Estimating Content Concreteness for Finding Comprehensible Documents

Shinya Tanaka[1], Adam Jatowt[1,2], Makoto P. Kato[1] and Katsumi Tanaka[1]

[1]Graduate School of Informatics,
Kyoto University, Yoshida-honmachi, Sakyo-ku
606-8501 Kyoto, Japan
{s.tanaka,adam,kato,tanaka}@dl.kuis.kyoto-u.ac.jp

[2]Japan Science and Technology Agency
4-1-8 Honcho, Kawaguchi-shi, Saitama
332-0012 Tokyo, Japan

ABSTRACT

Document comprehensibility is one of key factors determining document quality and, in result, user's satisfaction. Relevant web pages are of little utility if they are incomprehensible or impose too much cognitive burden on readers. Traditional measures of text difficulty focus often on syntactic factors of text such as sentence length, word length, syllable count, or they utilize fixed list of common terms. However, document comprehensibility depends on many factors, of which concreteness and the ease of concept visualization are crucial ones. In this paper, we first propose a method for predicting the concreteness of terms using SVM regression. We then extend it to calculating document concreteness level. The experimental results indicate satisfactory accuracy in estimating both term and document concreteness as well as demonstrate positive correlation between the document concreteness and comprehensibility. Our ultimate goal is to enable comprehension-driven search, which will return both relevant and comprehensible results.

Categories and Subject Descriptors

H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval

General Terms

Algorithms, Measurements

Keywords

concreteness, abstractness, comprehensibility, readability, information retrieval

1. INTRODUCTION

Comprehensibility, defined as "the ease of understanding", is one of the significant factors of the utility of documents. The gap between a text and a reader measured, for example, in school grade levels, determines whether the text is readable or not. Texts about complex topics (e.g., on scientific, philosophical or legal issues) can be easily comprehensible for educated and skilled readers; however, for a significant fraction of readers, they constitute considerable cognitive burden. According to the National Adult Literacy Survey (NALS) [21] about 21% of the adult population in USA (around 40 million) have low literacy skills, which are defined as reading at the 6 grade level or below. Another 27% (around 50 million) have limited literacy ability, defined as lacking reading and numeric proficiency to function adequately in society. This situation effectively limits information accessibility causing documents in collections like Britannica or Wikipedia, and many others, to be poorly comprehensible for relatively large population of users, especially, when it comes to difficult topics.

While comprehensibility depends on different factors such as syntactic difficulty measured by surface text features (e.g. sentence or word length) or document coherence, we focus on concreteness which is regarded as a key aspect of content comprehensibility [23, 26]. Consider the following two passages about symptoms of Parkinson's disease.

1. Parkinson's disease affects movement, producing motor symptoms. Non-motor symptoms, which include autonomic dysfunction, neuropsychiatric problems (mood, cognition, behavior or thought alterations), and sensory and sleep difficulties, are also common.[1]

2. Patient will begin to notice early signs of Parkinson's. These may include mild shaking or tremors in one limb. Patients may experience occasional loss of balance. Patient may begin to experience symptoms on both sides of their body. Shaking may regularly occur in all limbs. Uncontrollable shaking affects patient's ability to walk, stand, and maintain balance. Patient may encounter difficulty handling otherwise simple tasks.[2]

Passage 2 contains more concrete words and seems easier to read than the passage 1. It should help non-expert users understand and also imagine the symptoms of Parkinson's

WSDM'13, February 4–8, 2013, Rome, Italy.
Copyright 2013 ACM 978-1-4503-1869-3/13/02 ...$15.00.

[1] http://en.wikipedia.org/wiki/Parkinson's_disease [27th November, 2012]

[2] http://www.googobits.com/articles/1527-guide-to-parkinsons-disease.html [27th November, 2012]

disease to greater extent than the passage 1, which has been actually taken from Wikipedia. It would be beneficial for many users to receive documents with contents characterized by similar comprehensibility degree to the one of passage 2 for difficult queries such as “Parkinson’s disease”.

Concreteness does not only affect understanding but it also has direct impact on user interest and attitude to text. Readers may find a text with too many generalizations and abstractions boring, confusing or vague. Specific content seems to have greater impact on readers than abstract, general one, because it involves their memories of sensory experiences; often, virtually making them feel, see, hear, touch, smell, and taste through reading. The good style of writing is thus to capture reader’s attention by arousing his or her senses with many concrete words [23, 26].

In addition, interest and attitude to information seem to have direct relation to its remembering as human memory is known to be driven by emotions [2]. Consider the case when one tries to teach a child to be careful. Simply saying “It is dangerous to go out at night” will have lesser effect than providing the child with information on concrete, specific experiences. While it is well-known that humans learn by abstracting concrete cases through generalization, which allows them to apply the obtained knowledge to similar situations, when provided only with abstract explanations they often have problems with comprehension and remembering.

In principle, concrete words (e.g. “car”, “table”) refer to physical entities that can be observed by at least one of the senses [22]. On the other hand, abstract words indicate abstract objects, such as ideas or concepts (e.g. “democracy” or “love”) which are hard to be materially perceived by humans. According to previous studies [22, 26] concrete terms can be represented by two psycholinguistic attributes: *perceivability* and *imageability*. Perceivability is defined as the ability to sense the object, while imageability is the ability to imagine the object easily and quickly. In this paper we propose estimating the concreteness degree of words through SVM regression [29] equipped with a range of diverse features that are related to the concepts of perceivability and imageability. As a training set we use manual labels stored in Medical Research Council Psycholinguistic Database (MRCDB)[3] which is an established source in psycholinguistic research field. Next, we extrapolate the term-level concreteness to the document-level one to enable estimation of document comprehensibility. Although the effect of concreteness on comprehensibility has been already studied in psycholinguistic field, to the best of our knowledge, this is the first attempt for estimating term- and document-level concreteness in an automatic fashion.

We make the following contributions in this paper: (a) we propose a new research problem of evaluating concreteness of documents in order to improve reading comprehension, (b) we describe method for estimating term-level concreteness and then extend it into evaluating concreteness of documents, (c) we show the results of the experimental evaluation of both the methods and, lastly, (d) we perform additional experimentation for investigating the co-occurrence of concrete/abstract terms on Wikipedia and for estimating query concreteness.

The reminder of this paper is organized as follows. Section 2 surveys the related work. In Section 3 we describe our methodology for estimating concreteness of terms and documents. In Sections 4 and 5 we report experimental results for the term- and document-level concreteness estimation, respectively, while in Section 6 we provide general discussion. Finally, Section 7 concludes the paper and outlines our future work.

2. RELATED WORK

2.1 Psychological Studies of Concreteness

Concrete and abstract concepts are commonly defined by the reference to perceivability [8, 17]. Concrete entities are considered to be physical entities with characteristic shapes, parts, materials, and so on, whereas abstract entities lack physical attributes [8, 17]. Concreteness is also defined by imageability [26, 22], which measures how quickly and easily people can imagine a given referent. Some terms arouse a sensory experience such as a mental picture or sound very quickly and easily, whereas others may arouse a sensory experience slowly or with difficulty. For example, think of the terms “banana” or “fact”. “Banana” would probably arouse an image relatively easily, whereas “fact” would most likely do so with difficulty.

