N-gram Fragment Sequence Based Unsupervised Domain-Specific Document Readability

Shoaib Jameel Xiaojun Qian Wai Lam
Department of Systems Engineering and Engineering Management,
The Chinese University of Hong Kong.
{msjameel, xjqian, wlam}@se.cuhk.edu.hk

ABSTRACT
Traditional general readability methods tend to underperform in domain-specific document retrieval because they fail to effectively differentiate the reading difficulty of the individual domain-specific terms and the semantic associations between the textual units in a document. On the other hand, recently proposed domain-specific readability methods have relied upon an external knowledge base which may be unavailable in some domains. We develop a novel unsupervised framework for computing domain-specific document readability. Our model does not require an ontology or a knowledge base to capture domain-specific terms in a document. The sequential flow of terms in a document is modeled as a connected sequence of n-gram fragments in the latent concept space. We investigate an automatic sequential n-gram determination scheme that aids in capturing appropriate n-gram fragments which are semantically associated with the document's theme and cohesive with the context. The domain-specific readability cost of a document is computed based on n-gram cohesion and n-gram specificity guided by the latent concepts. The cost can be employed to re-rank the search results generated from an information retrieval (IR) engine. The experimental results demonstrate that our framework achieves significant improvement in ranking documents against the state-of-the-art unsupervised comparative methods.

KEYWORDS: Term Sequence, LSI, IR, Domain-specific Readability, Ranking, Dynamic programming.

Proceedings of COLING 2012: Technical Papers, pages 1309–1326,
COLING 2012, Mumbai, December 2012.

1 Introduction

Readability assessment is an important issue in NLP (Tanaka-Ishii et al., 2010). Traditional general readability methods (Dubay, 2004) have been applied to several problem tasks such as matching books with grade levels (Collins-Thompson and Callan, 2005; Fry, 1969). However, the problem of readability has not been well explored in Information Retrieval (IR) (Kim et al., 2012). Recently researchers have started looking at the problem in IR and several motivations for incorporating readability in an IR system have already been clearly laid out in (Kim et al., 2012; Tan et al., 2012; Nakatani et al., 2009; Yan et al., 2006; Collins-Thompson et al., 2011; Jameel et al., 2011; Zhao and Kan, 2010; Kumaran et al., 2005). What has lacked is a thorough investigation into the problem of reading difficulty in domain-specific IR because traditional unsupervised general readability formulae tend to underperform in domain-specific document retrieval (Yan et al., 2006; Jameel et al., 2011). Domain-specific IR is important because many people are searching for information outside their domain of expertise (Bhavnani, 2002; Yan et al., 2006).

The most affected group who regularly experience difficulties in retrieving documents based on readability are children (Collins-Thompson et al., 2011) and other users who are not domain experts (Yan et al., 2006). Hence, most of them will look for domain-specific documents which they can easily comprehend (Bhavnani, 2002). Domain experts employ complex search strategies such as usage of jargon and complex phrases to successfully retrieve documents based on their reading level (White et al., 2009). Moreover, domain experts know their target destinations such as the ACM Digital Library or Google Scholar using which they can successfully retrieve a document satisfying both relevance and their reading level (White et al., 2009). In contrast, non-domain experts face immense difficulties in formulating a query due to less exposure to the domain-specific terminologies (Vakkari et al., 2003; Paek and Chandrasekar, 2005).

Works which have looked into the problem of domain-specific readability (Yan et al., 2006; Zhao and Kan, 2010) have remained constrained to certain domains only, for instance, the Medical domain because of the required reliance on some knowledge bases to find domain-specific terms in a document. To circumvent this limitation, we propose a novel unsupervised framework for computing domain-specific document readability. The main factor that makes our work superior compared with the existing domain-specific readability methods is that our method does not require an external ontology or lexicon of domain-specific terms. We have previously proposed two terrain models (Jameel et al., 2011, 2012) in Latent Semantic Indexing (LSI) (Berry et al., 1995) to predict the technical difficulty of documents in domain-specific IR using heuristic methods based on conceptual hops between the unigrams and the evaluation is done on one domain only. In this paper, we present an n-gram sequence determination based method which captures appropriate n-gram fragments which are semantically linked with the document automatically while traversing forward following the term sequence in the document. We test the ranking effectiveness of our model on more than one domain. One application of our method is in domain-specific vertical search engines.

Our contributions are as follows: 1) We develop a novel framework to capture suitable n-gram fragments in a domain-specific document by optimizing n-gram fragment sequence connections and taking into account n-gram fragment specificity and cohesion. 2) Our method does not require a domain-specific knowledge base. 3) We conduct extensive experiments on domain-specific document collections in order to show the readability ranking performance.

2 Related Work

Unsupervised heuristic readability methods: Much research has been done in measuring the reading level of text (Qumsiyeh and Ng, 2011). A detailed description about important heuristic readability methods such as Dale-Chall (Dale and Chall, 1948), Automated Readability Index (ARI) (Senter and Smith, 1967), SMOG (McLaughlin, 1969), Coleman-Liau (Coleman and Liau, 1975) etc, can be found in (Dubay, 2004). These methods compute the vocabulary difficulty of a textual discourse. Their readability prediction is based on computing the number of syllables in a term, number of characters etc, which are the surface level features of text. Heuristic readability methods consist of two components linearly combined into a single formula. The components are - syntactic and semantic. These methods have long been in existence and still remain a dominant tool for computing the reading difficulty of traditional documents. In fact, many popular word processing packages use them today. However, readability methods tend to perform poorly on domain-specific texts (Yan et al., 2006) and web pages (Collins-Thompson and Callan, 2005). There are other shortcomings (Bruce et al., 1981) which undermine their importance. In (Nakatani et al., 2009) the authors described an unsupervised method to re-rank the search results of a web search engine in descending order of their comprehensibility using the Japanese Wikipedia but they failed to address the shortcomings in the readability formulae.

Why readability methods underperform on domain-specific documents? Consider a short sentence, "In its simplest form, a star network consists of one central switch, hub or computer, which acts as a conduit to transmit messages." The Flesch reading ease score for this sentence is 62.11, which according to the score is not a difficult sentence. However, the sentence carries a deep technical meaning which requires domain-specific knowledge for proper comprehension. Terms such as "star", "network", and "switch" are domain-specific terms in this example but the readability formula has detected them as easy due to the surface level features.

