A Study of Concept-based Document Readability in Domain Specific Information Retrieval

Xin Yan
School of Information Technology and Electrical Engineering
The University of Queensland
QLD 4072, Australia
yanxin@itee.uq.edu.au

Dawei Song
Knowledge Media Institute & Centre for Research in Computing
The Open University
Walton Hall, Milton Keynes, MK7 6AA, United Kingdom
d.song@open.ac.uk

Xue Li
School of Information Technology and Electrical Engineering
The University of Queensland
QLD 4072, Australia
xueli@itee.uq.edu.au

ABSTRACT

Domain specific information retrieval has become in demand. Not only domain experts, but also average non-expert users are interested in searching domain specific (e.g., medical and health) information from online resources. However, a typical problem to the average users is that the search results are always a mixture of documents with different levels of readability. Non-expert users may want to see the documents with higher readability on the top of the list. Consequently the search results need to be re-ranked in a descending order of readability. It is often not applicable for domain experts to manually label the readability of documents for large databases. Computational models of readability have been investigated. However, traditional readability formulas are designed for general purpose text and insufficient to deal with technical materials for domain specific information retrieval. More advanced algorithms such as textual coherence model are computationally expensive for re-ranking a large number of retrieved documents. In this paper, we propose an effective and computational tractable concept-based model of text readability. In addition to the textual genres of a document, our model also takes into account domain specific knowledge, i.e., how the domain-specific concepts contained in the document affect the document's readability. Three major readability formulas are proposed and applied to health and medical information retrieval. Experimental results show that our proposed readability formulas can make remarkable improvements over four traditional readability measures.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Retrieval Models

Keywords

Readability, Readability Formula, Coherence

1. INTRODUCTION

Domain specific information retrieval, particularly in the health and medical area, has become more and more in demand. Not only domain experts, but also average (i.e., non-expert) users are interested in searching domain specific information from online resources. According to the Pew Internet & American Life Project [4], for example, 52 million American adults have used the Web to get health or medical information, and 83% percent of them thought that online materials affected their decisions about treatment and care for relatives and themselves. In addition, 91% health information seekers have searched for information about physical illness.

Unlike general-purpose information retrieval systems, domain specific information retrieval is dealing with retrieval of domain specific documents, often a mixture of documents with different levels of readability. A recent investigation [5] found that, for 70% of 80 internet health web sites, users will need at least 2 years of college level education to comprehend their posted privacy statements. Another study [11] shows that much of the medical information for the public is written at a "10th grade, 2nd month reading level", higher than the average reading level of the population of patients. Therefore, to average users with little domain knowledge and low education level or the sicks and elders under physical, psychological, and emotional stress, there is a need for an IR system to find documents not only relevant to a query but also with higher degrees of readability.

Domain specific search imposes two major requirements to the readability computation. First, domain specific information contains a large number of technical materials, however, commonly used readability formulas were not developed for technical materials [9]. Second, in online information retrieval, readability measurements are expected to be efficient enough to deal with a large number of documents in real time. There are many existing techniques to measure the readability of textual documents. However, none of them meets the above requirements. Some complex readability measures are even more computationally expensive.

To address the problem, our work is focused on building

an effective and computational tractable model of document readability by taking advantages of both traditional readability formulas and domain knowledge. The latter may be defined in a domain-specific controlled vocabulary or taxonomy. We propose a novel concept-based readability model which takes into account the role of domain-specific concepts, contained by a document for determining the readability of the document. Two basic features of documents, cohesion and scope [17, 16], are adopted together with a traditional readability formula to build a computation model of word level document readability for domain specific materials. We propose three readability formulas based on this model and test them in health and medical information retrieval. Through the correlation analysis between users' readability judgements and the computations of readability formulas, our proposed measures demonstrate the outstanding performances in comparison with traditional readability formulas.

The rest of this paper is organized as follows: Section 2 introduces different measures of readability including traditional readability formulas and a advanced coherence model. Section 2.3 discusses the major problems of applying readability measures on domain specific information retrieval. Our work on a concept-based readability model are proposed in Section 3. Section 4 reports experimental setup and results. Section 5 gives conclusions and highlights some future research directions.

2. MEASURES OF READABILITY

Readability measures aim at matching the difficulty degrees of understanding a piece of text against the reading levels of readers. Through decades, many readability metrics were proposed to help the writers and authors compose texts which can be understood easily by the target readers. According to the theory of discourse [8], these metrics can be categorized into three types: surface code level metrics, text-base and the situation model based measures.

2.1 Surface Code Level Metrics

Given a piece of text, its surface code level is measured by wording and syntax of sentences. Documents full of "difficult" words are apparently more difficult for readers to understand than documents with simple words. Moreover, a large portion of long sentences with complex syntax will surely make a document more difficult to read. The word difficulty and sentence difficulty can be computed in various straightforward ways. Most of them such as Flesch Reading Ease Score, Flesh-Kincaid Grade Level, SMOG Index, Gunning-Fog Index, Automated Readability Index (ARI), Coleman-Liau and Dale-Chall will generate a numeric readability score. This score can correspond to an educational grade level. Thereby, these surface code level metrics are also called "grade-level formulas".