Most of the previous works focused on defining the concreteness of terms and measuring it on the basis of manual evaluation. For instance, Paivio *et al.* [22] asked subjects to rate concreteness and imageability of words on 7-point scale. The main finding was that abstract words are harder to understand than concrete words. Three theories have been proposed to explain this phenomenon: the *dual-coding theory*, *age of acquisition hypothesis* and *context availability model*. The *dual-coding theory* [22, 24] is the oldest one and assumes the existence of two parallel systems: the logogen denoting the verbal system and the imagen - the human’s image system. The difference between the concrete and abstract terms is that the former activate both the systems while being more strongly associated with the imagen system, whereas abstract words tend to activate only the logogen system. On the other hand, the *age of acquisition hypothesis* [24] binds the difficulty of abstract terms with their later acquisition during the language development process. Due to the time gap subjects have simply less exposure to them and hence experience greater difficulty in comprehending and remembering abstract concepts. The last theory, the *context availability model* [24], models word comprehension by means of a complex information retrieval system operating on human’s knowledge base. The slower retrieval of abstract terms is due to their weaker associations with contextual knowledge in comparison to more concrete words.

2.2 Readability Measures

Several readability measures have been proposed due to the practical needs for estimating text reading ease (e.g., textbooks, legal contracts or technical manuals). The widely accepted definition of readability can be found in [9]: “The sum total (including all the interactions) of all those elements within a given piece of printed material that affect the success a group of readers have with it. The success is the extent to which they understand it, read it at an optimal speed, and find it interesting”. This definition is quite general and emphasizes the impact of diverse factors on the ease of understanding.

[3] http://www.psy.uwa.edu.au/mrcdatabase/uwa_mrc.htm [27th November, 2012]

Despite the composite, multi-construct character of comprehensibility, most traditional readability indexes such as Flesch Reading Ease [12] or Coleman Liau Index [5] are calculated by the category of simple, syntactic measures such as the count of syllables, words and the length of sentences. Flesch Reading Ease, one of the earliest standard measures, defines readability as a function of word length and sentence length. This approach is simple to implement but sensitive to outliers. Another category of metrics focuses on difficulty levels of words themselves to achieve a more robust measure. A typical approach here is to use a predefined list of common or easy words (e.g., New Dale-Chall Formula [4]). However, due to a dynamic character of language, the static list requires continuing updates.

Other studies cast the problem to classifying texts according to reading levels. For example, Collins-Thompson and Callan [6] use statistical language modeling and multinomial Naïve Bayes classification for readability measure of English and French. This line of research exploits various features ranging from surface text ones (e.g., word length) to discourse-level features (e.g., the number of entities involved in a text) and from a manually compiled list of vocabulary to statistical language modelling. Contextual approach to readability estimation was proposed by Akamatsu *et al.* [1]. The authors employed biased random walk for finding easy or difficult pages on the Web based on the hypothesis that pages with similar comprehensibility levels are relatively more likely to be linked with each other than those with very different levels.

Other studies have investigated automatic and adaptive models [7] for personalizing web search results by their readability levels. However, none of the above works focused explicitly on estimating abstractness and concreteness qualities of terms or documents.

2.3 Information Retrieval and Concreteness

To the best of our knowledge, concreteness has not been explicitly incorporated into document retrieval, although some works proposed including readability measures into web search or returning content depending on user age and thus related to his or her mental capabilities.

Eickoff *et al.* [10] classified web pages into those suitable for children and those not following the topical and non-topical approach. They employed a range of diverse features such as common readability measures, presentation styles, numbers of multimedia on pages, link ratios, etc. Nakatani *et al.* [20] measured word detailedness and topicality based on analysis of links in Wikipedia in order to employ them in web search. The main hypothesis was that a Wikipedia article about a detailed word has most of their inlinks originating from articles in the same category as the one of the target article. Coh-Metrix [14] is a tool that comprehensively analyzes texts based on various measures of cohesion, language and readability. Abstractness and text vagueness are discussed in the context of this work. Although average imageability and concreteness scores of texts are provided by the system, these are based on the fixed list of judged words obtained from MRCDB database; hence, the scores are not algorithmically determined.

Considering image retrieval, current image search engines return results which are of good quality when the query contains concrete terms. On the other hand, search results for abstract query such as "summer" or "happiness" are often not satisfactory. To improve image search results for an abstract term query, Kato *et al.* [16] proposed to retrieve more concrete terms which are associated with the input abstract concept and then to substitute them for this concept. Sun and Bhowmick [25] proposed *image tag visual clarity* to evaluate the effectiveness of a tag in describing the visual content of its annotated images. The objective was to improve the effectiveness of image search engines. Larson *et al.* [19] predicted sizes of visually-depicted objects in images based on mining linguistic patterns within web search results such as "the <tag> in his hand", "the <tag> in her pocket", "the <tag> on the horizon", etc. where <tag> is a given object.

3. MEASUREMENT OF CONCRETENESS

In this section, we describe features used to estimate the concreteness of terms with SVM regression model and then propose the measure for document concreteness. Note that in the current implementation we have focused only on nouns which are key information bearing elements of text.

3.1 Estimating Concreteness of Terms

Automatically judging how much terms are concrete and abstract is not trivial. An intuitive approach would be a topic-based one by assuming that common terms in documents, for example, on psychology or philosophy tend to be abstract. However, this solution would not be practical as it would require large datasets of annotated documents on high number of diverse topics. In addition, not all documents and terms in abstract domains are actually abstract. We decided to follow a non-topical approach and determine term concreteness/abstractness based on their non-topical, and hence, more general features. We selected 21 features grouped in 8 categories: (1) visual representativeness and popularity, (2) diversity of annotations, (3) co-occurrence with sense verbs, (4) number of senses, (5) depth in ontology tree, (6) number of hyponyms, (7) sentiment level, and (8) term length. We describe each group in turn.

Visual Representativeness and Popularity
Imageability of terms indicates how easily and quickly people can imagine the referent of terms. To estimate word's imageability we propose to analyze the degree of *visual representativeness* of a term, which we define as the extent to which the term is used for describing photos or, in general, images. The reasoning behind this hypothesis is that words used frequently to describe photos or images have high probability to be concrete, since the photos and images usually show concrete objects as perceived by a sense of vision. We necessarily make an assumption here that in most of the cases photos or images are annotated with terms that represent the displayed objects. We measure the popularity of using a given word to annotate photos, $\text{Popularity}_{\text{photo}}(t)$ and images, $\text{Popularity}_{\text{image}}(t)$. In addition, we also capture the word's popularity on the web, $\text{Popularity}_{\text{web}}(t)$, according to the hypothesis that common words could represent concrete objects which people see in everyday life.