Domain-Specific Readability Methods: To address the shortcomings inherent in the heuristic readability methods, (Yan et al., 2006) proposed concept based readability ranking method where they have used a domain-specific ontology to capture the domain-specific terms in a document. Their method has a serious drawback in that it requires an ontology for every domain. The authors have only shown the application of their method in one domain. (Kim et al., 2012) described concept readability method in the medical domain. They have used average term and concept familiarity scores from the OAC CHV knowledge base to compute the difficulty of terms and concepts. (Zhao and Kan, 2010) presented domain-specific iterative readability method based on grade levels. Their method is influenced by two popular web link structure based algorithms which are HITS (Kleinberg, 1999) and SALSA (Lempel and Moran, 2001). A limitation of their approach is that they need some seed concepts to initialize their algorithm. This can sometimes be cumbersome as one has to search for a lexicon for every domain. In (Nakatani et al., 2010) the authors used Wikipedia to build a list of some technical terms. In contrast, our proposed framework in this paper does not require an ontology or seed concepts, which can be regarded as a major innovation. The two heuristic terrain models proposed in (Jameel et al., 2011, 2012) computed the technical difficulty of text documents and re-ranked the results obtained from a general purpose IR system. A limitation of the terrain models is that they cannot capture n-grams such as *random access memory* etc. Moreover, they lack a solid theoretical foundation. In this paper, we introduce a principled approach to n-gram fragment determination scheme where we first find the best n-gram sequence in a document automatically. Domain-specific readability cost model is then developed for a document based

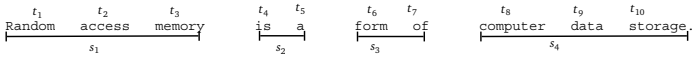

Figure 1: A sentence with four n-gram fragments (s_1, s_2, s_3, s_4). This sample sentence has been taken from one of the Wikipedia documents in our Science test collection. The fragmented sentence has been obtained using Equation 5.

on a new formulation of cohesion and specificity.

Supervised Methods for Readability: Although our proposed framework is completely unsupervised, some supervised methods for computing the reading difficulty of text have been proposed as well (François and Miltsakaki, 2012). Supervised learning approach for readability can be considered as a classification problem. In (Liu et al., 2004), the authors have used Support Vector Machines (SVM) (Vapnik, 1995) for recognizing the reading levels of texts from user queries. They have used syntactic and vocabulary based features to train the classifier. Language modeling has been applied to readability (Collins-Thompson and Callan, 2005) where the authors described a smoothed unigram model for computing the readability of text documents such as web pages. In (Si and Callan, 2001), the authors also used unigram language model to predict readability. Topic familiarity is different from traditional general readability (Kumaran et al., 2005), where the authors studied re-ranking of a search engine result based on familiarity. They also studied the importance of stopwords in their familiarity classifier (FAMCLASS). In (Leroy et al., 2008), classification of health related documents into three levels, namely, Beginner, Intermediate and Advanced is discussed. The authors achieved high classification accuracy using their classifier. In (Schwarm and Ostendorf, 2005; Petersen and Ostendorf, 2009) the authors combined word level features with other textual features. They have used SVM together with several word level features to classify documents based on readability. In (Heilman et al., 2008) the authors introduced a k-Nearest Neighbor classifier based on grammatical features such as sentence length and the patterns of the parse tree. (Bendersky et al., 2011) used several features including readability to improve relevance ranking of the web search results.

Readability is a relative measure (Van Oosten and Hoste, 2011). In order to cater to the results on an individual user basis, recently, methods using query log analysis have been proposed. Search engine query log mining and building individual user profile classifier can also help to solve the problem to some extent as done in (Collins-Thompson et al., 2011; Tan et al., 2012; Kim et al., 2012). But this requires confidential and proprietary query log data with private user session details (Silverstein et al., 1999). Many users might not want their sessions to be recorded or used due to privacy concerns (Jones et al., 2007).

Readability has also been studied in computational linguistics (Leroy and Endicott, 2012). In (Kate et al., 2010) the authors used several linguistic and language model features to build a classifier to predict readability of texts. Language model features were found out to be important to their classifier. (Pitler and Nenkova, 2008) used several textual features in their classifier. Their result shows that word features and average sentence length are strong predictors but the strongest ones are the discourse features. One major limitation of the supervised methods is that one needs a large amount of expensive annotated data (Kanungo and Orr, 2009). Language modeling approaches cannot capture domain-specific concepts in a domain (Zhao and Kan, 2010). In contrast, our method does not need any annotated data.

3 Background and Overview

We tackle the problem of domain-specific readability of text documents. We take as input a document collection of related documents of a domain. We compute the domain-specific readability of each document in the collection. Given a query, a similarity-based IR system retrieves documents. We then re-rank the retrieved top-k results automatically based on readability.

3.1 Overview

We address the problem of domain-specific document readability based on an automatic scheme for finding appropriate n-grams in the Latent Semantic Indexing (LSI) (Berry et al., 1995) latent concept space. As in the previous domain-specific readability approaches, the task lies in an effective capturing of domain-specific terms in the document and their individual specificity scores or scopes (e.g. document scope in (Yan et al., 2006)). In the LSI latent space, n-gram fragments which are central to a document, mainly the domain-specific terms, come close to their document vectors because the semantic fabric of the n-gram fragments inclines best with the technical content of the document (Bellegarda, 2000; Jameel et al., 2011). Common n-grams will not be coherently linked with the document semantics. They can be considered as non-central in that technical storyline (Bellegarda, 2000). N-grams which are semantically similar in meaning will cluster close to each other in the latent space forming a word cloud of semantically related terms (Berry et al., 1995).

We denote the sequence of unigrams in a document d as $(t_1, t_2, \cdots, t_W)$. In Figure 1, an ordered sequence of terms $(t_1, t_2, \cdots, t_{10})$ is shown. Using our proposed methods, the sequence of unigrams can be formed into a sequence comprising of n-gram fragments such as "random access memory" or a connective such as "is a", which are examples of n-gram fragments of order 3 and 2 respectively. An example of a sequence of variable length n-gram fragments is also shown in Figure 1, which is composed of $S = (s_1, s_2, s_3, s_4)$. The sequential flow of terms in a document is modeled as a sequence of n-gram fragments. In this paper, we investigate an automatic sequential n-gram determination scheme that aids in capturing the appropriate n-grams which are semantically associated with the document's theme and cohesive with the context. Cost expended during n-gram formation step can be regarded as a reading difficulty cost of a document. Our proposed n-gram sequence cost optimization linearly combines the effect of both cohesion and specificity, which play a dominant role in determining the domain-specific readability of a document. Some domain-specific readability methods have similar consideration of cohesion and term's domain-specific importance such as (Jameel et al., 2011, 2012; Yan et al., 2006), but they have some serious limitations as mentioned in Section 2.

The first step of our framework is to generate all n-gram fragments (i.e. unigrams, bi-grams, tri-grams etc.) in a document with a predefined maximum value of n, and subsequently we construct a weighted n-gram fragment-document matrix using the product of normalized n-gram frequency and inverse n-gram document frequency (formulae given in (Salton and Buckley, 1988)), where rows are represented by n-gram fragments and columns by documents. Then we perform LSI and obtain a low-dimensional representation of the original vector space. The main computation in LSI is Singular Value Decomposition (SVD) (Golub and Reinsch, 1970). Computing the SVD of a matrix is generally computationally expensive both in space and time (Berry et al., 1995). But with the fast development of better algorithms to compute the SVD, such concerns have been addressed (Wang et al., 2011; Zha et al., 1998).

