prime focus of IR being ranking [25]. Moreover, both works have adopted linguistic means to extract features for classifier. Also, in [24], the authors have presented their work as a proof-of-concept and the authors have stated that such feature extraction still does not exist. In [26], the author proposes document ranking based on text quality.

In [13], the authors illustrate the shortcomings of readability metrics when applied to technical documents. They use an external domain-specific knowledge base which is an "ontology tree". They describe the notion of document scope and document cohesion and hypothesize that these two factors aid in determining the concept level readability of text. One problem with their method is that they need a domain-specific ontology, where as our method does not need any ontology for technical term detection. Zhao et al. [27] describe an iterative method for computing the readability of text but the limitation of their approach is that their method needs some seed technical concepts for initialization. We have previously described a model [28], [29] using latent semantic indexing [5] in which they compute document cohesion and technicality. The model and empirical results are quite preliminary.

In summary, our method is better than previous works in the following ways: 1) Our method does not need any domain-specific ontology to detect technical terms in a document. 2) Our method does not require any private user data for personalization. We have an option to let user specify her own reading level. 3) Our method improves upon the shortcoming of the readability metrics by capturing the semantics of text than surface-level features. 4) Our method can be effectively applied on domain-specific search systems such as PubMed search and vertical search engines, which mainly deal with domain-specific searching. 4) Our method does not need any expensive annotated data for learning.

III. The Conceptual Terrain Model

To effectively model the technical readability of a document, we exploit the latent concept information embodied in a document. We therefore consider the Latent Semantic Indexing (LSI) approach. In our work we compute a term's technical difficulty in the latent space and then connect them sequentially. The setup can be viewed as a terrain. One observation in the latent space is that terms which are coherently linked with the document are close to their document vectors in the latent space derived from a set of domain-specific documents [30] and semantically similar documents and terms come close to each other. The major factor that aids in bringing semantically coherent terms close their document vectors is that in a large corpus, terms normally overlap across several documents. Common terms will occur in several documents whereas domain-specific terms will occur rarely across several documents in the collection. Thus, the more the term is shared across several documents, the lesser important it becomes. This property helps techniques such as LSI to bring out latent structural information embodied in the document together by coherently linking the terms with the documents. In domain-specific documents coherent terms are mainly jargon because the domain-specific documents mainly describe them and thus they remain central to the technical theme of that document in the entire body of the document.

In LSI, computing the SVD of a matrix was generally computationally expensive both in space and time complexity [31]. But with the fast development of better algorithms to compute the SVD, such concerns both in terms of time and space complexities have been addressed [32], [33], [34]. Some methods do not even compute the SVD and adopt a completely different approach for faster latent concept finding such as Random Projection method [35].

The LSI model has some limitations such as losing important structural information from the document which could prove useful in leveraging extra information from the text [36]. The process of technical discourse comprehension involves the identification of the inherent meaning of technical terms and continuity of the same topical theme across the language of discourse [37] so that a reader is able to relate with other parts of the text. Hence we compute the term difficulty and term cohesion which measures cohesion in sequence so that we can capture the major term difficulties and conceptual leaps. The greater the technical readability of a term, the more cognitive load a user needs to expend in order to find out the inherent meaning of the term. In technical texts, cognitive difficulties will mainly arise in places where a reader encounters a domain-specific jargon or a phrase expressing certain complex concept [38]. Comprehending the discourse also becomes tougher when the topical correlation with the surrounding texts is low [39]. In reality, the human reading process and comprehension may be more complicated than what we have assumed here. For example, we can consider prior knowledge such as considering the past contexts which she has already read, for instance, past paragraphs, last n terms, n-grams, even previous documents already read related to the current one. Such propositions would indeed substantially increase the computational complexity of the model. Thus we consider the past one term as a contextual history which we believe can help capture major conceptual leaps when transiting from one term to another in sequence and such consideration is computationally less expensive for large datasets. In [18], the authors state that texts frequently exhibit varying degrees of cohesion in different sections. The start of the text cannot be cohesive with the later sections. Based on such consideration,

we develop a terrain model where we maintain the order of the terms.