2.1.1 Word Difficulty Computation

McCallum *et al* [12] summarized that six features reflecting word difficulty can be used in readability formulas. The most commonly used features include length of words, number of syllables in words, and popularity of words. For example, the Flesch Reading Ease Score and Flesh-Kincaid Grade Level measure the average syllables per word. The SMOG Index and Gunning-Fog Index consider the number of words in a document containing no less than 3 syllables. The Automated Readability Index (ARI) and Coleman-Liau take the number of characters per word into account. Finally, a common word list is used in Dale-Chall's Readability Index formula. Words which are not in the list are regarded as difficult words. The readability of document is counted mainly by the percentage of difficult words in the document.

2.1.2 Sentence Difficulty Computation

Sentence difficulty is computed by measuring the syntactic features of sentences. The most popular feature is the length of sentence, usually a calculation of number of words per sentence. This method has been used in all the popular surface code level metrics.

2.2 Text-Base and Situation Model

According to Kintsch's theory [8], there are at least three levels cognitive representation to be built by user trying to understand something from the text. They are source code, text-base and situation model. The source code is the literal words and the way they are organized as sentences. The text-base consists of the surface meanings of clauses presented by the source code. The situation model is a user's mental model built on text-base with user's background knowledge for the purpose of understanding what the text is about. It may consist of something behind the surface meaning of the text (i.e. text-base) such as time, place, event and causality. The construction of situation model needs inferences based on the reader's knowledge and understanding of text-base. Current theory of discourse believes that in general the more coherence the situation model is, the more text comprehension the reader can achieve. In full awareness of the difficulty of detecting reader's knowledge level, researches are mainly focused on the cohesion of text-base by using the statistical or natural language processing methods. for example, the Latent Semantic Analysis (LSA) [10]has been used to represent semantic content and measure the text-base cohesion [3, 14].

2.3 The Problems of Readability in Domain Specific IR

Applying readability measures in information retrieval process is an interesting and challenging topic. By incorporating readability analysis, we expect automated domain-specific information retrieval to be more beneficial to average non-expert searchers. However, can the existing readability measures cope with domain specific information retrieval? In this subsection, we will discuss this problem in detail.

2.3.1 Readability of Technical Materials

Three problems arise when the existing word difficulty measures are applied to domain-specific materials.

One: technical word difficulty cannot be measured by number of syllables or characters in word.

There are not only common words but also technical or professional terms appearing in technical materials. We discover that the word difficulties of technical or professional terms cannot be effectively measured by the commonly used syllable counting and word length counting. It is an incorrect assumption that the more syllables or more characters a word contains, the more difficulty readers may encounter in order to read and understand the word.

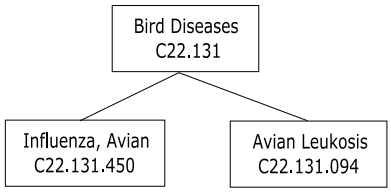

Figure 1: A Fragment of MeSH

Table 1: MeSH Levels and Word Difficulty Features

Level	Syllables	Characters
1	2.72	8.05
2	2.85	8.43
3	3.16	8.94
4	3.38	9.36
5	3.40	9.21
6	3.56	9.59
7	3.40	9.23
8	3.40	9.24
9	3.10	8.11
10	3.46	8.49
11	3.00	7.00

We verify the above statement by a study on Medical Subject Headings[1] (MeSH), the National Library of Medicine's controlled vocabulary thesaurus. It is used to index health and medical materials in the MEDLINE, an online database containing more than 11 million citations and abstracts from health and medical journals and other sources. The MeSH consists of sets of descriptors (concepts) in a hierarchical structure. The descriptors are organized to form a concept hierarchy by broader and narrower relationships. Figure 1 shows a fragment of the MeSH structure.

In this example, the descriptors "Avian Leukosis" and "Influenza, Avian" are two types of "Bird Diseases". It is obvious that the deeper a descriptor is in the MeSH hierarchy, the more technical it is and in turn the more difficult it is for a non-expert reader to read and understand. According to this feature of MeSH, our study aims to find out if the two traditional word difficulty measures still hold, i.e., whether the average number of characters and the average number of syllables per descriptor on a MeSH level will increase when the level goes deeper. Table 2.3.1 shows the relationship among the level of MeSH, the number of syllables per descriptor and the number of characters per descriptor in the level.

From Table 2.3.1, obviously the numbers of syllables and characters per descriptor do not go up with the levels (from one to eleven). Therefore, it gives evidence that these two variables are insufficient to measure word difficulty for domain specific materials.

Two: grade level metrics is unsuitable for technical materials.

As we know, readability measurement is a mapping between the reading difficulties of textual materials and the grade levels of readers. However, it is unsuitable to use grade level metrics for technical materials. Current grade level metrics were originally designed to measure children's school books rather than domain specific materials [13]. Moreover, some of the readers of technical materials, such as the sicks and elders under physical, psychological, and emotional stress, may have high grade levels but low abilities to read domain specific materials. Therefore, as a main idea in this paper, we mainly take the relative readability scores rather than the absolute grade levels as the outputs of our proposed readability measure of technical materials.

Three: "common" words are not always common.

As we know, the classic Dale-Chall's Readability Index measures word difficulty by counting the percentage of words which are out of Dale-Chall Word List[2], a list of 3,000 words which were known to be familiar to most of the fourth-grade readers. However, "common words" are not always common. Redish *et al* [13] questioned that in some technical or professional materials, common words will become technical terms. For example, the word "shock" is a common word frequently used in everyday life. It is regarded as a common word in Dale-Chall Word List. However, in medical and health materials, the meaning of "shock" could be "a pathological condition that can suddenly affect the hemodynamic equilibrium, usually manifested by failure to perfuse or oxygenate vital organs"[3].