In (Yan et al., 2006) the authors introduced two components in determining the overall

domain-specific readability of a document which are "document scope" and "document cohesion". The hypothesis is that readability of a document will not only depend on the reading difficulty of the individual terms but also on how the terms in the document are related to one another in the document. Hence a document comprising of many domain-specific terms will be difficult to read and also if the terms are not related to each other (low cohesion) in the same document then the reader will face difficulties in relating different concepts of a domain (Yan et al., 2006). However, the computation of document scope and cohesion in (Yan et al., 2006) is accomplished using an ontology tree which requires an ontology for every domain. Our proposed computation of document scope is different as that in (Yan et al., 2006). We introduce "n-gram specificity" which is able to capture document scope more effectively. As stated earlier, an n-gram fragment which is coherently linked with the technical storyline of a document will be close (Bellegarda, 2000; Jameel et al., 2011, 2012) to the document in the LSI latent space and thus will be more *specific* to that technical storyline. We use the notion of closeness of an n-gram vector to the document vector in the LSI latent space to compute the n-gram specificity.

Psychologists have studied various aspects which lead a reader to comprehend particular piece of discourse (Graesser et al., 1997). An important aspect in text comprehension is cohesion between texts. Cohesion is the property of text in which the units are semantically related to each other and describe one theme. Cohesion is important because the interpretation of one element of text depends on that of another. Texts frequently exhibit varying degrees of cohesion in different sections and hence the start of the text will not be cohesive with the end of the same text (Halliday and Hasan, 1976). Our work is inspired by this observation and we maintain the term order in the document in order to compute cohesion between the n-gram fragments in sequence. Text cohesion has been discussed in (Yan et al., 2006; Ferstl and von Cramon, 2001; Morris and Hirst, 2006; Moe, 1979; Graesser et al., 2004) which establish relation between cohesion and comprehension. Accurate comprehension of technical texts requires accurate identification of the technical meaning of the terms and connections between the terms with the surrounding parts of the text (Freebody and Anderson, 1983). Describing too many difficult non-cohesive terms in sequence will make the reading path of the reader troublesome and thus will affect discourse comprehension (McNamara et al., 1996).

One may argue why we have used a conceptual model instead of the original high-dimensional vector space? An obvious bottleneck in the original vector space is the curse of dimensionality as one has to deal with the space which will be enormous. The high dimensionality would lead to huge computational cost. Moreover, obtaining the n-gram fragment and document semantic relationships directly from the vector space is not possible (Berry et al., 1995) unless additional techniques such as LSI (Deerwester et al., 1990) or the method described in (Yamamoto and Church, 2001) is applied. LSI also handles issues related to data sparsity.

4 Sequential N-gram Connection Model (SNCM)

We present our model that establishes sequential n-gram connections in the document and eventually a cost is expended in order to make such a connection of n-grams in sequence. The aggregated cost expended in the document can be regarded as a document's domain-specific readability cost. If the textual units are not cohesive, then the reader faces cognitive difficulties in comprehension. In addition, if the individual textual units are difficult, then the reader expends cognitive load in figuring out the inherent meaning of the textual unit while reading the text (Yan et al., 2006). This phenomenon can be captured in a cost computation

model described below.

Let s be an n-gram fragment. Let d be the document where this n-gram fragment occurs. Let this fragment be represented as a vector in the LSI latent space as $\vec{s}$ and the document vector as $\vec{d}$. We compute the n-gram specificity, $\vartheta(\vec{s},\vec{d})$, using the following formula:

$$\vartheta(\vec{s},\vec{d}) = \text{cosine\_sim}(\vec{s},\vec{d}) \tag{1}$$

where cosine_sim$(\vec{s},\vec{d})$ is the cosine similarity (formula given in (Bellegarda, 2000)) between the n-gram vector $\vec{s}$ and the document vector $\vec{d}$ in the latent space. An n-gram fragment will obtain a high cosine similarity if it is close to the document vector in the LSI latent space and a low cosine similarity value if it is not close. Therefore, what we expect is that domain-specific n-grams will obtain a high cosine similarity (Jameel et al., 2012) value compared with common/general n-gram fragments. In (Park et al., 2002), they named a domain-specific term extraction scheme as *Degree of Domain-specificity*. However, their method deals with a completely different problem task.

Suppose $T = (t_1, t_2, \cdots, t_W)$ is the term sequence and $S = (s_1, s_2, \cdots, s_K)$ is one particular n-gram fragmented sequence of T, W is the total number of terms in the document d, K is the number of n-grams in S. We denote n-gram cohesion, $\eta(\vec{s_i}, \vec{s_{i+1}})$, between the two n-grams s_i and s_{i+1} at positions i and $i+1$ in sequence whose vectors are represented as $\vec{s_i}$ and $\vec{s_{i+1}}$ respectively for a particular document as:

$$\eta(\vec{s_i}, \vec{s_{i+1}}) = \text{cosine\_sim}(\vec{s_i}, \vec{s_{i+1}}) \tag{2}$$

When the cosine similarity between the two consecutive n-grams is high, the n-gram fragments are cohesive. The reason is that in the latent concept space n-gram fragments which appear under similar storylines and similar semantic meaning will cluster close to each other (Berry et al., 1995; Jameel et al., 2011, 2012). Hence the closer they are, the more semantically related they tend to become. A reader will be able to semantically relate those n-gram fragments (which are cohesive) easily (Halliday and Hasan, 1976) and will comprehend a piece of textual discourse well. A document tends to be semantically readable if the constituent terms are simple (i.e. low specificity values) with reference to the document vector in the LSI latent space. Therefore, we hypothesize that specificity values will be directly proportional to the document's overall domain-specific reading difficulty. In fact, in order to compute document generality and readability, (Yan et al., 2006, 2011) have made a similar hypothesis and have evaluated their hypothesis through experiments. We also hypothesize that cohesion is inversely related to the document domain-specific readability. Again, this assumption is in line with the assumptions made in (Yan et al., 2006, 2011).

4.1 Our First Model (SNCM1)

Our framework works towards determining a least cost n-gram connected sequence in the document where at each forward transition sequential n-gram cohesion is minimized. The sequence of terms consisting of variable n-gram fragments is considered while traversing forward. For a particular document, suppose the term sequence T and its n-gram fragmented sequence S are defined in a similar manner as above. The cost of the n-gram fragment sequence S, $C_1^{(d)}(S)$, can be written as:

$$C_1^{(d)}(S) = \sum_{k=1}^{K} \left(\frac{1}{\eta(\vec{s_{k-1}}, \vec{s_k}) + 1} \right) \tag{3}$$

1 is added in the denominator to handle the cases where the n-gram vectors are orthogonal to each other. Our goal is to minimize this cost, $C_1^{(d)}(S)$. The rationale for such minimization

formulation is to fit the most cohesive n-gram fragment in sequence which matches with the sequential storyline of the document. We achieve this using the following optimization scheme given in Equation 4.