In particular, we measure the frequency of a term t in the search index of the Bing web search[4], $\text{Freq}_{\text{web}}(t)$, the frequency of a term t in the Bing image search, $\text{Freq}_{\text{image}}(t)$, and the frequency of a term t in Flickr[5], $\text{Freq}_{\text{photo}}(t)$. As these frequencies range vastly we represent the features using

[4] http://www.bing.com [27th November, 2012]
[5] http://www.flickr.com [27th November, 2012]

logarithms.

$$\begin{array}{lll} \textbf{f1} & \text{Popularity}_{\text{web}}(t) & = \log_{10}(\text{Freq}_{\text{web}}(t)+1) \\ \textbf{f2} & \text{Popularity}_{\text{image}}(t) & = \log_{10}\left(\text{Freq}_{\text{image}}(t)+1\right) \\ \textbf{f3} & \text{Popularity}_{\text{photo}}(t) & = \log_{10}\left(\text{Freq}_{\text{photo}}(t)+1\right) \end{array}$$

Since some words, especially, the popular ones, could be used for annotating images just by chance, we also normalized $\text{Freq}_{\text{image}}(t)$ and $\text{Freq}_{\text{photo}}(t)$ by the popularity of words in text domain approximated by their popularity on the web, $\text{Popularity}_{\text{web}}(t)$. Then, additional features **f4** and **f5** are calculated as below.

$$\begin{array}{lll} \textbf{f4} & \frac{\text{Popularity}_{\text{image}}(t)}{\text{Popularity}_{\text{web}}(t)} & = \frac{\log_{10}\left(\text{Freq}_{\text{image}}(t)+1\right)}{\log_{10}(\text{Freq}_{\text{web}}(t)+1)} \\ \textbf{f5} & \frac{\text{Popularity}_{\text{photo}}(t)}{\text{Popularity}_{\text{web}}(t)} & = \frac{\log_{10}\left(\text{Freq}_{\text{photo}}(t)+1\right)}{\log_{10}(\text{Freq}_{\text{web}}(t)+1)} \end{array}$$

The last two features can be interpreted as the degrees of word's visual representativeness, or, the rate of word's popularity in the visual vs. textual domain.

Diversity of Annotations

In shared image databases users usually add annotations for stored contents. We hypothesize that when many diverse annotations are added for photos related to a given term t, the term might be abstract. This is because an abstract concept can have many interpretations and instantiations, and, often, there are several ways to represent its meaning. For example, a tag "happiness" could be used for annotating an image of a smiling child playing with a dog, a couple walking together, a person driving elegant car, a luxurious house and so on. We describe the number of annotations which are added for contents related to a term t as $\#\text{Annotations}(t)$ and the corresponding number of unique annotations as $\#\text{Annotations}_{\text{uniq}}(t)$.

For capturing these measures, we used social tagging data derived from Flickr. Since in Flickr users sometimes submit multiple photos with exactly same tags due to their similar theme or just for convenience, we applied a simple filtering scheme. A given photo is skipped if there is another photo submitted by the same user that has identical tag set.

After applying the filtering we took up to 500 top ranked photos from the search results for a query t, denoted as $\text{Photos}(t) = \{photo_1,\ photo_2,\ \cdots,\ photo_n\}$ where $0 \leq n \leq 500$. The number of tags of photos in $\text{Photos}(t)$ is $\#\text{Tags}(t)$ and the number of unique tags is $\#\text{Tags}_{\text{uniq}}(t)$. Note that tags containing stop words and the search query t itself were eliminated. The features **f6** and **f7** were calculated as follows.

$$\begin{array}{lll} \textbf{f6} & \#\text{Annotations}(t) & = \dfrac{\#\text{Tags}(t)}{n} \\ \textbf{f7} & \#\text{Annotations}_{\text{uniq}}(t) & = \dfrac{\#\text{Tags}_{\text{uniq}}(t)}{n} \end{array}$$

Co-occurrence with Sense Verbs

As mentioned before, concreteness is often defined in terms of perceivability. Concrete terms should commonly occur with verbs which denote senses such as *see*, *hear*, or *taste*. We use verbs representing sensory experience, called sense verbs, to represent 5 basic senses: sight, hearing, taste, smell, and touch. The average co-occurrence rate of a term t with the sense verbs is defined as $\text{SenseVerbs}_{\text{avg}}(t)$ and is measured using the window size of a single sentence.

For instance, let a set of verbs related to sight be $\text{Verbs}_{\text{sight}} = \{\text{see, sees, saw, seen}\}$. The co-occurrence between a term t and verbs related to sight is calculated as:

$$\text{Cooc}_{\text{sight}}(t) = \sum_{v \in Verbs_{\text{sight}}} \frac{\text{Freq}_{\text{web}}(``v * t")}{\text{Freq}_{\text{web}}(v \vee t)}$$

In the implementation, we use phrase search. An asterisk * matches more than 0 terms or does not match any terms in the same sentence. $\text{Cooc}_{\text{hearing}}$, $\text{Cooc}_{\text{taste}}$, $\text{Cooc}_{\text{smell}}$, and $\text{Cooc}_{\text{touch}}$ are calculated in a similar way.

Note however that some terms co-occur with only one sense. For example, a noun "photo" often co-occurs with a verb *see*, while rarely co-occurring with the other sense verbs. We thus use not only the average but also the maximum co-occurrence rating of a term t with sense verbs denoted as $\text{SenseVerbs}_{\text{max}}(t)$. The features **f8** and **f9** are calculated as follows.

$$\begin{array}{lll} \textbf{f8} & \#\text{SenseVerbs}_{\text{max}}(t) & = \max(\text{Cooc}_{\text{sight}}(t),\ \text{Cooc}_{\text{hearing}}(t), \\ & & \quad \text{Cooc}_{\text{taste}}(t),\ \text{Cooc}_{\text{smell}}(t), \\ & & \quad \text{Cooc}_{\text{touch}}(t)) \\ \textbf{f9} & \#\text{SenseVerbs}_{\text{avg}}(t) & = \frac{1}{5}(\text{Cooc}_{\text{sight}}(t) + \text{Cooc}_{\text{hearing}}(t) \\ & & \quad +\text{Cooc}_{\text{taste}}(t) + \text{Cooc}_{\text{smell}}(t) \\ & & \quad +\text{Cooc}_{\text{touch}}(t)) \end{array}$$

Number of Senses

A term often has more than one meaning. For example, according to WordNet [6], the nouns "reason" and life have 6 and 14 senses, respectively, while "tree" and "plant" have 3 and 4 senses, respectively.

We hypothesize that when the number of senses of a term t is high, t might be abstract. Hence we propose the number of senses obtained from WordNet, $\#\text{Senses}(t)$, that a term t has as another feature (**f10**).

Depth in Ontology Tree

The relation of concreteness of terms with their depth in an ontology tree has been pointed out in [11]. Similar idea was investigated in [18] for the purpose of subjectivity classification. We consider ontology to be a hierarchical structure constructed of relations such as is-a and part-of relations. Note that the depth of a term t in the ontology is usually different for its different senses. For example, the 4 senses of a noun "plant" in the WordNet ontology tree have depths of 7, 5, 9, and 11. We denote the depth of the most frequently used sense of a term, t as $\text{Depth}_{\text{freq}}(t)$ (**f11**) and the average depth of the senses of a term t as $\text{Depth}_{\text{avg}}(t)$ (**f12**). Both mean the distance in the ontology tree from the root to a target sense.

Number of Hyponyms

The number of hyponyms a term has appears to be related to the level of its generality. We hypothesize that when the number of hyponyms which a term t has is large, t might be abstract. Same as for the features based on the depth in the ontology tree, the number of hyponyms also depends on different senses of a term. We thus use two features, the number of hyponyms of the most frequently used sense, $\#\text{Hyponyms}_{\text{freq}}(t)$ (**f13**) and the average number of hyponyms of all the senses of t, $\#\text{Hyponyms}_{\text{avg}}(t)$ (**f14**).