Nevertheless, we argue that it's not a problem of Dale-Chall's Readability Index to measure the readability of technical materials. On the one hand, the case like "shock" is true but not quite usual. On the other hand, to the best of our knowledge, no user study has been conducted to investigate this drawback of Dale-Chall's Readability Index formula. In this paper, we will evaluate the effectiveness of Dale-Chall's Readability Index formula in the context of technical materials.

2.3.2 Readability and Document Ranking

Document ranking is a fundamental feature for almost all information retrieval (IR) systems. It may tell user how important a document is to the query. Due to the large amount of search results and the limitation of user's time and patience, it is impractical for user to review all the retrieved documents and judge their relevance. Study about user's searching behavior [6] in Google reveals that user's attention to search result drops down promptly when browsing the retrieved documents. Thereby in what order to present relatively highly readable documents is a major problem in order to benefit average non-expert users.

The research on searching behavior [6] also reveals that searchers usually spend little time, around 5 seconds only, in selecting documents to read from a ranked list of search results. Only 200 to 300 milliseconds eyes fixation on average are spent to acquire information from a piece of description of a document in order to judge its relevance. It consequently reveals that searcher will not read the abstract of a retrieved documents sentence by sentence. The major reading difficulty a searcher may encounter is at word level. Therefore, an efficient word-based document readability measurement is in great demand. In this paper, we propose a concept-based readability model which focuses on

[1]http://www.nlm.nih.gov/pubs/factsheets/mesh.html

[2]http://www.corelearn.com/PDFS/Articles/Dale-Chall%20Word%20list.pdf

[3]http://www.nlm.nih.gov/pubs/factsheets/mesh.html

the estimation of word level readability of domain specific search results (documents).

3. A CONCEPT-BASED DOCUMENT READABILITY MODEL

We investigate the problem of readability measurement for domain specific information retrieval from the following four perspectives.

The first is to consider the measurement of relative readability. Since the goal of our research is to find an approach to readability measurement in document ranking process, the relative score of document readability is more sensible than an absolute one in this case.

The second is to consider word level readability rather than sentence level. The difficulty in browsing a list of retrieved documents is mainly based on the word level, as we mentioned in the previous section.

The third is to consider the concept-based readability in a given knowledgebase, such as WordNet and MeSH, where concepts are organized from general to specific in a hierarchical structure. We will show that the coverage of domain concepts in a document and their relatedness are two major features of domain specific readability. The computations of these two features are respectively named as document scope and document cohesion, which are derived from our previous work [17, 16] about document generality.

The forth is to use Dale-Chall's common word list to measure the word level based readability in technical documents. The concept-based readability measure may have a limitation since it is often the case that a knowledgebase cannot cover all the words in documents. Word level based readability can be measured by calculating the percentage of words in document which are out of the Dale-Chall's common word list. Dale-Chall's method is a typical measure which doesn't take the number of syllables in words and the length of words as the variables to calculate word level readability, which have certain drawbacks as we mentioned above.

The hypothesis underlying our proposed concept-based readability is given as follows:

- *Document Scope* (DS) - A document is considered as a collection of terms. The scope of a document is regarded as the coverage of the domain concepts in the document. The more terms of the document are identified as domain concepts in the given domain knowledgebase, i.e., a concept hierarchy, the less readable the document tends to be. Also, within the concept hierarchy, the deeper the identified concepts appear, the more difficult the document is to read.

- *Document Cohesion* (DC) - When there is a focused topic or theme discussed in a document, the terms are often closely correlated in a certain context. The cohesion of a document can be computed by the relatedness between the identified concepts. The relationships between concepts are reflected by the links in the given concept hierarchy. Derived from Kintsch's theory [8], the more cohesive the concepts are, the more readable the document tend to be.

In following subsections, we will present our approach to the computation of concept-based word level readability in the context of medical and health information retrieval. Here we regard the document terms (including compounds) which have a match in a conceptual hierarchy as domain concepts. The terms which cannot be found in the conceptual hierarchy are referred to non-domain terms.

3.0.3 Document Scope

Document scope is a major characteristic of concept-based readability. For example, consider the following two definitions of SARS. Definition 1 comes from ABOUT [4], a web information service for daily life. Definition 2 is an official definition from the Department of Health in Hong Kong [5].

1. (d_1) A viral respiratory illness that was recognized as a global threat in March 2003.

2. (d_2) A viral *respiratory infection* caused by a *coronavirus* ($SARS\text{-}CoV$).

We may identify 3 domain concepts: "respiratory infection", "coronavirus" and "SARS-CoV" in Definition 2. However, in Definition 1 which is more public-oriented does not contain any domain concept. Instead, "Respiratory illness" is used to broadly describe SARS rather than a more specific concept "respiratory infection". It is obvious that the scope of the general definition (i.e. Definition 1) is larger than the scope of the professional definition (i.e. Definition 2).

We use a function to sum up the tree depths of all the individual concepts in the document to calculate its scope. Both domain concepts and non-domain terms need to be considered. Operationally, the tree depth of a domain concept is measured by the distance between that concept and the root in the MeSH hierarchy. The tree depth of a non-MeSH term is set to be zero. Therefore the document scope function is a monotonic decreasing function of the mean of tree depths of all concepts in document.