$$\min_{S} C_1^{(d)}(S) \tag{4}$$

This scheme ensures that an n-gram will be the least cost match at that position if it cohesively fits in that sequential discourse. We now describe a dynamic programming method to find the optimal cost. We define $C_1^{(d)}(T_i)$ as the optimal cost from the beginning until the term t_i in the document. Since the accumulated cost is the sum of the local costs, it can be decomposed in the same way as its predecessors and the local cost accumulated with the n-gram itself. To obtain the optimal path cost, we have to select the predecessor with the minimum total cost. Another issue is that we need to set a maximum bound for the number of terms in an n-gram. In principle, this bound could be set to any number m. Let $\vec{S_X}$ be a unigram composed of t_i, $\vec{S_Y}$ be a bigram composed of (t_{i-1}, t_i) and $\vec{S_Z}$ be an m-gram composed of $(t_{i-m+1}, \cdots, t_i)$. $\vec{S_{X-1}}$, $\vec{S_{Y-1}}$ and $\vec{S_{Z-1}}$ represent the particular n-gram (where n may be from 1 to m) in the optimal sequential path that appears just before $\vec{S_X}$, $\vec{S_Y}$ and $\vec{S_Z}$ respectively. The optimal cost for all the terms from t_1 until position t_i, (denoted as $C_1^{(d)}(T_i)$) can be written as:

$$\begin{aligned} C_1^{(d)}(T_i) = \text{minimum}\Big(& C_1^{(d)}(T_{i-1}) + \frac{1}{\eta(\vec{S_{X-1}}, \vec{S_X}) + 1}, \\ & C_1^{(d)}(T_{i-2}) + \frac{1}{\eta(\vec{S_{Y-1}}, \vec{S_Y}) + 1}, \\ & \cdots, \\ & \cdots, \\ & C_1^{(d)}(T_{i-m}) + \frac{1}{\eta(\vec{S_{Z-1}}, \vec{S_Z}) + 1}\Big) \end{aligned} \tag{5}$$

The final reading difficulty of a document will not only depend on cohesion but also specificity. Therefore, to compute the final readability cost of a text document, we linearly combine specificity values of the n-grams formed during sequential linear n-gram determination scheme using Equation 5. The overall document domain-specific readability cost, $E_1^{(d)}$, is given in Equation 6 where α ($0 \leq \alpha \leq 1$) is a parameter controlling the relative contribution of cohesion and specificity. A higher cost indicates that the document is difficult to read and a low cost is indicative of the ease in reading the document. We shall use the cost values to re-rank the search results obtained from a general purpose IR system.

$$E_1^{(d)} = \frac{\alpha C_1^{(d)}(T_W) + (1-\alpha)\sum_{i=1}^{K} \vartheta(\vec{s_i}, \vec{d})}{W} \tag{6}$$

where W is the total number of terms in the document and it removes the document length bias. Note that W in the denominator is more suitable than K because the reading difficulty of a document is not dependent on the number of n-gram fragments formed.

4.2 An Extended Model: SNCM2

We now modify our previous model in Equation 3 further and extend to the case where we combine the effect of both cohesion and specificity. In this model, we linearly combine the effect of specificity along with cohesion during n-gram fragment determination phase. We design the cost, $C_2^{(d)}(S)$, of an n-gram fragment sequence formation S as:

$$C_2^{(d)}(S) = \sum_{k=1}^{K} \left(\beta\, \vartheta(\vec{s_k}, \vec{d}) + (1-\beta) \frac{1}{\eta(\vec{s_{k-1}}, \vec{s_k}) + 1} \right) \tag{7}$$

β $(0 \leq \beta \leq 1)$ is a parameter controlling the relative weights of the two components. Our goal is to minimize the total cost $C_2^{(d)}(S)$ as follows:

$$\min_{S} C_2^{(d)}(S) \tag{8}$$

We can apply similar dynamic programming methodology. Let the optimal cost for all the terms from t_1 until position t_i be $C_2^{(d)}(T_i)$.

$$\begin{aligned} C_2^{(d)}(T_i) = \text{minimum} \\ \Bigg(& C_2^{(d)}(T_{i-1}) + \beta\, \vartheta(\vec{S_X}, \vec{d}) + (1-\beta) \frac{1}{\eta(\vec{S_{X-1}}, \vec{S_X} + 1)}, \\ & C_2^{(d)}(T_{i-2}) + \beta\, \vartheta(\vec{S_Y}, \vec{d}) + (1-\beta) \frac{1}{\eta(\vec{S_{Y-1}}, \vec{S_Y} + 1)}, \\ & \cdots, \\ & \cdots, \\ & C_2^{(d)}(T_{i-m}) + \beta\, \vartheta(\vec{S_Z}, \vec{d}) + (1-\beta) \frac{1}{\eta(\vec{S_{Z-1}}, \vec{S_Z} + 1)} \Bigg) \end{aligned} \tag{9}$$

The overall document domain-specific readability cost $E_2^{(d)}$ is given in Equation 10. We rank the documents based on an optimal cost obtained at the end of the n-gram sequence formation.

$$E_2^{(d)} = \frac{C_2^{(d)}(T_W)}{W} \tag{10}$$

Intuitive Justification: Specificity helps us in finding an n-gram fragment which matches with the technical storyline of the entire document. Cohesion on the other hand helps in finding the best linked n-gram fragments in the sequential discourse based on the context. In Equation 9 our objective is to find a least cost n-gram fragment that is a best match in the n-gram sequence which considers both components, namely, cohesion and specificity simultaneously in the document. The intuition behind adopting this strategy is to find n-gram fragments in a document which are semantically linked with the document's thematic structure and are cohesive with the context. However, in Equation 5 we are minimizing only cohesion between the n-gram fragments in sequence. This strategy may help us find cohesive n-grams with the context but the n-grams may not be thematically linked with the technical storyline of the document because of the absence of specificity during n-gram sequence determination phase. A closer look at the two variants of our model paints some interesting pictures. When $\beta = 0$ and $\alpha = 1$ the two variants (`SNCM1` and `SNCM2`) behave equivalently. Note that the cost expended to fit a specific n-gram fragment will be more in comparison to an n-gram fragment which is common/general. The longer the contextual history m, the better n-gram fragment prediction will be. However, longer contextual histories shall bring about additional computational burden especially for large datasets as ours.

5 Empirical Evaluation

5.0.1 Experimental Setup

Testbed Data: Current IR evaluation test sets such as TREC, and CLEF cannot be used in our experiments due to lack of readability annotations. Currently, they only contain relevance

judgments. So we chose two popular domains 1) Psychology, and 2) Science and subsequently we crawled a large number of web pages in these domains. We enlist some of the important resources from where we crawled the majority of the web pages as enlisting every crawled resource would be too long. Psychology web pages were crawled from: 1) Wikipedia, 2) Psychology.com, 3) Psychology Today 4) Simple English Wikipedia. Science web pages were crawled from: 1) ScienceDaily, 2) ScienceForKids, 4) Simple English Wikipedia 5) Wikipedia, 6) About.com and some more related web resources. We also included Computer Science documents in the Science domain. The reason for choosing these online resources was mainly due to their popularity and high quality content. By crawling web pages from different resources available online we are able to collect domain-specific documents which match diverse genres and audience. In all, our test collection includes 170,000 documents in Psychology with 154,512 n-grams in the vocabulary, and 300,000 documents in Science consisting of 490,770 n-grams in the vocabulary. No term stemming was performed as we wish to keep the original words. In fact, traditional unsupervised readability methods do not consider stemmed terms. We prepared two sets of document collection, one with stopwords[1] kept and another with stopwords removed. The objective is to study the role of stopwords in domain-specific readability (Refer Section 2). Note that we conduct experiments in each domain separately.