Sentiment Level

Intuitively, abstract terms tend to arouse positive or negative sentiments. For example, according to SentiWordNet[7], which is a lexical resource providing sentiment values to

[6] http://wordnet.princeton.edu/ [27th November, 2012]

[7] http://sentiwordnet.isti.cnr.it/ [27th November, 2012]

Wordnet synsets, “opportunity” arouses positive sentiment, whereas “regret” arouses negative sentiment. On the other hand, concrete terms (e.g., “tree”, “road”) might be more objective and to lesser extent associated with sentiment than abstract ones. We select positivity, negativity and objectivity values of terms as other features for SVM regression. Same as for some of the previously introduced features, the values can differ for different term senses. We thus define the positivity, negativity and objectivity values of the most frequently used sense of a term t as $\mathrm{Positivity}_{\mathrm{freq}}(t)$ (**f15**), $\mathrm{Negativity}_{\mathrm{freq}}(t)$ (**f16**), and $\mathrm{Objectivity}_{\mathrm{freq}}(t)$ (**f17**), respectively. The average positivity, average negativity, and average objectivity values of the senses of a term t are represented by $\mathrm{Positivity}_{\mathrm{avg}}(t)$ (**f18**), $\mathrm{Negativity}_{\mathrm{avg}}(t)$ (**f19**), and $\mathrm{Objectivity}_{\mathrm{avg}}(t)$ (**f20**), respectively.

Term Length

According to the rule of the least-effort in human communication as proposed by Zipf [30], the most frequent words tend to be short. These are often words describing common objects surrounding humans. In English, many abstract nouns are formed by adding noun-forming suffixes (-ness, -ity, -tion, -ism) to adjectives, verbs or other nouns (e.g. “happiness”, “circulation”, “serenity” and “communism”). We thus set a hypothesis that the longer the number of characters in a term, the more abstract the term might be. The number of characters of a term t, #Characters (t), is selected as feature **f21**.

3.2 Estimating Concreteness of Documents

In this section we estimate the concreteness of documents by using the calculated concreteness levels of terms. We propose two methods: (1) *average concreteness* and (2) *maximum concreteness*. The first one is based on the hypothesis that concrete documents contain many concrete terms. Let, $Conc(t)$ be the concreteness score of term t, D be a document and $|D|$ be the number of terms in D:

$$\mathrm{Conc}_{\mathrm{doc}}^{\mathrm{avg}}(D) = \frac{1}{|D|}\sum_{t\in D}\mathrm{Conc}\,(\mathrm{t}) \tag{1}$$

The second method, *maximum concreteness*, is based on the assumption that documents consist of abstract paragraphs and concrete paragraphs. The intuition behind this method is that often a short, yet concrete paragraph, such as the one with concrete examples or instances of abstract concepts makes a difference between document understanding and confusion.

The estimation of document concreteness consists then of 2 steps: (1) calculating the concreteness of paragraphs in a document, and (2) using the calculated paragraph scores to derive the concreteness of the document.

Let a document D be considered as a set of paragraphs: $D = \{P_1,\ P_2,\ \cdots,\ P_L\}$ where P_i is a paragraph in D. We then represent P_i as a sequence of terms $P_i = (t_1,\ t_2,\ \cdots,\ t_M)$. For each paragraph P in D we calculate the concreteness of P as follows.

$$\mathrm{Conc}_{\mathrm{par}}(P) = \frac{1}{|P|}\sum_{t\in P}\mathrm{Conc}\,(\mathrm{t}) \tag{2}$$

Finally, the concreteness of a document D is estimated as follows.

$$\mathrm{Conc}_{\mathrm{doc}}^{\max}(D) = \max\left(\mathrm{Conc}_{\mathrm{par}}(P_1),\ \cdots,\ \mathrm{Conc}_{\mathrm{par}}(P_N)\right)$$

4. EVALUATION OF TERM-LEVEL CONCRETENESS

4.1 Dataset and Evaluation Metrics

We use the Medical Research Council Psycholinguistic Database (MRCDB) version 2.0 as a dataset and WordNet 3.0, Flickr, and SentiWordNet 3.0 as external knowledge bases. MRCDB contains 150,837 terms and provides information on 26 different linguistic properties, although, information about every property is not available for each term. In MRCDB 8,288 terms have perceivability ratings[8], and 9,240 terms have imageability ratings. The ratings were derived from merging 3 sets of norms: Paivio [22], Toglia and Batting [27], and Gilhooly and Logie [13]. The 3 sets of norms correlated highly and were merged by adjusting both the means and standard deviations. The scores were obtained from two groups of evaluators and are expressed as integer values ranging in the interval [100, 700]. 28 evaluators (12 males) estimated perceivability, and 30 evaluators (15 males) estimated imageability. Perceivability was rated in 7-point numerical scale from 1 (one cannot experience the referent of a target term by his or her sense) to 7 (one can easily experience the referent of a target term, such as objects, materials, and people). Similarly, the imageability of terms was rated in 7-point numerical scale from 1 (one cannot imagine the referent of a target term quickly and easily) to 7 (one can easily and quickly imagine the referent of a target term). In the experiments, we used nouns which have perceivability rating and imageability rating in MRCDB, and which, at the same time, are contained in WordNet as nouns. In total, there were 3,455 nouns satisfying these conditions. We then used both the perceivability ratings and the imageability ratings of these terms as the target values for SVM regression. As the ratings range in the interval [100, 700] we have normalized them to fit into the interval [0, 1] by min-max normalization.

As mentioned before, the earlier studies defined concreteness by two psycholinguistic attributes, perceivability and imageability. Therefore we define the concreteness as follows:

$$\mathrm{Conc}_{\mathrm{MRC}}(t) = \frac{1}{2}\left\{\mathrm{Perc}_{\mathrm{MRC}}(t) + \mathrm{Imag}_{\mathrm{MRC}}(t)\right\}$$

$\mathrm{Perc}_{\mathrm{MRC}}(t)$ is the perceivability rating of a term t in MRCDB, and $\mathrm{Imag}_{\mathrm{MRC}}(t)$ is the imageability rating of t in MRCDB. Table 1 shows the minimum rating, the maximum rating, the average rating, and the standard deviation of perceivability, imageability rating and the combined rating of 3,455 nouns in MRCDB. The Pearson’s correlation coefficient between the perceivability and imageability ratings of 3,455 nouns in MRCDB is 0.826. This represents a strong positive correlation between the both measures.

Table 1: The statistics of 3,455 nouns in MRCDB used for the experiment.

Measure	min	max	Avg.	S.D.
Perceivability	0.097	0.950	0.578	0.195
Imaginability	0.048	0.945	0.620	0.175
Concreteness	0.152	0.923	0.609	0.177

[8] In fact, perceivability scores are termed “concreteness” in MRCDB although their sense is closer to the notion of perceivaility as indicated in the description of user study.

In order to evaluate the correlation between two lists of values, we used three measurements: Pearson's correlation coefficient, Kendall's τ, and Root Mean Square Error (RMSE). The values for Pearson's correlation coefficient and Kendall's τ range in the interval [-1, 1]. Large or small values imply that there is positive or negative correlation between two input lists, while a value of 0 implies the lack of such correlation. Values for RMSE are more than or equal to 0. Small RMSE values imply small differences between the input value pairs.

4.2 Results

We used SVM$^{\text{light}}$ [15] implementation with standard parameterization for SVM regression. SVM was trained first on imageability scores and then on perceivability scores.