It is often the case that a document contains large percentage of non-domain terms but just small portion of domain concepts. Consequently the scope values of this kind of documents may be very skewed. To make the scope computation more sensitive to the different average tree depths of the domain concepts in different documents, we employ an exponential function.

$$Scope(d_i) = e^{-\left(\sum_{i=1}^{n} depth(c_i)\right)} \quad (1)$$

In Equation 1, the function $depth(c_i)$ gets the depth of concept c_i in the MeSH hierarchy. The maximum value of document scope is 1 when a document contains only non-domain terms. The time complexity of scope-based readability measurement of m retrieved documents is $O(m \times n)$, where n is the maximum number of terms (i.e. both domain concepts and non-domain terms) per document.

3.0.4 Document Cohesion

Document cohesion is another feature of concept-based readability. It measures the relatedness degree of concepts in a document. The intuition of our approach is that the more cohesive the domain concepts of a document are, the more readable the document is.

[4] http://about.com
[5] http://www.info.gov.hk

Consider two sentences, the first is the title of a journal paper from AIDS, an official journal of the international AIDS society[6], and the second from MeSH.

1. (d_1) *AIDS* Events Among Individuals Initiating *HAART*: Do Some *Patients* Experience a Greater Benefit From *HAART* Than Others?
2. (d_2) *HIV* is a non-taxonomic and historical term referring to any of two species, specifically *HIV-1* and/or *HIV-2*.

In d_2, three domain concepts can be identified: "HIV", "HIV-1" and "HIV-2". In d_1, "AIDS", "HAART" and "Patients" are three identified domain concepts. They both contain the same number of domain concepts, thereby their readability cannot be distinguished in terms of their document scope. However, there is a stronger cohesion in d_2 than in d_1. In other words, concepts in d_2 are more strongly associated to each other than those in d_1. Specifically, "HIV-1" and "HIV-2" are two types of "HIV", i.e., there are direct links between them in MeSH. People who even doesn't know HIV-1 and HIV-2 can easily understand this sentence. However, in d_1, "AIDS" is a kind of symptom but "HAART" is a therapy for "AIDS". There is no direct link between "AIDS" and "HAART" in MeSH. People with low domain knowledge can hardly understand what HAART is and what the relationship is between HAART and AIDS.

The cohesion of a document is computed as the associations (semantic relatedness) between the domain concepts in the document. The more closely the concepts are associated, the higher degree of cohesion the document has. Since the cohesion calculation is exactly the same as the one in our previous work about document generality [17, 16], we just briefly list all the equations here.

$$Cohesion(d_i) = \frac{\sum_{i,j=1}^{n} Sim(c_i, c_j)}{Numberof Associations}, \; where\, n > 1, \; i < j \tag{2}$$

$$Sim(c_i, c_j) = -log\frac{len(c_i, c_j)}{2D} \tag{3}$$

$$Numberof Associations = \frac{n(n-1)}{2} \tag{4}$$

In Equation 2, n is the total number of domain concepts in a document d_i. $Sim(c_i, c_j)$ is a function computing the Leacock-Chodorow semantic similarity of concept c_i and c_j. $len(c_i, c_j)$ is the function to calculate the shortest path between c_i and c_j in the MeSH hierarchy. $Numberof Associations$ is the total number of mutual associations among domain concepts, which is defined in Equation 4.

In Equation 3, D is the maximum tree depth in the concept hierarchy. In our experiments, D is 11. The scope of Equation 2 is thus $[0, -log(\frac{1}{22})]$. For a document with less than one domain concept, its document cohesion is 0. For a documents with strongest associations among all the concepts within the document, its cohesion is $-log(\frac{1}{22})$, the maximum value. The time complexity of cohesion-based readability measurement on m retrieved documents is $O(m \times n^2)$, n is the maximum number of domain concepts in documents.

[6]http://www.medscape.com

3.0.5 Overall Concept-based Readability Score

We propose the following Equation 5 to calculate overall concept-based readability score (CRS), which is proportional to document scope, document cohesion and the reciprocal of Dale-Chall's Readability Index. Equation 6 is Dale-Chall's Readability Index.

$$CRS(d_i) = Scope(d_i) + Cohesion(d_i) + DaCw(d_i)^{-1} \tag{5}$$

$$DaC(d_i) = (0.0496 * AvgSL) + (0.1579 * PDW) + 3.6365 \tag{6}$$

$$DaCw(d_i) = PDW \tag{7}$$

In Equation 5, $DaCw$ is the simplified Dale-Chall's Readability Index function. In Equation 6, $AvgSL$ is the average length of sentence in document d_i. PWD is the percentage of difficult words in d_i. Difficult words are all the words out of the Dale-Chall Word List[7]. Since we are mainly focusing on the word level readability, the sentence level complexity is removed from Equation 6 to get a simplified Equation 7.

Dale-Chall's Readability Index is a popular readability formula. It takes the commonness of words into consideration rather than the syllables and length of words which were mentioned previously to be insufficient to act as features of word level readability. The simplified Dale-Chall's Readability Index can be a complement of the concept-based scope and cohesion calculation. Since the large the Dale-Chall Readability Index is, the less readable the document is, we join the reciprocal of the simplified Dale-Chall Readability Index into the Equation 5.

4. EXPERIMENTS AND EVALUATIONS

In order to evaluate our proposed concept-based readability measures, the user oriented experiments and evaluations are performed. Human judgements are treated as ground truth in our experiments. The proposed concept-based readability measures with other four widely used classical measurements of word level readability are then compared with the ground truth.