We used Zettair[2] to conduct document retrieval and obtained a ranked list using Okapi BM25 (Robertson et al., 1996) ranking function. BM25 retrieves documents based on relevance and the retrieved list contained a mix of easy readable and difficult to read documents. We then selected top-k documents retrieved from the ranked list where $k = 10$ for evaluation purpose. Selecting a higher value of k would lead to a huge cognitive load on the human subjects (which we describe later in the text) during annotation. Therefore, we keep this number as low as possible. Also, a previous study has found that users generally look at the first page of the search results containing the top ten documents (Silverstein et al., 1999).

Domain-specific information needs: Topics are queries posed to an IR system. We strictly followed topic development guidelines laid out in "INEX 2009 Topic Development Guidelines"[3]. We had asked two undergraduate students possessing beginner level knowledge in both Science and Psychology to generate domain-specific topics which represent real information needs. They generated 110 topics in Psychology and 150 topics in Science. We enlist some sample information needs in two domains here: Science: 1) "x-ray machine", 2) "acid and alkali" 3) "why the sky is blue", Psychology: 1) "depression", 2) "bad dreams", 3) "school of Psychology".

Annotations and metrics: To obtain the ground truth of domain-specific readability of the documents for evaluation purpose, two human annotators who were undergraduate students having varied background were invited. They had basic knowledge about Science and Psychology. They were asked to rate the documents based on relative domain-specific readability judgment of the documents. For the judgment, the selection was among "very low domain-specific readability" (i.e. difficult to read), "reasonably low domain-specific readability", "average domain-specific readability", "reasonably high domain-specific readability" and "very high domain-specific readability". These options were further translated to integer gains ranging from 4 to 0. A simple readable document obtained a score of 4 whereas the most difficult obtained a score of 0. In the beginning we acquainted them with the main aim of the study

[1] http://jmlr.csail.mit.edu/papers/volume5/lewis04a/a11-smart-stop-list/english.stop
[2] http://www.seg.rmit.edu.au/zettair/
[3] http://www.inex.otago.ac.nz/tracks/adhoc/gtd.asp

and showed them some sample documents from our test collection so that they could get an idea about the relative difficulty levels of the documents in the collection. Overall, the annotators annotated 983 documents in Psychology and 1442 documents in Science. In order to ascertain whether the manual annotation that we collected was feasible and reproducible, we assessed the inter-annotator agreement by computing the Cohen's Kappa coefficient. We found that there was an acceptable agreement between the annotators (approximately 0.8) in both Psychology and Science domains.

The evaluation metric is NDCG (same formula as in (Cai et al., 2011)) which is widely used for IR ranking effectiveness measurement. We computed the NDCG@i for each annotator and aggregated the final NDCG by taking the average. 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. NDCG@i scores will directly correlate with the readability annotation of the documents given by humans. Such scores can measure the quality of difficulty ranking of documents based on readability judgments provided by humans. If NDCG is high, it means the ranking function correlates better with the human judgments.

Result re-ranking scheme: We automatically re-rank the search results obtained from an IR system from *simple* to *difficult* readable documents using our proposed model as well as comparative methods. The reason is that domain experts normally employ complex search strategies to successfully retrieve documents based on their reading level (refer Section 2). They can find their material of interest easily but non-experts face difficulty in locating their content as they have to sift through the ranked list carefully to locate a document which can match their domain-specific reading level. In addition, previous related approaches have also followed similar re-ranking scheme (i.e. re-ranking from simple to advanced without integrating the readability scores in the initial retrieval score) such as (Yan et al., 2006; Nakatani et al., 2009; Kumaran et al., 2005) and in (Yan et al., 2011), the authors re-rank the top-k documents obtained from a baseline IR system based on decreasing specificity.

Comparative methods: We chose popular unsupervised general readability methods as our comparative models. They are ARI: Automated Readability Index, Coleman-Liau (denoted as C-L in the tables in our results), Flesch Reading Ease formula, Fog, LIX and SMOG. More details about these readability methods can be found in (Dubay, 2004). Our model does not have a syntactic component. Hence, it would be more appropriate to compare with the semantic components of the general readability methods (similar scheme also adopted in (Yan et al., 2006; Collins-Thompson and Callan, 2005). More details about the semantic components can be found in (Yan et al., 2006; Collins-Thompson and Callan, 2005; Dubay, 2004). For each readability formula; it computes a readability score for every document. Then the documents are re-ranked in descending order of the readability score. In addition, we also chose some recent unsupervised comparative methods described in (Collins-Thompson and Callan, 2005) such as Mean Log Frequency (denoted as `MLF`) and %UNK. We also compare our method with `CHM` described in Jameel et al., (Jameel et al., 2011). We denote their method as `CHM`.

In addition, we compare our method by manually collecting an extensive list of domain-specific concepts from online resources. The collection process was indeed cumbersome and time consuming but it will help us evaluate our method by mimicking the working principle of the recently proposed domain-specific readability methods which rely on some external knowledge-bases. Overall we collected about 900 domain-specific concepts in Science and about 600 in Psychology. In this method, we count how many times domain-specific terms occur in the

(a) Psychology

	NDCG@3	NDCG@5	NDCG@7	NDCG@10
ARI	0.515	0.548	0.582	0.618
C-L	0.525	0.553	0.584	0.612
Flesch	0.449	0.490	0.537	0.579
Fog	0.513	0.547	0.577	0.612
LIX	0.516	0.550	0.584	0.619
SMOG	0.517	0.550	0.579	0.616
CHM	0.465	0.456	0.473	0.482
Counts	0.551	0.575	0.603	0.649
MLF	0.530	0.554	0.581	0.631
%UNK	0.558	0.585	0.611	0.653
SNCM1	0.537	0.571	0.602	0.651
SNCM2	**0.581***	**0.607***	**0.635***	**0.680***

(b) Science

	NDCG@3	NDCG@5	NDCG@7	NDCG@10
ARI	0.524	0.547	0.562	0.564
C-L	0.541	0.551	0.572	0.576
Flesch	0.554	0.560	0.566	0.574
Fog	0.593	0.508	0.538	0.640
LIX	0.541	0.562	0.583	0.585
SMOG	0.584	0.538	0.500	0.523
CHM	0.400	0.406	0.407	0.412
Counts	0.595	0.563	0.564	0.627
MLF	0.557	0.584	0.611	0.657
%UNK	0.562	0.590	0.619	0.660
SNCM1	**0.617***	**0.645***	**0.672***	**0.713***
SNCM2	0.602*	0.625*	0.650*	0.702*

Table 1: Comparison of SNCM variants when $\alpha = \beta = 0.5$ against the comparative methods in both domains. * denotes statistically significant results for all comparisons according to paired t-test ($p < 0.05$).