A. Term Centrality

Term Centrality is a measure of cohesion between the term and document vectors in low dimensional latent space. The closer a term is to the document in the latent space, the more central it is in the technical context. In the latent space, a term which contributes with more synergy will be close to its document vector and unrelated terms will be relegated as unimportant under that technical theme [30]. For example, a document which describes about "earthquakes" will have terms such as "shock", "seismic", "shake", "earth" as central terms (if the document contains these terms). We denote "term centrality" as $\Delta_{t_n}^{(d)}$ for the term t at position n in the document d whose corresponding vectors in the latent space (i.e., term and document vectors) are separated by the Euclidean distance, $\rho_{t_n}^{(d)}$, as follows:

$$\Delta_{t_n}^{(d)} = \frac{1}{\rho_{t_n}^{(d)} + \psi} \tag{1}$$

where ψ is a very small constant added to accommodate the case when $\rho_{t_n}^{(d)} = 0$ and in general $\psi << \min(\rho_{t_n}^{(d)})$. The lesser the distance between the term vector and the document vector, the higher will be $\Delta_{t_n}^{(d)}$. The centrality of general terms will be low because of their large semantic distance from the document vectors.

B. Term Cohesion

In the LSI latent space, terms with similar semantic meaning tend to cluster close to each other. In order to comprehend a piece of text, semantic associations between the terms with other terms in the vicinity is essential. It is so because the alignment of meaning in a textual discourse depends the contextual history [40]. This results in the continuity of an idea in text. Studies have been conducted about such cohesive phenomenon in texts, for example, in [37], [40]. In our model, cohesion is computed between the two consecutive terms in a document. Let d be a particular document in the corpus. We use the Euclidean distance formula in the LSI latent space to compute the term cohesion between the term t_n and the term t_{n+1}, which we denote as as $\hat{\theta}_{(t_n,t_{n+1})}$.

By hopping/traversing from one term to another i.e., two consecutive terms in the terrain, we attempt to capture conceptual leaps while traversing the terrain. The more the conceptual leap, the more difficult will be the path of the reader. This affects the overall technical readability of a document. If a sequence terrain is comprised of series of domain-specific terms which are separated by large semantic distance, then a typical reader will probably leave the document and search for a more technically simpler and semantically cohesive piece of discourse.

In order to normalize the semantic distance between the two consecutive terms t_n and t_{n+1}, we proceed this way:

Step 1: Compute in the latent space the Euclidean distance $\hat{\theta}_{(t_n,\tau)}$ between t_n and each term τ which follows t_n in the entire corpus.

Step 2: Normalize the values such that the total sum of the normalized distances is 1. We denote the normalized distance as $\theta_{(t_n,t_{n+1})}$, which is the normalized semantic distance between the terms t_n and t_{n+1}. We use the following formula:

$$\theta_{(t_n,t_{n+1})} = \frac{\hat{\theta}_{(t_n,t_{n+1})}}{\sum_{\tau} \hat{\theta}_{(t_n,\tau)}} \tag{2}$$

C. Term Difficulty

Term difficulty is the relative technical readability of a term with respect to other terms in the document which is characterized by the document frequency count and its centrality to a document. The term difficulty score $\hat{\xi}_{t_n}^{(d)}$ for every term in the document is formulated as:

$$\hat{\xi}_{t_n}^{(d)} = idf_{t_n} \times \Delta_{t_n}^{(d)} \tag{3}$$

where $\Delta_{t_n}^{(d)}$ is the centrality component given in Equation 1. idf_{t_n} is the Inverse Document Frequency as in [41].

Documents describing about one particular technical discourse will have related terms close to it in the latent space. However, cases might arise that a document is specific about a technically simple term like "food", "water", etc which do not contribute considerably in technical readability. Technical difficulties of such terms need to be discounted and the idf_{t_n} has been introduced as a discounting component for those less technical central terms. Equation 3 ensures that the term difficulty of a general non-central term is always low because both the centrality component $\Delta_{t_n}^{(d)}$ and the idf_{t_n} are small. However, terms which are common but central will get relatively higher score when compared with non-central common function words as the centrality component $\Delta_{t_n}^{(d)}$ will be high. In cases where both the central and idf_{t_n} are high, that term will be highly difficult in that technical document.

Equation 3 gives non-standardized values of the term difficulties. We normalize $\hat{\xi}_{t_n}^{(d)}$ in such a way that the monotonicity of the global importance of terms is preserved. Hence, we transform the values to $[0, 1]$. The final term difficulty is computed as follows:

Step 1: Construct the term-document matrix such that terms are represented along the rows and documents are represented along the columns i.e., terms by document matrix. The elements of the term document matrix are the $\hat{\xi}_{t_n}^{(d)}$ values which denotes the term difficulty of term t_n in document d.