4.0.6 Test Measures

In our experiments, we consider the following three major scenarios of readability computation, where document scope, cohesion and simplified Dall-Chall Readability Index are combined in a reasonable manner. Those three cases are derived from our proposed Equation 1, 2 and 7. They are listed as followings.

DSChall

In this scenario, both document scope and simplified Dall-Chall Readability Index are considered to compute word level based readability. See Equation 8.

$$CRS(d_i) = Scope(d_i) + DaCw(d_i)^{-1} \tag{8}$$

[7]http://www.corelearn.com/PDFS/Articles/Dale-Chall%20Word%20list.pdf

DCChall

In this scenario, both document cohesion and simplified Dall-Chall Readability Index are considered to compute word level based readability. See Equation 9.

$$CRS(d_i) = Cohesion(d_i) + DaCw(d_i)^{-1} \quad (9)$$

DSDCChall

In this scenario, all the features, document scope, document cohesion and simplified Dall-Chall Readability Index, are considered to compute word level based readability. See Equation 1.

In order to individually study the performances of document scope and document cohesion in our proposed measures of word level based readability, other three scenarios are evaluated, they are:

DS

In this scenario, only document scope is considered to compute word level based readability, i.e. Equation 1.

DC

In this scenario, only document cohesion is considered to compute word level based readability, i.e. Equation 2.

DSDC

In this scenario, both document scope and document cohesion are considered to compute word level based readability. See Equation 10.

$$CRS(d_i) = Scope(d_i) + Cohesion(d_i) \quad (10)$$

There are four classical readability formulas which are widely used in different areas. They are: Automated Readability Index (ARI) [15], Flesch-Kincaid Grade Level(FKG) [2], Gunning-Fog Index (GFI) [7] and Dale-Chall Readability Index (DaC) [1]. They are simplified by removing the part of sentence level readability calculation in order to test their effectiveness of word level readability computations. All the four simplified formulas are listed as follows.

ARI

$$ARIw(d_i) = averageCharactersWord(d_i)^{-1} \quad (11)$$

Flesch

$$Fleschw(d_i) = averageSyllablesWord(d_i)^{-1} \quad (12)$$

Gunning

$$Gunningw(d_i) = percentageThreeSyllables(d_i)^{-1} \quad (13)$$

Chall

A simplified Dale-Chall's Readability Index formula, see Equation 7.

4.0.7 Test Materials

The test materials are all from the PubMed Central (PMC)[8], which is a free digital archive generated by U.S. National Institutes of Health (NIH). The PMC contains several-hundred thousands free full text articles about biomedical and life sciences.

The test materials are related to the frequently requested health topics in MedlinePlus[9], which is a site providing "several search mechanisms for searching for health topics of interest to the general public in PubMed"[10]. These topics are: Alzheimer's Disease, Back Pain, Breast Cancer, Cholesterol, COPD (Chronic Obstructive Pulmonary Disease), Depression, Diabetes, Fibromyalgia, Gastroesophageal Reflux/Hiatal Hernia, High Blood Pressure, Lupus, Parkinson's Disease, Pregnancy, Prostate Cancer and Sexually Transmitted Diseases.

We first submit TOPIC NAME + "FAQ" for each topic it to Google. From Google search results, we get a set of frequent answered questions for all the nominated health topics, from that we then generate a set of test queries.

We take advantage of the PubMed's feature of searching with MeSH database[11] in order to formulate the 18 queries. The main search topics in our 18 primitive queries are replaced by the corresponding domain concepts. All the stop words are removed according to the SMART 571 stop word list. For example, "what is depression?" is formalized as "Depression[MAJR:NoExp] ". MAJR means "restrict search to major topic headings only" and NoExp means "do not explode this term (i.e., do not include MeSH terms found below this term in the MeSH tree)". Other keywords which cannot be found in the MeSH are appended to the query by a boolean operator "AND".

Then in PubMed, the top 10 documents are retrieved for each query. Note that some queries may have less than 10 documents returned. Thus we form a data set containing titles and abstracts of a total number of 156 distinct documents retrieved by the 18 queries.

4.0.8 Users and Questionnaire

There are 14 users involved in the user experiments. They all have had higher education qualifications with no less than ten years English language training. All of them are requested to voluntarily complete a questionnaire about the retrieved documents for each of 18 queries. All users have only basic knowledge about health and medical topics. There is no time limit for users to complete the questionnaire.

The titles and abstracts of retrieved documents under each query are listed (see Figure 2) in an original order of PubMed's original similarity based ranking.

[8] http://www.pubmedcentral.nih.gov/
[9] http://medlineplus.gov/
[10] http://www.nlm.nih.gov/bsd/special_queries.html
[11] http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=mesh

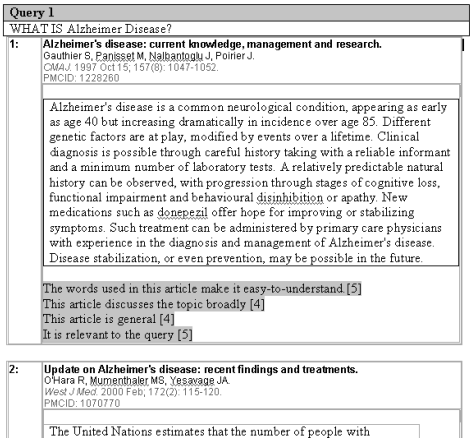

Figure 2: A Sample Questionnaire

Table 2: User Data Filtering Result

	Before Filtering	After Filtering
O1	24.22%	22.29%

In this questionnaire, users will score for the opinion denoted by O1 for every retrieved document under each of the 18 queries. The scores is in a range from 1 to 5. One for "strongly disagree", two for "disagree", three for "have no idea", four for "agree", and five for "strongly agree". The opinion which is related to readability study is:

1. O1: The words used in this document make it easy-to-understand.

Users' ratings of this opinion would reflect their judgements of word level based readability of the retrieved documents.