We denote the perceivability and imageability scores of term t estimated by using SVM regression as $\text{Perc}_{\text{SVM}}(t)$ and $\text{Imag}_{\text{SVM}}(t)$, respectively. We then estimate the concreteness of a term t, $\text{Conc}_{\text{SVM}}(t)$ using perceivability ratings and imageability ratings estimated by SVM regression.

$$\text{Conc}_{\text{SVM}}(t) = \frac{1}{2}\left(\text{Perc}_{\text{SVM}}(t) + \text{Imag}_{\text{SVM}}(t)\right) \quad (3)$$

Table 2 shows the results. The correlations between $\text{Perc}_{\text{MRC}}(t)$ and $\text{Perc}_{\text{SVM}}(t)$ measured by Pearson's correlation coefficient, Kendall's τ, and the Root Mean Square Error are 0.671, 0.492, and 0.145, respectively. The correlations between $\text{Imag}_{\text{MRC}}(t)$ and $\text{Imag}_{\text{SVM}}(t)$ are 0.675, 0.502, and 0.129, respectively. These results indicate reasonably strong positive correlation. It also appears that imageability can be estimated slightly more accurately. However, we managed to achieve better results on the concreteness measure. The correlations between the concreteness of terms using MRCDB, $\text{Conc}_{\text{MRC}}(t)$, and the proposed concreteness, $\text{Conc}_{\text{SVM}}(t)$, are 0.688, 0.508, and 0.128, respectively. These results indicate that our method can successfully estimate word concreteness scores as compared to the labelled data in MRCDB. For better understanding of the results, in Table 4 we show 10 terms which have the largest or smallest perceivability, imageability and the concreteness rating estimated by using SVM regression.

Table 2: Results of term-level concreteness estimation

Measure	Pearson's r	Kendall's τ	RMSE
Perceivability	0.671	0.492	0.145
Imageability	0.675	0.502	0.129
Concreteness	0.688	0.508	0.128

Next, we looked at the statistics of the scores generated by our approach. Table 3 shows the statistics of $\text{Perc}_{\text{SVM}}(t)$, $\text{Imag}_{\text{SVM}}(t)$ and the concreteness score, $\text{Conc}_{\text{SVM}}(t)$. We notice that the perceivability ratings estimated by SVM regression have smaller standard deviation than the target ratings in MRCDB (see Table 1). Similar smoothing effect is observed for imageability.

Table 3: The statistics of SVM regressions results

Measure	min	max	Avg.	S.D.
Perceivability	-0.016	1.010	0.601	0.132
Imaginability	0.105	1.017	0.623	0.116
Concreteness	0.085	1.001	0.612	0.122

Table 4: Top 10 terms having largest or smallest ratings estimated by SVM regression

Terms with the largest perceivability scores, $\text{Perc}_{\text{SVM}}(t)$
cattle, portrait, deer, cow, shrub, tree, robin, boat, moose, feline
Terms with the smallest perceivability scores, $\text{Perc}_{\text{SVM}}(t)$
competence, exactitude, ingratitude, comradeship, insufficiency, importance, integrity, infallibility, virtue, misconception
Terms with the largest imageability scores, $\text{Imag}_{\text{SVM}}(t)$
portrait, cattle, boat, sunlight, cow, grass, dusk, deer, sunset, mare
Terms with the smallest imageability scores, $\text{Imag}_{\text{SVM}}(t)$
besieger, exactitude, disparagement, ingratitude, increment, impotency, competence, profiteer, interposition, loquacity
Terms with the largest concreteness scores, $\text{Conc}_{\text{SVM}}(t)$
portrait, cattle, cow, boat, deer, robin, tree, grass, dusk, sunlight
Terms with the smallest concreteness scores, $\text{Conc}_{\text{SVM}}(t)$
exactitude, competence, ingratitude, insufficiency, besieger, infallibility, impotency, condescension, loquacity, comradeship

4.2.1 Feature Selection

We evaluate the importance of each feature with 5-fold cross-validation and 3 measurements: Pearson's correlation coefficient, Kendall's τ, and RMSE. The mechanism of feature selection is based on finding the least effective feature $\mathbf{f}$ in the set of all features $Features$, and then removing $\mathbf{f}$ from $Features$ step by step.

Table 5 lists the transition of the three evaluation metrics during the feature selection. The values in step 0 are the values which are estimated by SVM regression using all the 21 features. Each column in Table 5 shows the pairs of removed feature and the measured value of our evaluation metrics.

Feature **f20** (the average objectivity of the senses of a term) is the most important feature used in SVM regression for perceivability, and feature **f3** (the rating how much a term is related to photos) is the most important feature for imageability. Features **f3** and **f11** (the depth of the most common sense) are important for both perceivability and imageability. On the other hand, features based on capturing average values over different senses of a term such as **f9**, **f14** are not very helpful. The performance should thus improve if the actual sense of a word is considered. Surprisingly, **f5** (the ratio of popularity of a term in Flickr vs. popularity on the web) performs poorly, while a very similar feature **f4** (the ratio of popularity of a term in image search engine vs. popularity on the web) is quite significant. This is probably due to the fact that shared photo databases such as Flickr do not contain large variety of images when compared to the data captured by image search engines. For example, images of tourist places tend to be more common in Flickr than the images of less interesting objects such as hammer or sponge.

4.3 Term Co-occurrence and Concreteness

We complete the analysis of term-level concreteness by investigating another possible approach to its evaluation.

We expected a tendency that abstract nouns co-occur with other abstract nouns more frequently than with concrete nouns. Similarly, we also suspected that concrete nouns co-occur with concrete nouns more frequently than with abstract nouns. If these assumptions would be true, then the near context of a term could serve as an additional signal for concreteness estimation. To investigate the hypothesis regarding the correlation between the concreteness of terms and their co-occurrence we used the article content of En-

Table 5: The change in average Pearson's correlation coefficient, Kendall's τ, and RMSE

step	$\mathrm{Perc}_{\mathrm{SVM}}(t)$						$\mathrm{Imag}_{\mathrm{SVM}}(t)$					
	Pearson		Kendall		RMSE		Pearson		Kendall		RMSE	
0	–	0.668	–	0.489	–	0.145	–	0.671	–	0.499	–	0.130
1	**f5**	0.669	**f14**	0.490	**f5**	0.145	**f5**	0.673	**f6**	0.499	**f5**	0.129
2	**f14**	0.669	**f17**	0.490	**f14**	0.145	**f1**	0.673	**f20**	0.499	**f1**	0.129
3	**f16**	0.669	**f8**	0.490	**f16**	0.145	**f9**	0.673	**f8**	0.499	**f19**	0.129
4	**f19**	0.669	**f9**	0.490	**f19**	0.145	**f14**	0.673	**f1**	0.500	**f17**	0.129
5	**f8**	0.669	**f2**	0.490	**f8**	0.145	**f13**	0.673	**f5**	0.500	**f8**	0.129
6	**f9**	0.670	**f18**	0.490	**f9**	0.145	**f20**	0.673	**f9**	0.500	**f13**	0.129
7	**f2**	0.670	**f15**	0.490	**f2**	0.145	**f18**	0.674	**f17**	0.500	**f14**	0.129
8	**f6**	0.669	**f5**	0.490	**f18**	0.145	**f8**	0.674	**f13**	0.500	**f6**	0.129
9	**f18**	0.669	**f6**	0.489	**f17**	0.145	**f19**	0.674	**f15**	0.500	**f16**	0.129
10	**f17**	0.669	**f19**	0.488	**f6**	0.145	**f6**	0.674	**f14**	0.500	**f9**	0.129
11	**f13**	0.667	**f13**	0.487	**f13**	0.146	**f15**	0.674	**f19**	0.500	**f15**	0.129
12	**f10**	0.665	**f10**	0.485	**f10**	0.146	**f21**	0.673	**f12**	0.498	**f21**	0.129
13	**f15**	0.663	**f16**	0.483	**f15**	0.146	**f17**	0.671	**f18**	0.496	**f18**	0.130
14	**f12**	0.660	**f12**	0.479	**f12**	0.147	**f12**	0.669	**f4**	0.492	**f12**	0.130
15	**f7**	0.652	**f7**	0.470	**f7**	0.148	**f10**	0.666	**f21**	0.489	**f4**	0.130
16	**f21**	0.635	**f21**	0.459	**f21**	0.151	**f4**	0.664	**f10**	0.486	**f10**	0.131
17	**f4**	0.619	**f4**	0.443	**f4**	0.153	**f16**	0.653	**f16**	0.474	**f20**	0.132
18	**f1**	0.592	**f1**	0.422	**f1**	0.157	**f7**	0.631	**f7**	0.458	**f7**	0.136
19	**f11**	0.532	**f11**	0.359	**f11**	0.166	**f2**	0.608	**f11**	0.431	**f2**	0.139
20	**f3**	0.415	**f3**	0.287	**f3**	0.196	**f11**	0.344	**f2**	0.230	**f11**	0.448
21	**f20**	–	**f20**	–	**f20**	–	**f3**	–	**f3**	–	**f3**	–