Step 2: Normalize the values such that sum of all elements in each row is 1. Let D represent the total number of documents in the corpus. We adopt the following formula:

$$\xi_{t_n}^{(d)} = \frac{\hat{\xi}_{t_n}^{(d)}}{\sum_{i=1}^{D} \hat{\xi}_{t_n}^{(d_i)}} \tag{4}$$

D. Cost Computation to Facilitate Ranking

We conjecture that the readability is directly proportional to the term difficulty and also term cohesion i.e., semantic distance between the two consecutive terms. This leads us to the fact that the more the term difficulty and the greater the semantic distance, the more will be the technical readability

problem in that portion of the text where the sequence of terms appears. Assume a reader begins from term t_n in the document d with term difficulty denoted as $\xi_{t_n}^{(d)}$ and hops to the next term t_{n+1} in sequence (term difficulty score denoted as $\xi_{t_{n+1}}^{(d)}$ covering a semantic distance $\theta_{(t_n,t_{n+1})}$ in between the two terms in the latent space. We compute the traversal cost, $C_{(t_n,t_{n+1})}^{(d)}$ for each sequential term bigram in document d in the conceptual terrain which is expressed as:

$$C_{(t_n,t_{n+1})}^{(d)} = \alpha\left[\xi_{t_n}^{(d)} + \xi_{t_{n+1}}^{(d)}\right] + (1-\alpha)\theta_{(t_n,t_{n+1})} \quad (5)$$

where α ($0 \leq \alpha \leq 1$) is a parameter indicating the role that each of the components plays in determining the technical readability of a document. $\theta_{(t_n,t_{n+1})}$ is calculated as in Equation 2.

A closer look at Equation 5 reveals that transitions in the terrain take place considering two terms in sequence in the document's conceptual terrain. It is because transitions consist of the current history in context and one past history in memory in order to compute the relative conceptual leap that is needed to connect the two pieces of terms together. Consider a situation when there are two consecutive difficult terms located semantically distant from each other. In this case the overall conceptual difficulty for this transition will be high because of the technical difficulties of the two terms and the semantic separation between them. Another instance arises when two difficult terms are semantically close to each other. In that case, the individual term difficulty component plays a major role in conceptual difficulty of the discourse. The premise is that even though the terms are cohesive, their individual difficulties are high making discourse comprehension difficult for the reader. What we are measuring from Equation 5 is the role of term difficulty and inter-term hops coupled together to compute the overall bigram transition cost as both difficulty and inter-term hops play some role in the overall conceptual difficult of a discourse and locally at every transition.

E. Ranking

The prime motive of our approach is to find the relative technical readability (domain-specific readability) of a document d when moving over the text sequentially in fixed leaps between bigrams. The difficulty of terms is computed by the scores in the latent space that represent the deviation from common terms and the cost of computing cohesion between the terms in sequence. We aggregate the difficulties and transitions costs obtained from Equation 5 to come up with the document's technical readability, E_d formulated as:

$$E^{(d)} = \frac{\sum_{n=1}^{T_d-1} C_{(t_n,t_{n+1})}^{(d)}}{T_d - 1} \quad (6)$$

where $C_{(t_n,t_{n+1})}^{(d)}$ is defined in Equation 5 and T_d is the number of terms in d.

Existing similarity based IR systems are not designed to retrieve documents based on the technical comprehensibility and neither do they give any option to the user to specify the technical readability. In contrast, our re-ranking interface provides the choice where the user can specify "Beginner", "Intermediate" or "Advanced". This means whether the user wants to search for conceptually simple, conceptually intermediate or conceptually advanced level technical documents. This does away with several anomalies, for instance, if a user enters a difficult technical jargon, for example, *"ornithology"* but wants technically simple documents related to *"birds"*, or a user enters technically simple term, for example, *"money"* but wants technically advanced documents related to *"finance"*.