4.0.9 *Preprocessing and Analysis of Users' Ratings from Questionnaires*

The completed questionnaires may contain some users' hasty and careless inputs. In order to identify the outliers, a statistical filtering process is applied based on the statistics of the data, such as standard deviation, mean and coefficient of variation. Coefficient of variation is a basic statistical method to measure the relative scatter in user's ratings. It is calculated by dividing the standard deviation of all the 14 users' ratings (of an opinion for one of the 156 retrieved documents) by their mean value. The range of co-efficient of variation is from 0% to 100%. High co-efficient of variation value means that users cannot have a fairly consistent ratings about an opinion for a document retrieved. For each of the 156 retrieved documents, all the 14 users' ratings of the corresponding opinion are processed by using those statistical methods, the ratings of individual users which are two times or more than the standard deviation will be dropped.

Table 2 shows the improvement of average co-efficient of variations of users' ratings for 156 retrieved documents respectively under O1 after the statistical filtering process.

4.0.10 *Evaluation Methodology & Performance Indicator*

Our proposed three cases of concept-based readability measures together with other four simplified readability formulas are run on a data set of 156 documents retrieved by 18 queries. The output rankings of each query are then compared with 14 users' ranking. The correlation between the human readability judgements and computer readability measures is used in our study as the performance indicator of our test readability formulas. The higher the correlation is, the better the readability measure is. Spearman rank-order correlation coefficient is used for correlation analysis since it is a distribution free test considering the rank of a data item instead of its value.

Moreover, four levels of the significance (p) for a two-tailed test, namely critical values, are used to measure the confidence of the Pearson's correlation coefficient. The four levels are: 0.10, 0.05, 0.02 and 0.01. The critical value of a Pearson's correlation coefficient is fairly significant when it is less than 0.02, highly significant when it less than 0.01. It can be determined by using the critical values table when both the correlation value and the degree of freedom are given. The degree of freedom is the total number of data points in the correlation analysis minus 2. The greater both the degree of freedom and the correlation coefficient are, the higher the critical value can be achieved.

4.0.11 *Experimental Results*

There are totally 18 queries, 156 retrieved documents and 14 users involved in this experiment. In Table 3, the Spearman rank-order correlation coefficients are calculated between the selected cases of readability measures and users' ratings of word level based readability. Query# is the query number from 1 to 18. $n-2$ is the degree of freedom for a two-tailed test. n is the total number of documents which are retrieved by a particular query. Avg is the mean function. As to a specific readability measure, the mean value of the degree of freedom and the mean value of the correlation coefficient in all the 18 queries indicate the average performance of that measure.

The intuitive diagram of the correlations between users and selected formulas is presented in Figure 3. The X coordinator represents different readability measures. The Y coordinator represents the Pearson's correlation coefficient.

4.0.12 *Results Analysis*

In Table 3 it is obvious that our proposed three readability measures DSChall, DCChall and DSDCChall have strong correlations with human's readability judgements in almost all the 18 queries. Their average degree of freedom is 7. With this value, we found that the critical values of DSChall and DCChall are all below 0.01, and the critical value of DSDCChall is below 0.02. It shows that all those strong correlation values are highly confident. By comparison, the correlation between each of the four test readability measures (i.e. Automated Readability Index, Flesch-Kincaid Grade Level(FKG), Gunning-Fog Index and Dale-Chall's Readability Index) and human judgements is quite weak.

Through Figure 3, it is obvious that our proposed readability measures, DSChall, DCChall and DSDCChall, remarkably have a far and away improvement of the performance in readability calculation in comparison with all the other four word level based readability measures, which are, Automated Readability Index, Flesch-Kincaid Grade