Doc1	Specificity	Syl	Doc2	Specificity	Syl
earth science	0.74	2	**cancer**	0.71	2
earth	0.78	1	in-spite	0.12	2
any	0.09	1	**lung cancer**	0.71	3
mapped	0.51	2	management	0.18	4

Table 2: N-gram specificity values obtained from two separate documents in our collection using Equation 1. We compare the specificity values with the number of syllables in the n-gram (denoted as Syl).

document with respect to the list contained in the lexicon and then divide by the number of words in the document to remove document length bias. We name this comparative method as Counts.

We had set the value for m in Equations 5 and 9 to 3 in our experiments. This could help capture tri-gram fragments which we believe is a suitable number for large datasets. We used MATLAB to compute SVD of the matrix using the "svds" function. The number of latent concepts in LSI were 200 as previous studies have found that 150-200 dimensions give optimal performance (Dumais, 1995). Our term weighting scheme was the product of normalized n-gram count and inverse n-gram document frequency (formulae given in (Salton and Buckley, 1988)). Our main models are SNCM1 and SNCM2 with stopwords kept intact. Our objective is to study the performance of our proposed variants by keeping the entire documents sequence intact without removal of any of the features as term order plays an important role in our model. Moreover, general traditional readability methods also work on original texts. We had set the value of $\alpha = \beta = 0.5$ in our experiments, which means that both the components have equal weights in determining the final readability ranking order.

5.0.2 Results and Analysis

To enlighten the reader more about the specificity values, we show some specificity values obtained from our experimental dataset in Table 2. The technical storyline of $Doc1$ revolves around *Earth Science* and $Doc2$ deals with *Lung Cancer*. Domain-specific terms which we indicate in bold text have obtained higher specificity values compared to common n-gram fragments. We can also observe that the domain-specific terms appear simple in terms of the number of syllables (denoted as Syl). Such n-grams will appear common to any readability formula which relies on the number of syllables for text readability prediction. Note that for a readability formula, if the number of syllables is more, the word is difficult.

(a) Psychology

Method Name	Queries Improved		Average Improvement	
	SNCM1	SNCM2	SNCM1	SNCM2
ARI	53	59	17.56%	18.06%
C-L	61	61	22.84%	22.86%
Flesch	65	65	25.66%	25.66%
Fog	68	65	20.02%	17.12%
LIX	60	62	22.05%	24.03%
SMOG	58	60	23%	23.08%
CHM	86	88	36%	38%
Counts	29	40	1.02%	12.05%
MLF	49	60	2.01%	20.76%
%UNK	3	32	0	9.34%

(b) Science

Method Name	Queries Improved		Average Improvement	
	SNCM1	SNCM2	SNCM1	SNCM2
ARI	95	95	22.34%	22.01%
C-L	90	91	20.12%	20.36%
Flesch	92	92	21.56%	21.50%
Fog	80	80	17.90%	17.90%
LIX	90	90	20.19%	20.13%
SMOG	92	92	25.56%	26%
CHM	121	119	32%	29.99%
Counts	82	79	19.76%	17.55%
MLF	83	75	21.45%	19.23%
%UNK	77	69	17.55%	16.53%

Table 3: Performance comparison based on queries for SNCM1 and SNCM2.

NDCG@i	α=0	α=0.2	α=0.5	α=0.8	α=1
@3	0.506	0.524	0.537	0.566	0.573
@5	0.545	0.586	0.571	0.583	0.599
@7	0.579	0.605	0.602	0.630	0.627
@10	0.631	0.623	0.651	0.547	0.672

Table 4: Varying α for SNCM1 in Psychology with stopwords.

NDCG@i	α=0	α=0.2	α=0.5	α=0.8	α=1
@3	0.496	0.499	0.498	0.537	0.567
@5	0.534	0.525	0.545	0.577	0.587
@7	0.570	0.571	0.574	0.598	0.611
@10	0.624	0.631	0.666	0.640	0.658

Table 5: Varying α for SNCM1 in Psychology without stopwords.

NDCG@i	α=0	α=0.2	α=0.5	α=0.8	α=1
@3	0.603	0.604	0.617	0.545	0.599
@5	0.631	0.633	0.645	0.630	0.622
@7	0.657	0.646	0.672	0.639	0.647
@10	0.697	0.698	0.713	0.710	0.700

Table 6: Varying α for SNCM1 in Science with stopwords.

The main result of the document ranking performance is given in Table 1. In the Psychology domain, SNCM2 has performed better than other comparative methods whereas in the Science domain we notice that SNCM1 has fared better than all other models. We performed a paired t-test between the SNCM variants and the comparative methods. We have obtained statistically significant improvement with ($p < 0.05$) for all comparisons. Counts did not perform well in both domains. An obvious reason is that this method demands a longer list of technical terms, which is extremely time consuming to obtain. Readability methods have failed to give optimal ranking performance. It is because they fail to capture the inherent semantics of text which our method can effectively capture. CHM performed rather poorly in both the domains. One reason could be due to the weak model and incorporation of non-linearity using a heuristic approach. %UNK has shown some good performance. One reason is due to the use of an elaborate list of words (over 3000). In Table 3 we present results based on the improvement we have obtained on query basis. Results show that our models have obtained tangible improvement against the comparative methods.

Some interesting conclusions can be derived from the results in Tables 4,5,6,7,8,9,10 and 11. These results highlight the role of the two components, namely, cohesion and specificity in influencing the overall ranking of the results. Tables 4,5,8 and 9 show the effect of varying α and β in the Psychology domain. Our discussion will mainly focus on the results when $\alpha = 0$, $\alpha = 1$ and $\beta = 0$, $\beta = 1$ because these values portray the contribution that the two components, namely, cohesion and specificity individually make in the overall ranking of the search results. We notice in the Psychology domain that ranking of the search results is significantly dominated by cohesion (note the values close to $\alpha = 1$ and $\beta = 0$). This observation can be reasoned out from the usage of terms across the documents in the collection. We noticed in the Psychology corpus that the documents are more general than the Science documents. Science documents contain relatively more domain-specific terms than Psychology documents. Thus the contribution of specificity will be more uniform across documents in Psychology than in Science. Hence usage of terminologies will be almost at the same level.

We obtained some interesting results in the Science domain as well. In Tables 6 and 7, we note

that the gap in the results when $\alpha = 0$ and $\alpha = 1$ is not very wide. This means that both the components have approximately equal role in affecting the final ranking of the search results. However an interesting conclusion is that cohesion has a slightly more dominant effect than specificity but the importance of specificity cannot be completely disregarded (we conclude this when $\alpha = \beta = 0.5$). Even for SNCM2 in Tables 10 and 11 the observations remain the same where we note that both components have almost equal role in affecting the overall ranking of the results.

