Using Complexity Measures in Information Retrieval

Frans van der Sluis
Human Media Interaction, University of Twente
P.O. Box 217, 7500 AE
Enschede, The Netherlands
f.vandersluis@utwente.nl

Egon L. van den Broek
Human Media Interaction, University of Twente
P.O. Box 217, 7500 AE
Enschede, The Netherlands
vandenbroek@acm.org

ABSTRACT

Although Information Retrieval (IR) is meant to serve its users, surprisingly little IR research is *not* user-centered. In contrast, this article utilizes the concept complexity of information as the determinant of the user's comprehension, not as a formal golden measure. Four aspects of user's comprehension are applies on a database of simple and normal Wikipedia articles and found to distinguish between them. The results underline the feasibility of the principle of parsimony for IR: where two topical articles are available, the simpler one is preferred.

Categories and Subject Descriptors: H.1.1 [User/Machine Systems]: Value of information; H.1.2 [User/Machine Systems]: Human information processing; H.3.1 [Information Storage and Retrieval]: Content Analysis and Indexing; H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval

General Terms: Human Factors, Experimentation, Algorithms.

Keywords: Complexity Measures, Parsimony, Relevance, Comprehension.

"Frustra fit per plum quod potest fieri per pauciora".

What can be explained by the assumption of fewer things is vainly explained by the assumption of more things.

William of Ockham (1288–1348)

1. INTRODUCTION

Ockham's razor, also known as the principle of parsimony, states: where two theories explain the same data, the simpler theory is preferred. This paper applies this principle, a rule of thumb for science, to IR. Users judge the complexity of documents using the following criteria: understandability [24], comprehensibility [18], and accessibility [19]. Complexity has also been related to user experience: difficulty is a significant cause of (negative) affect [1] and has, accordingly, been posited as a primary antecedent of information retrieval experience [23]. In addition, a good balance between the user's skills and the complexity of the information can make information search intrinsically motivating [2]. Although formal indicators for document complexity have been identified in literature, a unambiguous user-centered notion on what constitutes (perceived) document complexity is still lacking. To achieve such, comprehension will be used as a starting point. Comprehension refers to a process allowing the user to understand and use information and is, for a large part, determined by: readability, amount of information, coherence, and content overlap.

The *readability* of a text has been shown to facilitate comprehension [20]. Readability is the ease of reading a text, and is dependent upon vocabulary, sentence structure and style of a text. Moreover, it relates to the reading skills and interests of the reader . Two seminal psychology works imply that the higher the *amount of information* of texts, the harder it is to process them. First, Miller [16] showed that humans share some similarity with a communication system, among which a channel capacity; i.e., a limited amount of input information that can be transmitted to a certain amount of (correlated) output information. Second, Hick's law [9] noted that more information leads to longer decision times. A higher textual *coherence* leads to better comprehension. However, in specific situations where the reader has an appropriate level of content overlap, less coherent texts may actually stimulate the deep processing of the reader [12]. From a linguistic perspective, coherence can be both on a grammar level and a semantic level. Semantic coherence has been divided in a microstructure and a macrostructure, which denotes respectively the local (i.e., sentence level) and global organization/connection of propositions. *Content overlap* refers to the overlap between the text and the reader's prior knowledge: texts too close to the reader's knowledge are redundant and texts too far away are too difficult [12].

Next, metrics will be introduced for document complexity based on the user-centered notion of comprehension. In Section 3, these metrics are tested on a database that includes texts from both Simple Wikipedia and English Wikipedia. Finally, in Section 4, the results of this endeavor are discussed.

2. METRICS OF COMPLEXITY

The determinants of comprehension do not allow to define *one* metric. However, some of them do allow the definition of a metric of complexity, as is reviewed next.

IIiX 2010, August 18–21, 2010, New Brunswick, New Jersey, USA.
Copyright 2010 ACM 978-1-4503-0247-0/10/08 ...$10.00.

2.1 Readability Formulae

The common approach to estimating the difficulty of a text is by readability measures. These are rough measures relating textual characteristics like content, style, design, and structure to averaged latent user characteristics about prior knowledge and reading skills [13], often through a linear regression model.