Table 3: Word Level Correlations between Users and Selected Formulas

Query#	n-2	ARI	Flesch	Gunning	Chall	DS	DC	DSChall	DCChall	DSDC	DSDCChall
1	8	0.1212	-0.0485	0.0848	0.7879	-0.0970	0.0606	0.9818	0.9697	0.0970	0.9455
2	8	0.2182	0.2182	0.2061	0.6545	-0.7152	-0.0242	0.9576	0.9091	0.2182	0.8242
3	7	0.1833	0.3000	0.1833	0.3333	0.7000	0.7000	0.9833	0.9667	0.3000	0.9833
4	8	0.6606	0.6606	0.4909	0.7091	0.1394	0.5152	0.9636	0.9394	0.6364	0.8545
5	8	0.6061	0.3394	0.6303	0.4364	0.0242	0.5091	0.9697	0.8848	0.3273	0.8000
6	8	-0.5515	-0.3939	-0.4061	0.2909	0.0727	-0.0485	0.7212	0.4545	-0.4303	0.2606
7	8	0.4545	0.3697	0.2606	0.2000	0.1879	0.6848	0.9879	0.8788	0.4424	0.8788
8	8	0.2909	0.4000	0.6061	0.6667	0.1697	0.1333	0.9818	0.8970	0.4000	0.8727
9	8	0.3697	0.5879	0.4303	0.3455	-0.0424	0.0182	0.9758	0.9394	0.4788	0.9394
10	8	0.5030	0.6727	0.3697	0.6727	0.4788	0.7091	0.9758	0.9636	0.6970	0.9636
11	8	0.3394	0.3394	0.2182	0.6545	0.5939	0.4606	0.9818	0.8848	0.4364	0.8121
12	8	0.1515	0.1515	-0.0667	0.5152	0.5394	-0.4061	0.9515	0.7818	0.1515	0.7091
13	8	-0.5212	-0.5455	-0.4848	0.1212	0.5212	-0.0364	0.9697	0.6303	-0.6061	0.4485
14	5	-0.2143	-0.3571	-0.2500	-0.1786	0.1071	0.4286	0.8571	0.3571	-0.3571	0.2500
15	2	1.0000	0.4000	0.4000	0.4000	-0.2000	-0.4000	1.0000	0.8000	0.8000	1.0000
16	2	0.4000	0.4000	0.2000	-0.2000	-0.8000	-0.4000	1.0000	1.0000	0.4000	1.0000
17	2	0.4000	0.2000	0.2000	0.4000	-0.4000	-1.0000	1.0000	0.8000	0.2000	0.8000
18	6	0.5238	0.5238	0.5952	0.1190	0.0476	0.4286	0.9286	0.9286	0.5238	0.9048
Avg	7	0.2742	0.2343	0.2038	0.3849	0.0737	0.1296	**0.9548**	**0.8325**	0.2620	**0.7915**

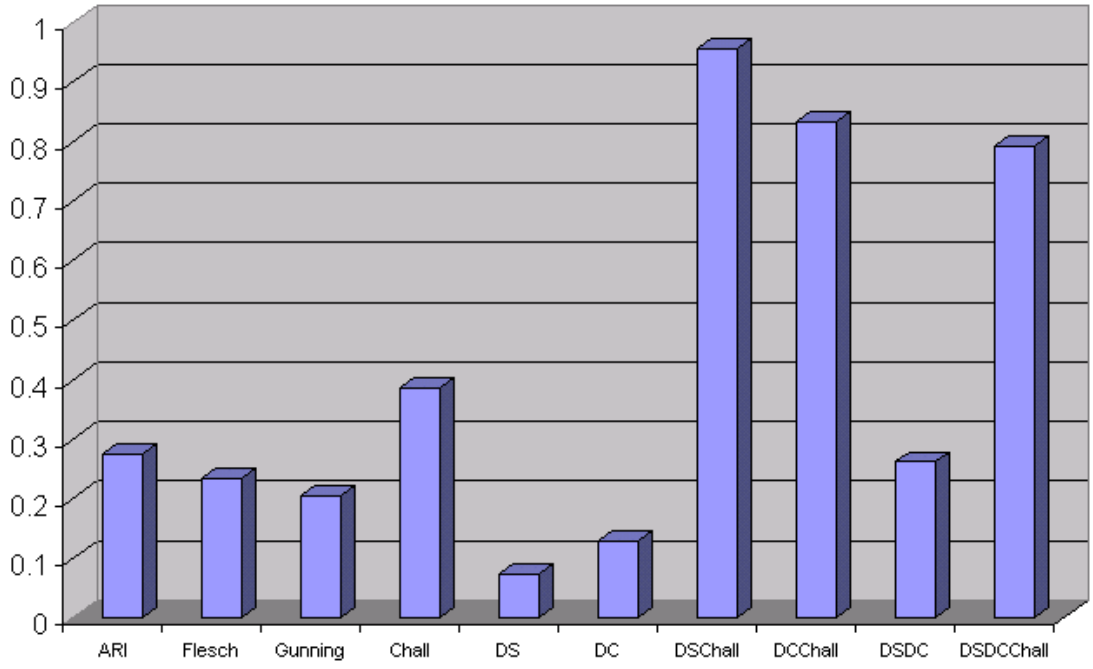

Figure 3: Comparison of Word Level Readability Measures

Level(FKG), Gunning-Fog Index and Dale-Chall's Readability Index. We performed a dependent t-test (Paired Two Sample for Means) which compares the paired Pearson's correlation coefficients between any of DSChall, DCChall, DSDCChall and any of the four word level based readability measures. With all the $p-value$ less than 0.01, it turns out that the improvement made by DSChall, DCChall and DSDCChall are all statistically significant.

It is noticeable that most of the four selected readability measures (i.e. Automated Readability Index, Flesch-Kincaid Grade Level(FKG) and Gunning-Fog Index) are using number of syllables or the length of words as the major features of word level readability. Thereby it proves our observation and argument that word difficulties of technical or professional terms cannot be effectively measured by the commonly used methods such as syllables counting and word length counting.

It is also noticeable that Dale-Chall's measure (i.e. Chall), DS, DC and DSDC will have a poor performance when they are used alone. However, when Dale-Chall's measure is combined with DS or DC (i.e. DSChall, DCChall and DSDCChall), the performance is boosted far and away. This result demonstrates the advantage of our proposed approach, in which domain specific knowledgebase together with Dale-Chall's word list are used as a solution to determine the word level based readability in the context of technical materials. Moreover, the results also support our argument against Redish *et al* [13]'s opinion that the common words in the Dale-Chall's Word List becoming special jargons in technical materials makes it useless in readability measurement of technical materials. In our special case of medical and health materials, it shows that a small portion of jargons in the Dale-Chall Word List will not affect its performance very much. According to the result of a dependent t-test (Paired Two Sample for Means), Dale-Chall's Readability Index even significantly outperforms Automated Readability Index, Flesch-Kincaid Grade Level(FKG) and Gunning-Fog Index with an acceptable performance.