glish Wikipedia [9]. In particular, we examined the relation between the co-occurrence of terms in the same sentences and their concreteness ratings as provided in MRCDB.

Let #Sentences (t) be the number of sentences which contain a term t. The Jaccard Coefficient between any two terms, t and u, is calculated by the following equation.

$$\text{Jaccard}\,(t,\ u) = \frac{\#\text{Sentences}\,(t \wedge u)}{\#\text{Sentences}\,(t \vee u)}$$

To proceed we make two sets that include adjustable portions of terms according to their perceivability or imageability scores (terms are ranked by their scores):

$$Terms_{\text{high}} = \{\text{terms in highest } x\%\}\,,$$
$$Terms_{\text{low}} = \{\text{terms in lowest } x\%\}\,.$$

The co-occurrence rate between the two sets of terms, $Terms_i$ and $Terms_j$, is then calculated as the average Jaccard coefficient between terms from these two sets.

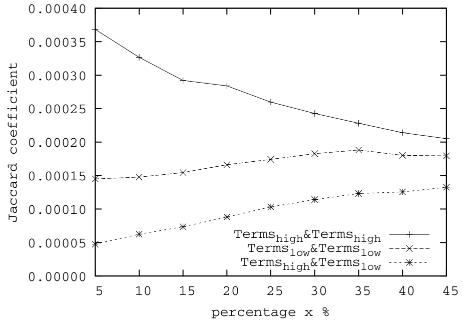

Figure 1: Average Jaccard Coefficient results of two sets of terms for perceivability

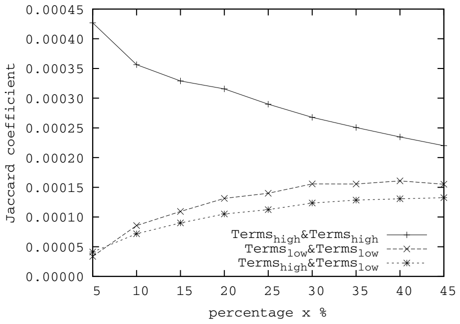

Figure 2: Average Jaccard Coefficient results of two sets of terms for imageability

Figures 1 and 2 show the results according to the used rate of top-scored terms in the two term sets (i.e., x%). We can observe significant differences between term co-occurrence levels at different sizes of term sets. For example, at $x =$ 5 the average co-occurrence between the little perceivable terms or between the highly perceivable terms is 3 and 7 times higher, respectively, than the average co-occurrence between terms in which one has high and the other has low perceivability scores. These results suggest that concrete terms may tend to co-occur with other concrete terms more than with abstract ones. The co-occurrence score of a term with manually labelled examples of concrete terms could be then another feature for SVM.

5. EVALUATION OF DOCUMENT-LEVEL CONCRETENESS

Although we obtained satisfactory results the estimation of term-level concreteness, we wanted to investigate whether it is feasible to estimate the concreteness of the whole documents based on the concreteness scores of their contained terms. In this section we report the results of the second experiment to measure the accuracy of document-level concreteness estimation.

5.1 Settings

To create dataset for the second experiment we have collected web pages from the Yahoo! web search engine[10]. We have used 50 queries, among which, 33 queries were manually selected from TREC Million Query Track 2007, 2008 and 2009 [3], and 17 were selected from the titles of

[9] Wikipedia dump of July 2011.

[10] http://www.yahoo.com [27th November, 2012]

Wikipedia articles. We have tried to collect both abstract (e.g., "rationalism", "good moral character") and concrete queries (e.g., "copper penny", "apple laptops") in order to have documents with different concreteness degrees. In total, we had 35 abstract and 15 concrete queries. The top 10 returned search results were collected for each query. Next, we removed HTML tags, Javascript and any multimedia files, and captured the core content of each document. Finally, we extracted two top paragraphs filtering out the ones with less than 400 characters.

Such processed documents were then shown to 5 evaluators who rated them in terms of concreteness and comprehensibility using the 5-point Likert scale ("very low", "low", "neutral", "high", and "very high"). The evaluators were in their 20's and 30's, had higher education degrees and were not experts in any of the topics of test queries. Besides the concreteness and comprehensibility judgements, the evaluators were also asked to rate the relevance of each document to its associated query. By this we wanted to eliminated spam or completely irrelevant documents. The relevance choices were as follows: "irrelevant (spam)", "neutral" and "relevant". After the documents have been evaluated, we removed those that were judged as irrelevant or neutral. The criteria for removal were as follows, (a) a document received at least one "irrelevant (spam)" vote from any of five judges or (b) at least three judges labelled the document as "neutral". After this step we had in total 305 labelled documents. Note that we have decided to choose a relatively high number of judges rather than extending the size of document collection according to the intuition that both concreteness and comprehensibility are subjective and may depend on user's knowledge. The inter-evaluator agreement of the scores assigned by the five judges was measured by average Kendall's τ coefficient. The result was 0.568 for concreteness and 0.624 for comprehensibility indicating reasonably high agreement among judges. To assign a single concreteness and comprehensibility score to each document we took the median of five scores provided by the evaluators.

5.2 Results

5.2.1 Correlation between Concreteness and Comprehensibility

We first investigate our initial assumption that concrete texts help users better understand documents. We show the results of the correlation between the ground truth scores (i.e., the median of the scores by five judges) of concreteness and comprehensibility. The Kendall's τ correlation coefficient between the concreteness and comprehensibility is 0.385 ($p < 0.0001$). This is moderate positive correlation indicating that concreteness and comprehensibility of documents are somehow correlated, although the correlation is not very high. Moreover, we have found that concrete documents are much more likely to be comprehensible than abstract ones as evidenced by the conditional probabilities, $P(\text{High Comp}|\text{High Conc}) = 0.971$, and $P(\text{High Comp}|\text{Low Conc}) = 0.491$. Similarly, incomprehensible documents are less likely to be concrete than comprehensible ones, $P(\text{High Conc}|\text{Low Comp}) = 0.067$, and $P(\text{High Conc}|\text{High Comp}) = 0.723$. High Conc and High Comp indicate here the scores, "high", and "very high", while Low Conc and Low Comp indicate the scores, "low", and "very low" for the concreteness and comprehensibility, respectively. These results indicate the usefulness of concrete texts when it comes to content understanding.