NDCG@i	α=0	α=0.2	α=0.5	α=0.8	α=1
@3	0.607	0.609	0.598	0.590	0.590
@5	0.633	0.633	0.620	0.606	0.617
@7	0.658	0.659	0.650	0.637	0.642
@10	0.699	0.701	0.698	0.691	0.693

Table 7: Varying α for SNCM1 in Science without stopwords.

NDCG@i	β=0	β=0.2	β=0.5	β=0.8	β=1
@3	0.573	0.576	0.581	0.573	0.509
@5	0.599	0.602	0.607	0.598	0.548
@7	0.627	0.630	0.635	0.630	0.582
@10	0.672	0.675	0.680	0.675	0.634

Table 8: Varying β for SNCM2 in Psychology with stopwords.

NDCG@i	β=0	β=0.2	β=0.5	β=0.8	β=1
@3	0.567	0.565	0.573	0.574	0.503
@5	0.587	0.585	0.591	0.592	0.542
@7	0.611	0.610	0.615	0.615	0.578
@10	0.658	0.656	0.660	0.661	0.630

Table 9: Varying β for SNCM2 in Psychology without stopwords.

NDCG@i	β=0	β=0.2	β=0.5	β=0.8	β=1
@3	0.599	0.602	0.602	0.603	0.618
@5	0.622	0.625	0.625	0.628	0.646
@7	0.647	0.649	0.650	0.651	0.670
@10	0.700	0.702	0.702	0.704	0.719

Table 10: Varying β for SNCM2 in Science with stopwords.

NDCG@i	β=0	β=0.2	β=0.5	β=0.8	β=1
@3	0.590	0.590	0.594	0.587	0.620
@5	0.617	0.617	0.620	0.615	0.645
@7	0.642	0.642	0.644	0.641	0.670
@10	0.693	0.695	0.697	0.694	0.717

Table 11: Varying β for SNCM2 in Science without stopwords.

We can now infer that cohesion has a more deep seated role compared to specificity in the two domains but specificity cannot be completely disregarded. We also studied the behavior of SNCM variants along with the role of stopwords in the two domains. From the results we note the stopwords have an important role to play in influencing ranking. This stands consistent with the prior findings discussed in Section 2.

6 Conclusions and Future Work

We have presented our SNCM models where we form a fragmented n-gram sequence in a document. We find a least cost path in the n-gram sequence. The cost reflects the domain-specific readability of a text document. We have shown that general readability methods and other state-of-the-art unsupervised methods are not effective to determine the readability of a text document. Experiments in two domains show the superiority of our proposed models. Our proposed approach is more scalable than recently proposed domain-specific readability methods because we do not use any external domain-specific ontology to capture domain-specific terms.

In the future, we would study how the hyperlink structure of the web can aid in determining the reading difficulty of text documents. The hypothesis is that a general web page would link with other general web pages (Akamatsu et al., 2011) as well. We would also explore other features which could help improve readability ranking performance such as a web page layout and content such as fonts, title fields, line and paragraph breaks, etc.

7 Acknowledgements

The work described in this paper is substantially supported by grants from the Research Grant Council of the Hong Kong Special Administrative Region, China (Project Code: CUHK413510) and the Direct Grant of the Faculty of Engineering, CUHK (Project Codes: 2050476 and 2050522). This work is also affiliated with the CUHK MoE-Microsoft Key Laboratory of Human-centric Computing and Interface Technologies.

References

Akamatsu, K., Pattanasri, N., Jatowt, A., and Tanaka, K. (2011). Measuring comprehensibility of web pages based on link analysis. In *Proc. of WI-IAT*, pages 40–46.

Bellegarda, J. (2000). Large vocabulary speech recognition with multispan statistical language models. *IEEE Transactions on Speech and Audio Processing*, 8(1):76 –84.

Bendersky, M., Croft, W. B., and Diao, Y. (2011). Quality-biased ranking of web documents. In *Proc. of WSDM*, pages 95–104.

Berry, M. W., Dumais, S. T., and O'Brien, G. W. (1995). Using linear algebra for intelligent information retrieval. *SIAM Rev.*, 37(4):573–595.

Bhavnani, S. K. (2002). Domain-specific search strategies for the effective retrieval of healthcare and shopping information. In *Human Factors in Computing Systems*, pages 610–611.

Bruce, B., Rubin, A., and Starr, K. S. (1981). Why readability formulas fail. *IEEE Transactions on Professional Communication*, pages 50–52.

Cai, P., Gao, W., Zhou, A., and Wong, K.-F. (2011). Relevant knowledge helps in choosing right teacher: active query selection for ranking adaptation. In *Proc. of SIGIR*, pages 115–124.

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

Collins-Thompson, K., Bennett, P. N., White, R. W., de la Chica, S., and Sontag, D. (2011). Personalizing web search results by reading level. In *Proc. of CIKM*, pages 403–412.

Collins-Thompson, K. and Callan, J. (2005). Predicting reading difficulty with statistical language models. *J. Am. Soc. Inf. Sci. Technol.*, 56(13):1448–1462.

Dale, E. and Chall, J. S. (1948). A formula for predicting readability: Instructions. *Educational Research Bulletin*, 27(2):pp. 37–54.

Deerwester, S., Dumais, S. T., Landauer, T. K., Furnas, G. W., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the Society for Information Science*, 41(6):391–407.

Dubay, W. H. (2004). The principles of readability. *Costa Mesa, CA: Impact Information*.

Dumais, S. T. (1995). Latent semantic indexing (LSI): TREC-3 report. In *Overview of the Third Text REtrieval Conference*, pages 219–230.

Ferstl, E. C. and von Cramon, D. (2001). The role of coherence and cohesion in text comprehension: an event-related fmri study. *Cognitive Brain Research*, 11(3):325 – 340.

François, T. and Miltsakaki, E. (2012). Do NLP and machine learning improve traditional readability formulas? In *Proc. of the First Workshop on Predicting and Improving Text Readability for target reader populations*, pages 49–57, Montréal, Canada.

Freebody, P. and Anderson, R. C. (1983). Effects of vocabulary difficulty, text cohesion, and schema availability on reading comprehension. *Reading Research Quarterly*, 18(3):pp. 277–294.

Fry, E. B. (1969). The readability graph validated at primary levels. *The Reading Teacher*, 22(6):pp. 534–538.

Golub, G. and Reinsch, C. (1970). Singular value decomposition and least squares solutions. *Numerische Mathematik*, 14:403–420.

Graesser, A., McNamara, D., Louwerse, M., and Cai, Z. (2004). Coh-metrix: Analysis of text on cohesion and language. *Behavior Research Methods*, 36:193–202.

Graesser, A. C., Millis, K. K., and Zwaan, R. A. (1997). Discourse comprehension. *Annual Review of Psychology*, 48(1):163–189.

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

Heilman, M., Collins-Thompson, K., and Eskenazi, M. (2008). An analysis of statistical models and features for reading difficulty prediction. In *Proc. of EANL*, pages 71–79.