However, DC is not as good as DS to describe the word level based readability. It results that DSChall outperforms DCChall and DSDCChall. The reason is that it simply considers semantic relationship between domain concepts only. Since there are a large percentage of non-domain terms in the documents, it is necessary to consider statistical relationships between domain concepts and non-domain terms. A possible solution is to use the frequency of co-occurrence to measure the relatedness of concepts (i.e. both domain and non-domain terms). The more often two concepts co-occur, the stronger their association is.

5. CONCLUSIONS AND FUTURE WORKS

In this paper, we studied a readability measure problem in domain specific information retrieval where document ranking where not only domain experts, but also average users are interested in searching domain specific information from online resources such as online health resources. A typical problem to the average users is that the search results are always a mixture of documents with different readability. It is sometimes difficult for average users to quickly sort out documents with relatively high readability. It is not applicable for domain experts to manually label readability for large database. Traditional readability formulas are oversimplified to deal with technical materials. More advanced algorithms such as textual coherence model are effective but computational expensive to apply on the re-ranking approaches for a large number of retrieved documents.

We have proposed a novel applicable model of domain specific readability by taking advantages of both traditional readability formula (i.e. Dale-Chall's Readability Index) and the knowledgebase. To do this, we developed the well discussed readability problem by introducing concept-based readability, that is, how the concepts in documents make text easy to read in terms of a given domain specific knowledgebase. Three major readability formulas are proposed and applied in the context of health and medical information retrieval. The readability predictions of our proposed formulas together with 4 traditional readability formulas are compared with 14 users' readability judgements on 156 retrieved health and medical article abstracts. The results show that our proposed readability formulas can make remarkable improvements against the traditional readability measures.

Since document cohesion is not quite a significant feature of word level based readability. We plan to study other factors of word level relatedness. So far we have considered quantifying only the semantic closeness amongst domain concepts in order to calculate the document cohesion. In our further study we will consider the statistical relationships between non-domain terms and domain concepts.

6. ACKNOWLEDGMENTS

The work reported in this paper has been funded in part by the Co-operative Centre for Enterprise Distributed Systems Technology (DSTC) through the Australian Federal Governments CRC Programme (Department of Education, Science and Training).

7. REFERENCES

[1] J. S. Chall and E. Dale. *Readability Revisited: The New Dale-Chall Readability Formula.* Brookline Books, 1995.

[2] R. Flesch. A new readability yardstick. *Journal of Applied Phychology*, 32:221–233, 1948.

[3] P. Foltz, W. Kintsch, and T. Landauer. The measurement of textual cohererence with latent semantic analysis. *Discourse Processes*, 25:285–307, 1998.

[4] S. Fox, L. Rainie, J. Horrigan, A. Lenhart, T. Spooner, M. Burke, O. Lewis, and C. Carter. The online health care revolution: How the web helps americans take better care of themselves. Technical report, The Pew Internet & American Life Project, 2000.

[5] M. A. GRABER, D. M. D'ALESSANDRO, and J. JOHNSON-WEST. Reading level of privacy policies on internet health web sites. *J Fam Pract.*, 51(7):642–5, 2002.

[6] L. Granka, T. Joachims, and G. Gay. Eye-tracking analysis of user behavior in www-search. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, Sheffield, UK, 2004. ACM Press.

[7] R. Gunning. *The technique of clear writing.* Mcgraw-Hill, New York, 1952.

[8] W. Kintsch. *Comprehension: A paradigm of cognition.* Cambridge University Press, New York, 1998.

[9] G. R. Klare. The formative years. In *Readability: its Past, Present and Future*. international Reading Association, 1988.

[10] T. K. Landauer, P. W. Foltz, and D. Laham. Introduction to latent semantic analysis. *Discourse Processes*, 25(2&3):259–284, 1998.

[11] G. MA, R. CM, and K. B. Readability levels of patient education material on the world wide web. *J Fam Pract.*, 48(1):58–61, 1999.

[12] D. R. McCallum and J. L. Peterson. Computer-based readability indexes. In *Proceedings of the ACM '82 conference*, pages 44 – 48. ACM Press, 1982.

[13] J. Redish. Readability formulas have even more limitations than klare discusses. *ACM Journal of Computer Documentation*, 24(3):132 – 137, 2000.

[14] M. E. Schreiner, B. Rehder, T. K. Landauer, and D. Laham. How latent semantic analysis (lsa) represents essay semantic content: Technical issues and analysis. In *Proceedings of the 19th annual meeting of the Cognitive Science Society*, page 1041, 1998.

[15] E. A. Smith and R. J. Senter. Automated readability index. *AMRL-TR-66-22*, 1967.

[16] X. Yan, X. Li, and D. Song. Document re-ranking by generality in bio-medical information retrieval. In *The 6th International Conference on Web Information Systems Engineering*, volume LNCS 3806, pages 376–389, New York City, New York, 2005.

[17] X. Yan, X. Li, and D. Song. Document generality: its computation for ranking. In *Australian Computer Science Communications*, volume 28, pages 109–118. ACM Digital Library, 2006.