In addition, we investigated correlation between readability metrics and comprehensibility. By this we wanted to check whether using readability measures only would be enough to determine document comprehensibility. We have used here the following measures, Flesch Reading Ease (FRE) [12], New Dale Chall (NDC) [4] and Lexical Density (LD) [28]. As mentioned before, FRE calculates document readability based on the combined measure of sentence and word length, while NDC estimates the expected school grade level of a text based on the average sentence length and the number of "difficult" words. To distinguish "easy" words from difficult ones NDC uses the list of 3,000 common words in English. Lastly, LD estimates the rate of content words (e.g., nouns, adjectives, verbs, etc.) to the total number of words in a document including grammatical words. Intuitively, documents with high lexical density should contain much information (e.g., academic papers) and may thus be poorly understandable by readers. To calculate LD we have used POS tagger available in the Natural Language Toolkit[11] (NLTK).

Table 6 shows the correlation between documents' comprehensibility and their readability. We can see that the assumption that readability formulas only such as FRE and NDC can successfully estimate how easy documents are is generally not correct. When compared to the above-reported correlation between the document concreteness and comprehensibility, these results suggest that concreteness is an important factor of document comprehensibility that cannot be simply captured by the surface document characteristics such as sentence length, word length, or POS tag distribution.

Table 6: Kendall's τ coefficient between the readability and the comprehensibility. A parenthetical value indicates the p value.

$\text{Conc}_{\text{doc}}^{\text{gt}}$	Readability Measures		
	FRE	NDC	LD
0.385 (0.000)	0.215 (0.000)	0.335 (0.000)	-0.111 (0.003)

5.2.2 Evaluation of Document Concreteness Measure

For evaluation we use two baselines, *name entity score* and *generality score*. The first one calculates the rate of named entities in text using NLTK toolkit according to the intuitive assumption that named entities usually indicate concrete entities. The second baseline calculates average term commonness in a document using a large English corpus. It is expressed as follows:

$$\text{Generality}\,(D) = \frac{1}{|D|}\sum_{t \in D} \ln \text{cf}\,(t) \tag{4}$$

where cf (t) is the term frequency of t in the Corpus of Contemporary American English[12] (COCA). COCA is a balanced, up-to-date corpus containing about 450 million words from documents of diverse genres. The intuition behind this choice is the assumption that documents containing popular words should be on average more concrete than the ones

[11] http://nltk.org [27th November, 2012]

[12] http://corpus.byu.edu/coca [27th November, 2012]

with less popular words. For example, a blog describing someone's daily life or dining experience should contain on average more common words than scientific paper on astrophysics or a discourse about philosophy. We are aware that this hypothesis may not be always true and there are still many documents that describe abstract topics using relatively popular words (e.g., "love", "sadness"). We note that as our work is the first attempt for automatic estimation of document concreteness we could not use any standard evaluation benchmark or other similar systems.

First, we calculate the correlation of the concreteness scores given by judges and the scores assigned by $Conc_{doc}^{avg}$ and $Conc_{doc}^{max}$, as described in Section 3.2. The results are shown in Table 7 and indicate that the proposed methods produce significantly better results than the baselines. Interestingly, the *name entity scoring* is characterized by the negative correlation with the concreteness scores. Thus it is not necessarily true that documents with many named entities are concrete. The results of the Mean Average Precision and nDCG measures (see Table 8) confirm also the higher effectiveness of our approach when compared to the baselines. However, we notice from these results that it is difficult to conclude which method, $Conc_{doc}^{avg}$ or $Conc_{doc}^{max}$, performs better.

For a more complete analysis we show the 11-points interpolated precision-recall graph in Figure 3. To plot the grap, first, we ranked all the documents based on the calculated concreteness score. Then we considered documents having "high" or "very high" median score as *true documents.* Looking at the Figure 3 we can notice that the precision at 10% recall drops significantly, although it is still relatively high (over 60%). We can explain it as an increased difficulty to predict correct concreteness for the documents that did not achieve top ranks. In other words, it is relatively easy to determine the correct concreteness levels of documents with plenty of concrete terms but more difficult to do it for documents with moderate number of terms. The precision values until the middle recall range remain relatively stable (e.g., Precision = 60% at Recall = 50%).

We note here that the relatively simple extension from the term-level to document-level concreteness produces already good results. One improvement of this approach would involve estimating concreteness levels of phrases and sentences. Consider, for example, the expression "lipstick on a pig". While both the nouns in this expression are concrete, their combination refers to rather abstract concept.

Table 7: Kendall's τ coefficient between the ground truth and concreteness predicted by each method. A parenthetical value indicates the p value.

Baseline methods		Proposed methods	
Named entity score	Generality score	$\mathrm{Conc}_{\mathrm{doc}}^{\mathrm{avg}}$	$\mathrm{Conc}_{\mathrm{doc}}^{\mathrm{max}}$
-0.0255 (0.507)	0.101 (0.009)	0.362 (0.000)	**0.364** (0.000)

Table 8: Average MAP and nDCG over all the queries.

	Baseline methods		Proposed methods	
	Named entity score	Generality score	$\mathrm{Conc}_{\mathrm{doc}}^{\mathrm{avg}}$	$\mathrm{Conc}_{\mathrm{doc}}^{\mathrm{max}}$
MAP	0.666	0.664	**0.796**	0.793
nDCG@1	0.693	0.750	0.830	**0.835**
nDCG@3	0.756	0.810	**0.847**	0.845
nDCG@5	0.826	0.855	0.886	**0.890**

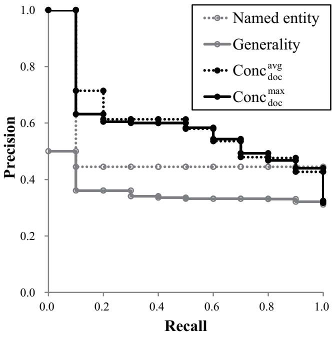

Figure 3: A 11-points interpolated precision curve.

5.3 Estimating Query Concreteness

We are also interested whether there is any correspondence between query concreteness and search results' concreteness. In other words, we wish to know whether abstract queries return on average abstract documents and, correspondingly, whether concrete queries result in on average concrete search results. To test these assumptions we have measured conditional probabilities of the average document concreteness/abstractness (using ground truth scores) depending on the query characteristics (as calculated by our method). The results indicate that documents returned in response to a concrete query are likely to be concrete ($P(ConcD|ConcQ) = 0.779$), while ones returned in response to an abstract query are likely to be abstract ($P(AbstD|AbstQ) = 0.685$). We think that the concreteness value of query itself could be a good predictor of the average concreteness level of the returned search results. This could be used by search engines to dynamically adapt their ranking mechanisms depending on the expected query "difficulty" represented by its concreteness/abstractness score in a similar way as approaches to query performance prediction (relevance viewpoint). This forms our future work.