A range of measures exists for determining how readable a text is, of which four common ones will be illustrated. These formulae have been shown to be highly correlated to each other, ranging from .90 to .99 [5]. First, the Flesch Reading Ease Scale [4], ranging from 0 to 100 [6]. Let WpS be the words per sentence and SpW the syllables per word:

$$S_f = 206.84 - 84.60\text{SpW} - 1.02\text{WpS}. \tag{1}$$

Second is the Flesch-Kincaid Readability formula [11]. It indicates the reading grade level (G_f) of a text, from grade 5 to college level:

$$G_f = 0.39\text{WpS} + 11.80\text{SpW} - 15.59. \tag{2}$$

Third, the Fog Index [8], representing the number of years of formal education needed (E_f) for a text. Let W be the number of words and PW the number of polysyllabic words:

$$E_F = 0.40\text{WpS} + \frac{\text{PW}}{W} \times 100. \tag{3}$$

Fourth, the SMOG Readability Formula [15], gives the minimal required reading grade (G_s) for a text. Let S denote the number of sentences:

$$G_s = 3 + \sqrt{30\frac{\text{PW}}{S}}. \tag{4}$$

Although these formulae do indeed differentiate between texts of different complexities [4], they do not aid understanding in what features make some documents are more difficult to understand than others (e.g., whether the problem is in syntactics or semantics).

2.2 Entropy

A metric for the amount of information in an ergodic signal is entropy. In particular, entropy measures the uncertainty or the informativeness of an observation. Text can be modeled as an ergodic signal. For example, every letter can be seen as a symbol s and the total of all letters of the text as A. The entropy for a sequence of symbols is:

$$H_n = -\sum_{B \in A^n, s \in A} p(B,s) \log_b \frac{p(B,s)}{p(B)}, \tag{5}$$

where b is the logarithmic scale (usually 2; i.e., bits), A^n is the collection of all sequences of length n, $p(B,s)$ is the probability of sequence B followed by symbol s, and $p(B)$ the probability of sequence B [21]. Please note that to evaluate a text on its content, it is more appropriate to use larger values of n (e.g., 3) and words instead of letters [7].

The information content indicates an important facet of comprehension: the amount of information that has to be processed. However, in relation to comprehension, the construct validity is less optimal: entropy is not a precise measure for the perceived amount of information.

2.3 Semantic Complexity

Using a lexicon (e.g., WordNet), Gervasi and Ambriola [7] provide an approach to indicate how much knowledge is needed to comprehend what is read. The method counts, for all words in a text, the number of concepts that are within n steps related to a word w. The higher this count, the less complex the text is expected to be.

Let W be the lexical database (e.g., WordNet), φ_w be all the synsets of which the word w is part of (i.e., lexical categories such as synonym sets in WordNet), and $r(\varphi, \varphi')$ be a boolean function indicating the relationship between synset φ and synset φ‘. Then, all the synsets (A) related in n steps to a word w are given by:

$$\begin{aligned} A_0(w) &= \{\varphi \in W | w \in \varphi\} \\ A_{n+1}(w) &= A_n(w) \cup \{\varphi \in W | r(\varphi, \varphi') \wedge \varphi' \in A_n(w)\}. \end{aligned} \tag{6}$$

The method can be defined for a whole text T as well:

$$A_n(T) = \bigcup_{w \in T} A_n(w). \tag{7}$$

A normalized versions of this method will be reported. Namely, $\overline{A}(T)$ is A normalized for the size of the text, by dividing A with $A_0(T)$.

As a measure of meaningful information, this measure of semantic complexity differentiates on a facet of complexity not touched on by either readability or entropy. The measure closely relates to the required semantic knowledge for a surface comprehension of a text.

2.4 Semantic Coherence

The semantic coherence indicates how sentences are linked together. It refers to the use of repetition of words or the use of closely related words over sentences. The coherence is measured by the average similarity between each pair of succeeding sentences. Each sentence is represented as a bag of words.