IV. Experiments and Results

A. Data Set and Text Preprocessing

Existing standard IR test collections do not fulfill our purpose of evaluation as we need technical readability judgment on each document and current IR test collections have relevance judgments. In order to show the full operational characteristics of our model, we build a large test collection of technical web pages of our own. We chose two popular domains 1) Psychology and, 2) Science. In all, we crawled about 170,000 web pages in Psychology and 300,000 web pages in Science. Enlisting every crawled resource would be too long but we name a few popular resources from where we crawled the web pages: 1) Wikipedia, 2) Psychology.com, 3) PubMed research papers[2], 4) ScienceDaily, 5) ScienceForKids, 6) PhysicsForKids, 7) Kids Wikipedia, and some more related web resources. By crawling web pages from different resources available online we are able to collect technical contents which fit the understanding level and difficulty for diverse backgrounds of people. No term stemming was performed. We prepared two sets of documents, one with stop words[3] kept and another with stop words removed. Removing stop words breaks the natural semantic structure of the document, but this will capture conceptual leaps between the sequences of content words. Moreover, stop words normally do not aid in technical readability of text but they connect the conceptual terrain which lends some meaning to text.

B. Experimental Setup

We evaluate the effectiveness of our model and compare with other state-of-the-art approaches in terms of technical readability prediction and ranking. We also investigate whether our method can perform significantly better in different domains. In our setup we selected some queries from AOL query logs in each domain and used an IR engine to conduct document retrieval. After that, the retrieved documents were re-ranked automatically. First, we searched the query logs for queries containing Science and Psychology jargon. For example, simply searching for the term *"science"* in the AOL query logs retrieves 39026 query/URL pairs. Then we manually sampled out a subset of queries. The queries had two to three terms on average. Moreover they were not ambiguous and were informational in nature. Another criterion for selecting the queries was that a good match can be found

[2]http://www.ncbi.nlm.nih.gov/pmc/tools/ftp/

[3]http://jmlr.csail.mit.edu/papers/volume5/lewis04a/a11-smart-stop-list/english.stop

between the terms present in the query and the documents that we had crawled in our test collection. This would increase the number of relevant documents retrieved by an IR system. We had selected 110 queries in Psychology and 150 in Science. We used Zettair[4] to conduct retrieval and obtained a ranked list using the Okapi BM25 [42] ranking function. We then selected the top-k documents retrieved from the ranked list where k=10 in both Science and Psychology for evaluation purpose. The reason for selecting the top-k documents for evaluation is that we observed that these top-k documents from Zettair system were all relevant to the query and were in mixed order of technical readability as Okapi BM25 is not designed to rank documents based on technical readability. These documents were then re-ranked automatically from conceptually simple to difficult using our proposed models as well as some existing models for comparison. Similar kind of experimental setup and document re-ranking scheme has been adopted in [13], [16]. The reason for re-ranking from conceptually simple to advanced in our experiments is as follows. According to the studies undertaken relating to the behavior of novices and expert web searchers, it has been found that an increasing number of users are searching for information in unfamiliar domains [43]. Hence, most of them will probably look for introductory level documents. A study has also found that domain experts employ complex search strategies such as usage of jargon, complex phrases to successfully retrieve documents based on their technical readability level [44], [45], [46]. Therefore, ranking from conceptually simple to advanced fits most of the users. As stated previously in [13], [16], the authors also ranked documents from introductory to advanced when they tested their model on users possessing average level of knowledge about health care but have compared their method only against readability methods.

We refer "Terrain (Stop)" as our terrain model keeping the stop words intact and "Terrain (No Stop)" as our terrain model with stop words removal. The number of latent concepts in LSI was 200. Normally 150-200 factors have shown to give good performance [31]. The parameter α in Equation 5 was set to 0.5 so that the two components viz. the term difficulty and term cohesion could equally contribute in determining the overall document's conceptual difficulty. The existing unsupervised methods used for our comparative experiments include: 1) Okapi BM25 described in [42], 2) Cosine similarity based measure in the vector space [47], 3) Dirichlet Smoothed query likelihood Language Model [48] with default value for μ provided in Zettair.

We also compared with the popular unsupervised readability scores, namely, ARI: Automated Readability Index [49], Coleman-Liau [50], Flesch Reading Ease formula [51], Fog [52], LIX [53], and SMOG [54]. Our terrain model captures on the semantics of text and does not take into account its syntax. Hence it would be more appropriate to compare with the semantic components of the readability methods. More details about the semantic components of readability methods can be found in [13]. For each readability formula, we obtain a semantic readability score for every document. Then the documents were re-ranked in descending order of readability score.

[4]http://www.seg.rmit.edu.au/zettair/index.html

TABLE I

TECHNICAL READABILITY JUDGMENT GUIDELINES GIVEN TO THE HUMAN JUDGES.