6. DISCUSSION

6.1 Search Considering Document Concreteness

Conventional search engines do not seem to directly consider concreteness of web pages when ranking search results. Currently, users have to manually find web pages which contain concrete contents. Typically, they do it either by skimming through search results or by reformulating their initial queries to achieve more concrete results. Considering the latter, in the previous example of Parkinson's disease, users could append keywords, such as "experience" and "symptoms" to the initial query "Parkinson's disease". However, the ability to find such keywords depends on users' experience and expertise, hence, novice or inexperienced users may have difficulty with conceptualizing appropriate terms. Imagine a legal document about a newly established law written in a fairy abstract and complicated way. A novice, non-expert user may require texts about concrete, real-life cases to study the law in practice in order to better understand its meaning and application scope. Yet, it may be difficult for the user to come up with keywords leading to pages describing the specific cases of this law. While state-of-the-art web search engines offer query suggestions given the input query string, the returned candidates do not necessarily lead to retrieving more concrete documents. In general, there is no systematic way to retrieve concrete documents neither to support users in finding such documents, for example, by suggesting query extensions that would return more concrete results. We think that incorporating concreteness/abstractness judgements into IR systems could become a valuable enhancement. In the future we plan to propose concreteness/abstractness query expansion model in order to support users in the search for more comprehensible, concrete contents.

6.2 Abstract Documents

We focused in this work on finding concrete web pages. However, concrete documents may be sometimes too long and users might have difficulty to read and memorize them. Certain readers, such as more proficient or knowledgeable ones, may actually prefer to read abstract texts. Although, in general, concrete documents help readers understand the topic of documents, concrete documents lack information which abstract documents contain. For example, think of the documents about the effects of Parkinson's disease. In abstract documents, one of the effects of Parkinson's disease is *motor symptoms*. In concrete documents, the effects of Parkinson's disease are *loss of balance*, *shaking in limbs*, and *slow walk*. In many cases, it is difficult for users to induce abstract rules, theories, or concepts from concrete cases and examples. In the above example, *motor symptoms* covers *loss of balance*, *shaking in limbs* and *slow walk*, yet, many users could have difficulties to abstract into *motor symptoms* from these concrete descriptions. We believe that finding abstract web pages could be also useful sometimes and we do not exclude this case.

7. CONCLUSION AND FUTURE WORK

Concrete contents help to understand abstract, complex concepts, are more easily remembered and are often more attractive to users. However, despite these well-known phenomena, no reasonable solution has been proposed until now for supporting users in finding concrete and thus more comprehensible documents. In this paper, we describe method for evaluating the concreteness of words using machine learning and then extrapolate it to the estimation of concreteness on the document-level. A wide number of signals are explored and the evaluation is conducted on term-level as well as on the document-level. In addition, we discuss the problems and applications of the concreteness estimation in IR and provide additional experimentation for directing further studies. In the future we would like to focus on the previously mentioned directions as well as on the interplay between the concreteness and relevance of documents.

8. ACKNOWLEDGEMENTS

This work has been supported in part by the project: Grants-in-Aid for Scientific Research (No. 24240013) from MEXT of Japan.

9. REFERENCES

[1] K. Akamatsu, N. Pattanasri, A. Jatowt, and K. Tanaka. Comprehensibility of web pages based on link analysis. In *WI'11*, pages 40–46. IEEE, 2011.

[2] L. Cahill and J. L. McGaugh. A novel demonstration of enhanced memory associated with emotional arousal. *Consciousness and Cognition*, 4(4):410–421, 1995.

[3] B. Carterette, V. Pavlu, H. Fang, and E. Kanoulas. Million query track 2009 overview. In *TREC '09*, 2009.

[4] J. Chall. *Readability Revisited: The New Dale-Chall Readability Formula*. Brookline Books/Lumen Editions, 1995.

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

[6] K. Collins-Thompson and J. Callan. A language modeling approach to predicting reading difficulty. In *HLT-NAACL*, pages 193–200, 2004.

[7] K. Collins-Thompson, S. de la Chica, and D. Sontag. Personalizing web search results by reading level. In *CIKM'11*, pages 403–412. ACM, 2011.

[8] D. Crystal. *The Cambridge encyclopedia of the English language*. Cambridge University Press, 1995.

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

[10] C. Eickhoff, P. Serdyukov, and A. P. de Vries. A combined topical/non-topical approach to identifying web sites for children. In *WSDM '11*, pages 505–514.

[11] T. Faaß, L. Kaczmirek, and A. Lenzner. Psycholinguistic determinants of question difficulty: A web experiment. In *7th International Conference on Social Science Methodology*, 2008.

[12] R. Flesch. A new readability yardstick. *Journal of Applied Psychology*, 32(3):221–233, 1948.

[13] K. J. Gilhooly and R. H. Logie. Age-of-acquisition, imagery, concreteness, familiarity, and ambiguity measures for 1,944 words. *Behavior Research Methods*, 12(4):395–427, 1980.

[14] A. C. Graesser, D. S. McNamara, M. M. Louwerse, and Z. Cai. Coh-metrix: analysis of text on cohesion and language. *Behavior Research Methods, Instruments, and Computers*, 36(2):193–202, 2004.

[15] T. Joachims. *Making large-scale support vector machine learning practical*, pages 169–184. MIT Press, Cambridge, MA, USA, 1999.

[16] M. Kato, H. Ohshima, S. Oyama, and K. Tanaka. Can social tagging improve web image search? In *WISE'08*, pages 235–249. Springer, 2008.

[17] J. Krug and X. Xu. Imagery, context availability, contextual constraint, and abstractness. In *Proceedings of the 23rd Annual Conference of the Cognitive Science Society*, pages 1134–1139. Erlbaum, 2001.

[18] D. Lambov, G. Dias, and J. Graca. Multi-view learning for text subjectivity classification. In *1st Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, 2010.

[19] M. Larson, C. Kofler, and A. Hanjalic. Reading between the tags to predict real-world size-class for visually depicted objects in images. In *MM'11*, pages 273–282. ACM, 2011.

[20] M. Nakatani, A. Jatowt, and K. Tanaka. Easiest-first search: towards comprehension-based web search. In *CIKM'09*, pages 2057–2060. ACM, 2009.

[21] U. D. of Education Office of Educational Research and Improvement. Adult literacy in america, 2002. `http://nces.ed.gov/pubs93/93275.pdf` [27th November, 2012].

[22] A. Paivio, J. C. Yuille, and S. A. Madigan. Concreteness, imagery, and meaningfulness values for 925 nouns. *Journal of Experimental Psychology*, 76(1, Part 2):1–25, 1968.

[23] R. Rumbo. English composition 1: Using specific and concrete diction. `http://www2.ivcc.edu/rambo/eng1001/eng1001_diction.htm` [27th November, 2012].

[24] P. Schwanenflugel. *Why are Abstract Concepts Hard to Understand?*, pages 223–250. 1991.

[25] A. Sun and S. Bhowmick. Quantifying visual-representativeness of social image tags using image tag clarity. In *Social Media Modeling and Computing*, pages 3–23. Springer, 2011.

[26] J. T. and E. Richardson. *Imagery, concreteness, and lexical complexity*, volume 27 of *2*, pages 211–223. Psychology Press, 1975.

[27] M. P. Toglia and W. F. Battig. Handbook of semantic word norms. *John Wiley & Sons, Inc., One Wiley Drive, Somerset, New Jersey 18873*, page 152, 1978.

[28] J. Ure. *Lexical density and register differentiation*, pages 443–452. London: Cambridge Univ. Press, 1971.

[29] V. N. Vapnik. *The nature of statistical learning theory*. Springer-Verlag New York, Inc., 1995.

[30] G. K. Zipf. *Human behavior and the principle of least effort*. Addison-Wesley, Cambridge, (Mass.), 1949.