A semantic similarity measure determines the highest similarity sim between the synsets $A_0(w)$ (see Equation 6) related to a word w, averaged over all possible combinations of words in the compared sentences S_1 and S_2 [14]:

$$\text{sim}(S_1, S_2) = \frac{\sum_{w_1 \in S_1, w_2 \in S_2} \text{sim}(w_1, w_2)}{W(S_1) W(S_2)} \tag{8}$$

with

$$\text{sim}(w_1, w_2) = \text{argmax}_{\varphi_1 \in A_0(w_1), \varphi_2 \in A_0(w_2)} \text{sim}(\varphi_1, \varphi_2).$$

Here, W refers to the number of words, and argmax iterates over all synsets φ related to word w, selecting the most similar relation. Numerous implementations for the $\text{sim}(\varphi_\alpha, \varphi_\beta)$ function exist. In this paper, the St-Onge implementation is used, weighing and restricting direction changes next to the path length in a semantic network [10].

An extensive user study showed that the presented metric for semantic coherence correlates ($r = .32$) reasonably with perceived coherence [14]. Hence, it is a valid indication of coherence, although there is more to perceived coherence than the metrics show.

3. EVALUATION

To evaluate the metrics introduced in the previous section, a data set with a clear diversity in complexity was needed. The Wikipedia encyclopedia, available in both normal English (See Table 1a) and simple English (See Table 1b), perfectly suited this aim as the latter is explicitly targeted at readability, semantics, and coherence. The simple English

Table 1: List of resources.

a) English Wikipedia: `en.wikipedia.org`
b) Simple English Wikipedia: `simple.wikipedia.org`
c) WikiXMLJ: `code.google.com/p/wikixmlj`
d) Bliki engine: `code.google.com/p/gwtwiki`
e) Jericho HTML parser: `jericho.htmlparser.net`
f) Fathom: `www.representqueens.com/fathom`

set is expected to represent how the authors view complexity, making it a valuable, user-centered source. However, the simple English articles tend to be smaller than their English Wikipedia counterparts and, consequently, describe their topics in less depth.

3.1 Method

The data as used was retrieved on April 1, 2010 and consisted of two dumps, containing all articles encoded as wiki-text for both normal and simple English. Both sets were imported into a MySQL database, using WikiXMLJ (See Table 1c). Only articles were imported that were not a stub (i.e., an incomplete, article), special, disambiguation, or redirect page. Moreover, solely articles were selected that were found in both simple and normal English, allowing a pair-wise comparison.

The data was converted from wiki-text to normal text via HTML, in order to preserve the layout of the text. Wiki-text to HTML conversion was done using the Bliki engine (See Table 1d). The non-standard templates were not parsed; hence, items like menus and references were omitted. Finally, the HTML was converted to text using a customized Jericho HTML parser (See Table 1e). The resulting extracted text was parsed into sentences, words, and characters with the default Java text processing toolkit. The resulting data set consisted of $46,292$ articles, or $23,146$ pairs of both simple and normal English articles.

The readability features were analyzed on paragraphs consisting of at least one sentence and at least two words, to prevent any noise from headers. The number of syllables in a word was analyzed using the Fathom toolkit (See Table 1f). Entropy sequences of up to $n = 3$ were explored on a case-insensitive encoding of the text. The semantic features were based on WordNet version 3.0 [cf. 17]. For both semantic complexity and semantic coherence, only nouns were considered and only relations up to 5 steps were explored by the algorithms. Moreover, for semantic complexity, hyponymy relations were excluded to prevent the method from quickly converging on the whole lexicon. And, for the St-Onge [10] implementation of semantic coherence, the values of $C = 6.50$ and $k = .50$ were used.

The features were compared, besides using normal descriptive statistics, on their ability to distinguish between simple and normal English articles. Due to the large sample sizes, a normal statistical test cannot differentiate between the features. Therefore, we used two coefficients of statistical power: the point biserial correlation coefficient r_{pb} [22] with a pooled sample standard deviation and the Mann-Whitney U^N, a non-parametric coefficient of statistical power. The latter was used because the coherence metric does not follow a normal distribution, but is far skewed towards zero (i.e., no relation between sentences). Both functions have a range between 0 and 1, where 0 implies no difference and 1 a maximal difference.

Table 2: Features' descriptives and statistical power.