Annotation Guidelines	
The relative technical difficulty of the document that you are currently reading is:	
4	Very low.
3	Reasonably low.
2	Borderline.
1	Reasonably high.
0	Very high.

It is important to mention that rationale behind choosing the readability methods as one of the main comparative methods. One may argue that since these methods are quite old and may not be very effective. However, they are the closest comparative methods to compare with because they are completely unsupervised and do not require any domain-specific ontology as in [13] or some seed set of concepts to initialize their algorithm as in [27]. Hence, it will be more fair to compare with the readability methods. Although we know that ranking functions such as BM25 etc were not designed to handle readability problem, but we compare with these methods because they are purely unsupervised and are widely used. Moreover, the results will support our claim that they cannot rank documents by readability in practice.

C. Evaluation Metric

To obtain a ground truth of the technical readability of the documents for evaluation purpose, human annotators who were undergraduate students having varied background were invited. They had basic knowledge about Science and Psychology. The annotators were fluent in reading English passages. They gave annotations following the guidelines given in Table I. They were also asked to read the article sequentially without skipping any of the terms. In the beginning we acquainted them with the main aim of the study and also showed them some sample documents from our test collection so that they could get an idea about the relative readability levels of documents in the collection. The standard deviations of judgments among the annotators were 1.18 for Science and 1.23 for Psychology.

We evaluate our method using NDCG [55], which is widely used for IR ranking effectiveness measurement. NDCG is well suited for our task because it is defined by an explicit position discount factor and it can leverage the judgments in terms of multiple ordered categories. The NDCG formula is:

$$W(i) = \frac{1}{Z_n} \sum_{i=1}^{n} \frac{2^{r(i)} - 1}{\ln(1+i)} \tag{7}$$

where Z_n is the normalization constant such that a perfect list gets a score of 1; $r(i)$ denotes the rank label (readability label in our case) of the i^{th} document in the ranked list; n is the length of the ranked list. We computed the NDCG for each annotator and aggregated the final NDCG by taking the average.

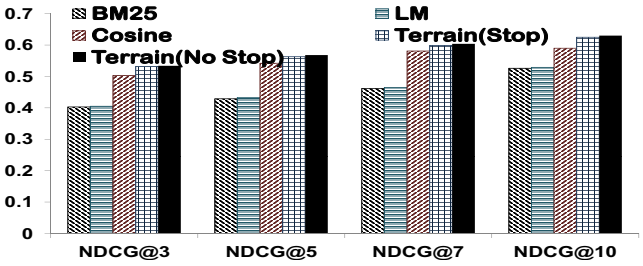

Fig. 2. Results obtained in Psychology domain when compared with standard unsupervised ranking functions. Our results are statistically significant using paired student t-test (p-value < 0.05)

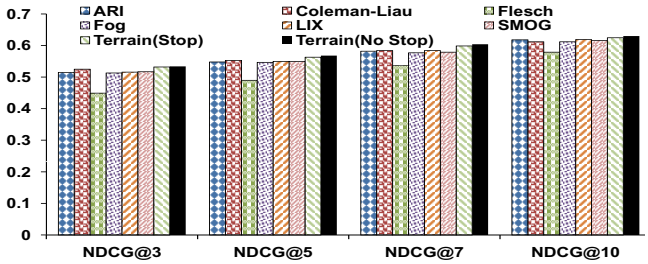

Fig. 3. Results obtained in Psychology domain when compared with standard unsupervised readability methods. Our results are statistically significant using paired student t-test (p-value < 0.05).

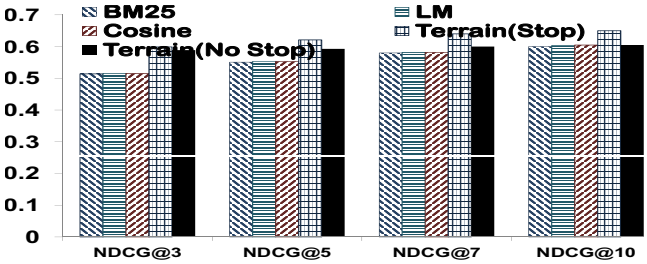

Fig. 4. Results obtained in Science domain when compared with standard unsupervised ranking functions. Our results are statistically significant using paired student t-test (p-value < 0.05).