Jameel, S., Lam, W., Au Yeung, C.-m., and Chyan, S. (2011). An unsupervised ranking method based on a technical difficulty terrain. In *Proc. of CIKM*, pages 1989–1992.

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

Jones, R., Kumar, R., Pang, B., and Tomkins, A. (2007). "I know what you did last summer": query logs and user privacy. In *Proc. of CIKM*, pages 909–914.

Kanungo, T. and Orr, D. (2009). Predicting the readability of short web summaries. In *Proc. of WSDM*, pages 202–211.

Kate, R. J., Luo, X., Patwardhan, S., Franz, M., Florian, R., Mooney, R. J., Roukos, S., and Welty, C. (2010). Learning to predict readability using diverse linguistic features. In *Proc. of COLING*, pages 546–554.

Kim, J. Y., Collins-Thompson, K., Bennett, P. N., and Dumais, S. T. (2012). Characterizing web content, user interests, and search behavior by reading level and topic. In *Proc. of WSDM*, pages 213–222.

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

Kumaran, G., Jones, R., and Madani, O. (2005). Biasing web search results for topic familiarity. In *Proc. of CIKM*, pages 271–272.

Lempel, R. and Moran, S. (2001). SALSA: The Stochastic approach for Link-structure Analysis. *ACM Trans. Inf. Syst.*, 19(2):131–160.

Leroy, G. and Endicott, J. E. (2012). Combining NLP with evidence-based methods to find text metrics related to perceived and actual text difficulty. In *Proc. of IHI*, pages 749–754.

Leroy, G., Miller, T., Rosemblat, G., and Browne, A. (2008). A balanced approach to health information evaluation: A vocabulary-based naive bayes classifier and readability formulas. *J. Am. Soc. Inf. Sci. Technol.*, 59(9):1409–1419.

Liu, X., Croft, W. B., Oh, P., and Hart, D. (2004). Automatic recognition of reading levels from user queries. In *Proc. of SIGIR*, pages 548–549.

McLaughlin, G. H. (1969). SMOG grading: A new readability formula. *Journal Of Reading*, 12(8):639–646.

McNamara, D. S., Kintsch, E., Songer, N. B., and Kintsch, W. (1996). Are good texts always better? interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction*, 14(1):pp. 1–43.

Moe, A. J. (1979). Cohesion, coherence, and the comprehension of text. *Journal of Reading*, 23(1):pp. 16–20.

Morris, J. and Hirst, G. (2006). The subjectivity of lexical cohesion in text. In Shanahan, J., Qu, Y., and Wiebe, J., editors, *Computing Attitude and Affect in Text: Theory and Applications*, volume 20 of *The Information Retrieval Series*, pages 41–47. Springer Netherlands.

Nakatani, M., Jatowt, A., and Tanaka, K. (2009). Easiest-first search: Towards comprehension-based web search. In *Proc. of CIKM*, pages 2057–2060.

Nakatani, M., Jatowt, A., and Tanaka, K. (2010). Adaptive ranking of search results by considering user's comprehension. In *Proc. of the 4th International Conference on Uniquitous Information Management and Communication*, pages 27:1–27:10.

Paek, T. and Chandrasekar, R. (2005). Windows as a second language: an overview of the jargon project. In *Proc. of the First International Conference on Augmented Cognition*.

Park, Y., Byrd, R. J., and Boguraev, B. K. (2002). Automatic glossary extraction: Beyond terminology identification. In *Proc. of COLING*, pages 1–7.

Petersen, S. E. and Ostendorf, M. (2009). A machine learning approach to reading level assessment. *Comput. Speech Lang.*, 23(1):89–106.

Pitler, E. and Nenkova, A. (2008). Revisiting readability: a unified framework for predicting text quality. In *Proc. of EMNLP*, pages 186–195.

Qumsiyeh, R. and Ng, Y.-K. (2011). ReadAid: A robust and fully-automated readability assessment tool. In *Proc. of the 23rd IEEE International Conference on Tools with Artificial Intelligence*, pages 539–546.

Robertson, S., Walker, S., Jones, S., Hancock-Beaulieu, M., and Gatford, M. (1996). Okapi at trec-3. pages 109–126.

Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Inf. Process. Manage.*, 24(5):513–523.

Schwarm, S. E. and Ostendorf, M. (2005). Reading level assessment using Support Vector Machines and statistical language models. In *Proc. of ACL*, pages 523–530.

Senter, R. and Smith, E. (1967). Automated readability index. *Cincinnati University Ohio*.

Si, L. and Callan, J. (2001). A statistical model for scientific readability. In *Proc. of CIKM*, pages 574–576.

Silverstein, C., Marais, H., Henzinger, M., and Moricz, M. (1999). Analysis of a very large web search engine query log. *SIGIR Forum*, 33(1):6–12.

Tan, C., Gabrilovich, E., and Pang, B. (2012). To each his own: personalized content selection based on text comprehensibility. In *Proc. of WSDM*, pages 233–242.

Tanaka-Ishii, K., Tezuka, S., and Terada, H. (2010). Sorting texts by readability. *Comput. Linguist.*, 36(2):203–227.

Van Oosten, P. and Hoste, V. (2011). Readability annotation: Replacing the expert by the crowd. In *Proc. of the 6th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 120–129.

Vakkari, P., Pennanen, M., and Serola, S. (2003). Changes of search terms and tactics while writing a research proposal a longitudinal case study. *Inf. Process. Manage.*, 39(3):445–463.

Vapnik, V. N. (1995). *The nature of statistical learning theory*. Springer-Verlag New York, Inc.

Wang, Q., Xu, J., Li, H., and Craswell, N. (2011). Regularized latent semantic indexing. In *Proc. of SIGIR*, pages 685–694.

White, R. W., Dumais, S. T., and Teevan, J. (2009). Characterizing the influence of domain expertise on web search behavior. In *Proc. of WSDM*, pages 132–141.

Yamamoto, M. and Church, K. W. (2001). Using suffix arrays to compute term frequency and document frequency for all substrings in a corpus. *Comput. Linguist.*, 27(1):1–30.

Yan, X., Lau, R. Y., Song, D., Li, X., and Ma, J. (2011). Toward a semantic granularity model for domain-specific information retrieval. *ACM Trans. Inf. Syst.*, 29(3):15:1–15:46.

Yan, X., Song, D., and Li, X. (2006). Concept-based document readability in domain specific information retrieval. In *Proc. of CIKM*, pages 540–549.

Zha, H., Marques, O., and Simon, H. (1998). Large-scale SVD and subspace-based methods for information retrieval. In Ferreira, A., Rolim, J., Simon, H., and Teng, S.-H., editors, *Solving Irregularly Structured Problems in Parallel*, volume 1457 of *Lecture Notes in Computer Science*, pages 29–42. Springer Berlin / Heidelberg.

Zhao, J. and Kan, M.-Y. (2010). Domain-specific iterative readability computation. In *Proc. of JCDL*, pages 205–214.