	Simple		English		Distance	
$\mathcal{F}$	$\mathcal{M}$	$\mathcal{SD}$	$\mathcal{M}$	$\mathcal{SD}$	r_{pb}	U^N
Length						
W	258.03	1971.52	2377.60	12439.22	.118	.449
S	17.57	123.70	138.76	791.89	.106	.434
Readability						
S_f	47.55	23.61	37.84	16.15	.233	.168
G_f	10.38	5.45	12.51	3.77	.222	.286
E_F	23.67	10.23	28.20	6.78	.252	.246
G_s	11.42	3.34	13.48	2.40	.333	.302
Entropy						
H_0^C	4.08	0.12	4.16	0.05	.376	.267
H_1^C	2.80	0.55	3.43	0.20	.604	.442
H_2^C	1.23	0.60	2.33	0.45	.720	.453
H_0^W	5.74	1.20	7.75	0.87	.690	.449
H_1^W	0.88	0.53	1.93	0.64	.667	.425
H_2^W	0.14	0.14	0.34	0.18	.532	.365
Semantic Complexity						
$\overline{A}_2$	22.32	12.67	9.49	5.51	.549	.354
$\overline{A}_3$	39.94	35.85	9.74	9.70	.498	.369
$\overline{A}_4$	64.34	68.13	11.46	14.99	.472	.373
$\overline{A}_5$	64.15	79.80	9.11	15.87	.431	.374
Semantic Coherence						
C	0.18	0.15	0.14	0.10		.066

Note. Features are specified in Section 2.

Abbrev. $\mathcal{F}$: features; $\mathcal{M}$: mean; $\mathcal{SD}$: standard deviation; H_n^C: character entropy, H_n^W: word entropy

3.2 Results

Table 2 shows the descriptive statistics of each of the features for both the simple and normal English Wikipedia articles. These features have been categorized by the metrics described in Section 2, and an extra category containing two features indicative of article length. As can be seen, all features behave according to expectation. For example, the Flesch Reading Ease Scale (S_f) decreases, the entropy shows an increase, and coherence a decrease. Also, semantic complexity decreases: a higher value of $\overline{A}$ indicates less prerequisite knowledge.

The differences between both Wikipedias are also shown in Table 2. The average article length of English Wikipedia is longer, both in words and in sentences. In order to control for the effects of article length, the average Pearson's correlation r for the length features compared to each of the metrics were determined. The correlations indicate that the different metrics correlate slightly with the article length: $r = .004$ for coherence, $r = .075$ for readability, $r = .093$ for semantic coherence, and $r = .181$ for entropy.

Table 2 gives two coefficients for the statistical power: the point biserial correlation coefficient, r_{pb}, and the Mann-Whitney U^N. Using Cohen's [3] rule-of-thumb, the effect sizes can be interpreted as small if $.100 < r_{pb} \leq .243$, medium if $.243 < r_{pb} \leq .371$, and large if $r_{pb} > .371$. Subsequently, the statistical power analysis indicates that entropy and normalized semantic complexity has strong, readability medium, and coherence has little differentiation power between both Wikipedias.

Since the goal is to benchmark the features, it is informative to look at the differences between the metrics as well. For this, the averaged cross-correlations between the metrics were determined. These show a strong indication that readability, entropy, semantic complexity, and semantic coherence do indeed measure different aspects of complexity, with $r_{pb} \leq .243$. An exception to this are entropy and semantic complexity that correlate strongly with $r_{pb} = .670$.

4. DISCUSSION

This paper reported on the feasibility of applying Ockham's razor to search results: preferring a simple text above a complex one. Founded on human information processing, four facets of comprehension were introduced (Section 1): readability, amount of information, coherence, and content overlap. Next, for each facet, Section 2 introduced accompanying metrics of complexity. Subsequently, these were tested on data distinctive in complexity: simple and normal English Wikipedia. The evaluation showed that most metrics could indeed differentiate between different levels of complexity. Moreover, the tests showed that the metrics measure different properties of complexity. This indicates that the four determinants of comprehension are indeed reflected by the metrics.

The Wikipedia data set implies two limits on their interpretation. First, the data was likely not very distinctive on coherence. Both simple and normal English Wikipedia present coherent articles, as is confirmed by the relatively low statistical power for the coherence metric. Second, the data was not only distinctive in the complexity of articles, but also in article length. To control for this effect, the correlations between length and the complexity metrics was computed. These correlations were low, except for entropy. These effects can be explained by the difference in length between the two sets of articles, as longer articles discuss more information (entropy).