D. Discussion

Figure 2 shows the results obtained in Psychology domain. Our terrain model has consistently outperformed major standard ranking functions. Although it is obvious that traditional unsupervised ranking functions are not designed to handle technical readability ranking and our results fit that intuition very well. We then compare our model with the traditional readability methods in Figure 3 because they are widely used unsupervised methods. Although there are other supervised methods proposed in literature (discussed in Section II) but they require training data for parameter learning, which is difficult to collect. Hence, we do not compare our method against those methods. Our model has also consistently outperformed traditional readability methods in ranking documents by technical readability. This is because readability methods failed to capture importance of Psychology terms with low syllable counts such as *school, thought, fear* etc.

Figure 4 shows the results where we compare against state-of-the-art ranking functions in Science domain. It again fits the general intuition that even in Science traditional ranking functions cannot rank documents based on technical readability. Figure 5 shows when we compare our model with the readability methods in the Science domain, our model has again outperformed readability methods significantly.

One noticeable observation is that the document set in the Science domain which contained stop words has performed better than the set where we removed stop words. Moreover, in the Psychology domain the performance between the terrain models in the two different document sets is mostly the same. This points to the fact that stop words have some role to play in determining the technical readability of domain-specific documents.

There are certain inherent qualities in our model which makes it more superior than the current state-of-the-art readability methods. First of all, we have addressed some of the major shortcomings present in the readability methods especially handling cohesion and contextual usage of the terms. Secondly, our model can capture the difficulty of the terms even when the terms have low syllable counts.

V. Conclusions and Future Works

We have presented our domain-specific readability model which has performed exceptionally well in two domains. By using a conceptual model to solve the problem, we have addressed several shortcomings inherent in the heuristic readability methods. We view the setup of the terms and their semantic relationships in sequence in the document space as a terrain where we compute the cost of traversal in that document terrain. An advantage that our model has is that it does not need any external domain-specific ontology or knowledge base to unearth the technical terms in a document. We have described to components in our model which are cohesion and difficulty. Cohesion measures semantic relationships between the terms in sequence whereas difficulty measures individual term difficulty in a domain. By maintaining the term order in the document and transiting/hopping from one term to another sequentially, our model captures term cohesion which is only dependent on the neighboring terms. In future, we would study how link structure of the web affects technical readability. We would also study different fields of a web page can be made use to determine the technical difficulty of the documents such as only considering the TITLE fields etc.

References

[1] X. Shen, B. Tan, and C. Zhai, "Implicit user modeling for personalized search," in *Proc. of 14th CIKM*, 2005, pp. 824–831.

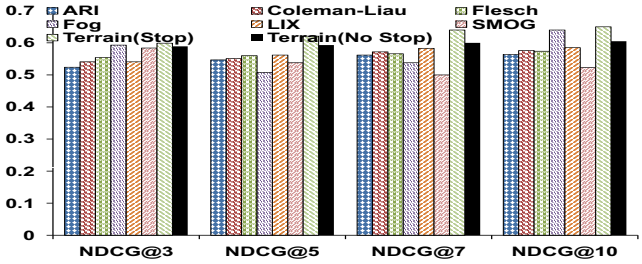

Fig. 5. Results obtained in Science domain when compared with standard unsupervised readability methods. Our results are statistically significant using paired student t-test (p-value < 0.05).

[2] B. Li, A. Ghose, and P. G. Ipeirotis, "Towards a theory model for product search," in *Proc. of 20th WWW*, 2011, pp. 327–336.

[3] A. Kobsa, "Tailoring privacy to users' needs," in *User Modeling 2001*, ser. LNCS, M. Bauer, P. Gmytrasiewicz, and J. Vassileva, Eds. Springer Berlin, 2001, vol. 2109, pp. 301–313.

[4] X. Yan, R. Y. Lau, D. Song, X. Li, and J. Ma, "Toward a semantic granularity model for domain-specific information retrieval," *ACM Trans. Inf. Syst.*, vol. 29, pp. 15:1–15:46, July 2011.

[5] M. W. Berry, S. T. Dumais, and G. W. O'Brien, "Using linear algebra for intelligent information retrieval," *SIAM Rev.*, vol. 37, pp. 573–595, December 1995.

[6] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *JASIST*, vol. 41, no. 6, pp. 391–407, 1990.

[7] M. B. W. Wolfe, M. E. Schreiner, B. Rehder, D. Laham, P. W. Foltz, W. Kintsch, and T. K. Landauer, "Learning from text: Matching readers and texts by latent semantic analysis," *Discourse Processes*, vol. 25(2/3), pp. 309–336, 1998.

[8] A. Broder, "A taxonomy of web search," *SIGIR Forum*, vol. 36, pp. 3–10, September 2002.

[9] G. Kumaran, R. Jones, and O. Madani, "Biasing web search results for topic familiarity," in *Proc. of CIKM*, 2005, pp. 271–272.

[10] J. S. Chall and E. Dale, *Readability revisited: the new Dale-Chall readability formula*, ser. Brookline Books (Cambridge, Mass.), 1995.

[11] J. P. Kincaid, R. P. Fishburne, R. L. Rogers, and B. S. Chissom, "Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel." Tech. Rep., Feb. 1975.

[12] W. H. Dubay, "The principles of readability," *Costa Mesa, CA: Impact Information*, 2004.

[13] X. Yan, D. Song, and X. Li, "Concept-based document readability in domain specific information retrieval," in *Proc. of 15th CIKM*, 2006, pp. 540–549.

[14] K. Collins-Thompson and J. Callan, "Predicting reading difficulty with statistical language models," *J. Am. Soc. Inf. Sci. Technol.*, vol. 56, pp. 1448–1462, Nov. 2005.

[15] B. C. Bruce, A. Rubin, and K. S. Starr, "Why readability formulas fail," *IEEE Trans. on Prof. Comm.*, pp. 50–52, March 1981.

[16] M. Nakatani, A. Jatowt, and K. Tanaka, "Easiest-first search: towards comprehension-based web search," in *Proc. of 18th CIKM*, 2009, pp. 2057–2060.

[17] X. Liu, W. B. Croft, P. Oh, and D. Hart, "Automatic recognition of reading levels from user queries," in *Proc. of 27th SIGIR*, 2004, pp. 548–549.

[18] M. A. K. Halliday and R. Hasan, *Cohesion in English (English Language)*. Longman Pub Group, May 1976.

[19] A. Hagerup-Neilsen, *The role of macrostructures and linguistic connectives in comprehending familiar and unfamiliar written discourse*. Univ. of Minnesota., 1977.

[20] C. Tan, E. Gabrilovich, and B. Pang, "To each his own: personalized content selection based on text comprehensibility," in *Proc. of WSDM*, 2012, pp. 233–242.

[21] K. Collins-Thompson, P. N. Bennett, R. W. White, S. de la Chica, and D. Sontag, "Personalizing web search results by reading level," in *Proc. of 20th CIKM*, 2011, pp. 403–412.

[22] J. Y. Kim, K. Collins-Thompson, P. N. Bennett, and S. T. Dumais, "Characterizing web content, user interests, and search behavior by reading level and topic," in *Proc. of WSDM*, 2012, pp. 213–222.

[23] L. Feng, M. Jansche, M. Huenerfauth, and N. Elhadad, "A comparison of features for automatic readability assessment," in *Proc. of Coling*. Stroudsburg, PA, USA: ACL, 2010, pp. 276–284.

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

[25] T.-Y. Liu, "Learning to rank for information retrieval," *Found. Trends Inf. Retr.*, vol. 3, no. 3, pp. 225–331, Mar. 2009.

[26] M. Bendersky, W. B. Croft, and Y. Diao, "Quality-biased ranking of web documents," in *Proc of WSDM*, 2011, pp. 95–104.

[27] J. Zhao and M.-Y. Kan, "Domain-specific iterative readability computation," in *Proc. of JCDL*, 2010, pp. 205–214.

[28] S. Jameel, W. Lam, C.-m. Au Yeung, and S. Chyan, "An unsupervised ranking method based on a technical difficulty terrain," in *Proc. of CIKM*, 2011, pp. 1989–1992.

[29] S. Jameel, W. Lam, X. Qian, and C.-m. Au Yeung, "An unsupervised technical difficulty ranking model based on conceptual terrain in the latent space," in *Proc. of JCDL*, 2012, pp. 351–352.

[30] J. Bellegarda, "A multispan language modeling framework for large vocabulary speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 5, pp. 456 –467, sep 1998.

[31] S. T. Dumais, "Latent semantic indexing (lsi): Trec-3 report," in *TREC*, 1994, pp. 105–115.