The aim of using complexity measures in IR is to retrieve information better suited to the user. However, the intrinsic relation with the user's percept of complexity is far from trivial. We pose that with this research, a first step is made towards an adaptive variant of IR: giving the user more control over the retrieved information and, consequently, increase the user experience in its broadest sense. This would pave the path towards the development of IR systems as they should be: truly user-centered.

5. ACKNOWLEDGEMENTS

We thank Claudia Hauff, Betsy van Dijk, Anton Nijholt, and Franciska de Jong for their helpful comments on this research. This work was part of the PuppyIR project, which is supported by a grant of the 7th Framework ICT Programme (FP7-ICT-2007-3) of the European Union.

References

[1] I. Arapakis, J. M. Jose, and P. D. Gray. Affective feedback: an investigation into the role of emotions in the information seeking process. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 395–402, New York, USA, 2008. ACM.

[2] J. Chen. Flow in games (and everything else). *Commun. ACM*, 50(4):31–34, 2007.

[3] J. Cohen. *Statistical power analysis for the behavioral sciences*. Lawrence Erlbaum, Hillsdale, NJ, 1988.

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

[5] D. B. Friedman and L. Hoffman-Goetz. A Systematic Review of Readability and Comprehension Instruments Used for Print and Web-Based Cancer Information. *Health Educ Behav*, 33(3):352–373, 2006.

[6] E. Fry. Readability versus leveling. *Reading Teacher*, 56(3):286, 2002.

[7] V. Gervasi and V. Ambriola. Quantitative assessment of textual complexity. In L. Merlini Barbaresi, editor, *Complexity in Language and Text*, pages 197–228. Pisa, Italy: Plus Pisa University Press, 2003.

[8] R. Gunning. *The technique of clear writing*. McGraw-Hill, 1968.

[9] W. E. Hick. On the rate of gain of information. *Quarterly Journal of Experimental Psychology*, 4(1):11 – 26, 1952.

[10] G. Hirst and D. St-Onge. Lexical chains as representations of context for the detection and correction of malapropisms. In C. Fellbaum, editor, *WordNet: An Electronic Lexical Database*, pages 305–332. MIT Press, 1998.

[11] J. P. Kincaid and et al. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, National Technical Information Service, Springfield, Virginia, 1975.

[12] W. Kintsch. Text comprehension, memory, and learning. *American Psychologist*, 49(4):294 – 303, 1994.

[13] G. R. Klare. The measurement of readability: useful information for communicators. *ACM J. Comput. Doc.*, 24(3):107–121, 2000.

[14] M. Lapata and R. Barzilay. Automatic evaluation of text coherence: models and representations. In *IJCAI'05: Proceedings of the 19th international joint conference on Artificial intelligence*, pages 1085–1090, San Francisco, CA, USA, 2005. Morgan Kaufmann Publishers Inc.

[15] G. H. McLaughlin. Smog grading - a new readability formula. *Journal of Reading*, 12(8):639–646, 1969.

[16] G. A. Miller. The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review*, 63(2):81 – 97, 1956.

[17] G. A. Miller. Wordnet: a lexical database for english. *Commun. ACM*, 38(11):39–41, 1995.

[18] S. Mizzaro. How many relevances in information retrieval? *Interacting with Computers*, 10(3):303 – 320, 1998.

[19] L. Schamber. Relevance and information behavior. *Annual Review of Information Science and Technology (ARIST)*, 29:3–48, 1994.

[20] T. Shanahan, M. L. Kamil, and A. W. Tobin. Cloze as a measure of intersentential comprehension. *Reading Research Quarterly*, 17(2):229–255, 1982.

[21] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 625–656, Jul, Oct 1948.

[22] R. F. Tate. Correlation between a discrete and a continuous variable. point-biserial correlation. *The Annals of Mathematical Statistics*, 25(3):603–607, 1954.