[32] H. Zha, O. Marques, and H. Simon, "Large-scale SVD and subspace-based methods for information retrieval," in *Solving Irregularly Structured Problems in Parallel*, ser. LNCS, A. Ferreira, J. Rolim, H. Simon, and S.-H. Teng, Eds., 1998, vol. 1457, pp. 29–42.

[33] S. Vigna, "Distributed, large-scale latent semantic analysis by index interpolation," in *Proc. of 3rd ICSIS*, ser. InfoScale, 2008, pp. 18:1–18:10.

[34] T. A. Letsche and M. W. Berry, "Large-scale information retrieval with latent semantic indexing," *Information Sciences*, vol. 100, no. 1-4, pp. 105–137, 1997.

[35] E. Bingham and H. Mannila, "Random projection in dimensionality reduction: applications to image and text data," in *Proc. of 7th SIGKDD*, ser. KDD '01, 2001, pp. 245–250.

[36] D. Metzler, "Beyond bags of words: effectively modeling dependence and features in information retrieval," *SIGIR Forum*, vol. 42, no. 1, p. 77, 2008.

[37] P. Freebody and R. C. Anderson, "Effects of vocabulary difficulty, text cohesion, and schema availability on reading comprehension," *Reading Research Quarterly*, vol. 18, no. 3, pp. pp. 277–294, 1983.

[38] W. C. Flick and J. I. Anderson, "Rhetorical difficulty in scientific english: A study in reading comprehension," *TESOL Quarterly*, vol. 14, no. 3, pp. pp. 345–351, 1980.

[39] E. Tarone and G. Yule, *Focus on the language learner: approaches to identifying and meeting the needs of second language learners*, ser. Oxford applied linguistics. Oxford University Press, 1989.

[40] W. Kintsch, "The role of knowledge in discourse comprehension: A construction-integration model," *Psychological Review*, vol. 95, pp. 163–182, 1988.

[41] S. Robertson, "Understanding inverse document frequency: on theoretical arguments for idf," *Journal of Documentation*, vol. 60, no. 5, pp. 503–520, 2004.

[42] S. E. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, and M. Gatford, "Okapi at trec-3," in *TREC*, 1994.

[43] S. K. Bhavnani, "Domain-specific search strategies for the effective retrieval of healthcare and shopping information," in *Proc. of CHI*, 2002, pp. 610–611.

[44] N. Belkin, H. Brooks, and P. Daniels, "Knowledge elicitation using discourse analysis," *Int'l Journal of Man-Machine Studies*, vol. 27, no. 2, pp. 127 – 144, 1987.

[45] H. (Iris) and Xie, "Patterns between interactive intentions and information-seeking strategies," *IP&M*, vol. 38, no. 1, pp. 55 – 77, 2002.

[46] R. W. White, S. T. Dumais, and J. Teevan, "Characterizing the influence of domain expertise on web search behavior," in *Proc. of WSDM*, 2009, pp. 132–141.

[47] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information Processing Management*, vol. 24, no. 5, pp. 513 – 523, 1988.

[48] C. Zhai and J. Lafferty, "A study of smoothing methods for language models applied to information retrieval," *ACM Trans. Inf. Syst.*, vol. 22, no. 2, pp. 179–214, Apr. 2004.

[49] R. Senter and E. Smith, "Automated readability index," *Cincinnati University Ohio*, 1967.

[50] M. Coleman and T. L. Liau, "A computer readability formula designed for machine scoring," *Journal of Applied Psychology*, vol. 60(2), pp. 283–284, 1975.

[51] R. Flesch, "A new readability yardstick." *Journal of Applied Psychology*, vol. 32, no. 3, pp. 221–233, 1948.

[52] R. Gunning, "The fog index after twenty years," *Journal of Business Communication*, vol. 6, no. 2, pp. 3–13, 1969.

[53] J. Anderson, "Lix and rix: Variations on a little-known readability index," *Journal of Reading*, vol. 26, no. 6, pp. pp. 490–496, 1983.

[54] G. H. McLaughlin, "Smog grading: A new readability formula," *Journal Of Reading*, vol. 12, no. 8, pp. 639–646, 1969.

[55] K. Järvelin and J. Kekäläinen, "Cumulated gain-based evaluation of IR techniques," *ACM Trans. Inf. Syst.*, vol. 20, pp. 422–446, October 2002.