[23] F. Van der Sluis, E. L. Van den Broek, and E. M. A. G. Van Dijk. Information Retrieval eXperience (IRX): Towards a human-centered personalized model of relevance. In *Third International Workshop on Web Information Retrieval Support Systems August 31, 2010 (Toronto, Canada)*, in press.

[24] Y. C. Xu and Z. Chen. Relevance judgment: What do information users consider beyond topicality? *J. Am. Soc. Inf. Sci. Technol.*, 57(7):961–973, 2006.

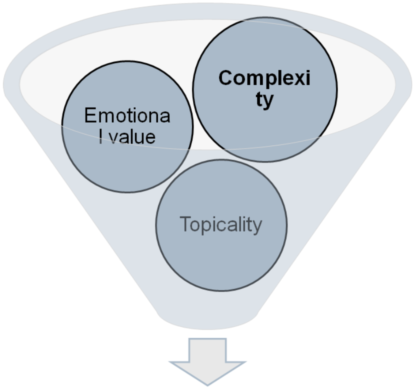
Using Complexity Measures in Information Retrieval

Frans van der Sluis and Egon L. van den Broek
f.vandersluis@utwente.nl and vandenbroek@acm.org
Human-Media Interaction (HMI), University of Twente

Examples of Complexity

“*Occam's razor (or Ockham's razor) is a principle from philosophy. It says that the simplest explanation is usually the best one. William of Ockham, a Franciscan friar who studied logic in the 14th century, first made it well known. The principle says that if there are several possible ways that something might have happened, the way that has the least guesses involved is probably the right one.*”	**Simple**
“*Occam's razor (or Ockham's razor), entia non sunt multiplicanda praeter necessitatem, is the principle that ``entities must not be multiplied beyond necessity'' and the conclusion thereof, that the simplest explanation or strategy tends to be the best one. The principle is attributed to 14th-century English logician, theologian and Franciscan friar, William of Ockham.*”	**Normal**

Information Retrieval eXperience (IRX)



Comprehension Experience

Comprehension and Complexity

Determinants of Comprehension	Metric of Complexity
Readability	Readability formulae
Amount of Information	Entropy
Content overlap / prerequisite knowledge	Semantic Complexity
Coherence	Semantic Coherence

Affiliations

Human Media Interaction

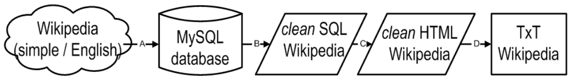
UNIVERSITY OF TWENTE.

Automatic Detection of Complexity: Method

Simple Wikipedia versus Normal Wikipedia:
- 46,292 articles.
- 23,146 pairs of articles.
- Each pair of article compared on complexity.

Simple Wikipedia authors aim for readability, semantics, coherence. Hence, a user-centered data set.

Data preprocessing in four steps:
A. Select and import decent quality articles from Wikipedia dump
B. Construct article pairs
C. Convert wiki-text to html
D. Convert html to structured text, maintaining layout of text (e.g., paragraphs)



Automatic Detection of Complexity: Results

Results:
- Most features differentiate very well between levels of complexity.
- Coherence seems to be equal in both Simple and Normal Wikipedia.
- Influence of article length marginal (r = .004 to r = .181).

Metric	Distance (r)	Power (Cohen’s interpretation)
Readability	.233 - .333	Small – Medium
Entropy	.376 - .720	Large
Semantic Complexity	.431 - .549	Large
Semantic Coherence	~	None

Conclusion

- Complexity is an intrinsic part of relevance and IRX.
- The presented model of complexity is founded on human information processing through the four facets of comprehension.
- Features were determined and validated on an extensive user-centered data set.
- The model aims towards IRX and, thus, information better suited to the user.
- The results show the feasibility of applying the principle of parsimony for IR.

Related Work

- F. Van der Sluis, E.L. Van den Broek, and E.M.A.G. Van Dijk. Information Retrieval eXperience (IRX): Towards a human-centered personalized model of relevance. In Third International Workshop on Web Information Retrieval Support Systems. August 31, 2010 (Toronto, Canada), in press.
- F. Van der Sluis and E. L. Van den Broek. Modeling user knowledge from queries: Introducing a metric for knowledge. In Proceedings of the 2010 International Conference on Active Media Technology, Lecture Notes in Computer Science. Springer, in press.