

Generality of Texts

R. B. Allen and Yejun Wu

College of Information Studies
University of Maryland
College Park, MD 20742, U.S.A.
rba@glue.umd.edu and wuyj@glue.umd.edu

Abstract. When searching or browsing a user may be looking either very general information or very specific information. We explored predictors for characterizing the generality of six encyclopedia texts. We had human subjects rank-order the generality of the texts. We also developed statistics from analysis of word frequency and from comparison to a set of reference terms. We found a statistically significant relationship between the human ratings of text generality and our automatic measure.

1 INTRODUCTION

In collection selection it would be helpful to have some measures of the characteristics of a collection. As part of our ongoing research to define the properties of documents and collections, we have explored a measure of the “scope” of collection, where scope is considered to be the “breadth” of a collection [1].

We believe that it would also be of interest to determine the scope of individual document. However, we were not able to get consistent inter-subject ratings of the “scope” of documents in pilot studies. One difficulty concerned the human subjects’ understanding of what constituted the “scope”. The subjects seemed particularly confused about how to compensate for document structure.

In this work, we explore a related property of document “generality”. By the “generality” we mean that a document “addresses general things or concepts”. We believe that documents generality is easier for individual to understand and the document we examined did not have distinct sections. We believe the searchers will find it useful to know whether a retrieved document is general or concrete. For instance, a general chemistry textbook would cover many areas of chemistry and would be distinguished from texts that focused on specialities in chemistry from such as organic chemistry and biochemistry.

2 GENERALITY OF TEXTS

2.1 Rating the Generality of Documents

We selected six one-page articles from www.encyclopedia.com which we judged covered a range from “general” to “concrete”. Eight subjects were asked to rate

these documents using the instructions in Figure 1. To encourage consistency, the subjects first made pair-wise comparisons between the documents and they used these judgements to rank-order the documents' generality. The subjects were paid \$12 for the approximately one-hour task.

You are asked to judge (have a sense of) the generality/concreteness of each document. If a document addresses mostly general things/concepts, we consider it has high generality; if a document addresses mostly concrete or specific things, it has low generality. Read through each article in 4-5 minutes and have a sense of its generality. Please feel free to take notes on the documents. Please plan to read through all the 6 documents in about 30 minutes. Take 2 documents at a time and compare their generality, assign either "more than" ($>$), or "equal to" ($=$), or "less than" ($<$) between the two documents on the attached work sheet. There are totally 15 comparisons. Please plan to finish all the comparisons in about 20-25 minutes. Finally, rank/sort the 6 documents in terms of generality (please indicate ties if any).

Fig. 1. Instructions to subjects.

Beyond the instructions, the subjects developed their own criteria. During the debriefing at the end of the experiment, they reported basing their ratings of concreteness on factors such as: "more numbers, places, enumeration of the aspects of an entity and detailed things". Figure 2 shows a selection from a general document while Figure 3 shows a selection from a concrete document.

In many countries extensive government programs control the planning, financing, and regulation of agriculture. Agriculture is still the occupation of almost 50% of the world's population, but the numbers vary from less than 3% in industrialized countries to over 60% in Third World countries.

Fig. 2. Selection for the text on "modern agriculture" from encyclopedia.com which was rated as most "general".

The average of the inter-subject correlation was 0.12. However, one subject was very different from the others and without that subject the inter-observer correlation would have been much higher.

3 GENERAL/CONCRETE TERMS

Following our work on the scope of collections [1], in which we were able to predict the distance between documents (and hence the scope of the collection)

In the 1960s and 1970s, a combination of 2,4-D and 2,4,5-T was widely used in Vietnam as a defoliant under the name Agent Orange . As a result of questions concerning the possible health effects of the use of Agent Orange, heightened awareness of possible ecological and health dangers attributable to herbicides has resulted in reevaluation of many compounds and has called indiscriminate use into question.

Fig. 3. Selection for the text on "herbicide" which was rated as most "concrete".

based on the distance between individual words, we attempted to predict the generality of documents from the generality of the terms in those documents. Therefore, we developed a measure for the generality of documents based on the terms in the documents.

4 Using Cooccurrence to Tell Whether a Target Word is General or Concrete

We believe that general terms should appear across a wide variety of contexts. A set of 64 terms was chosen to serve as a reference collection. An attempt was made to choose a wide ranges of terms from general to specific. Our technique allows all parts of speech to be measured as concrete or specific. For instance, we were able to judge adjectives as specific or general.

4.1 Measuring Co-Occurrence

Search engines such as Google and Altavista provide counts of "pages found" for their searches. As developed in Allen and Wu [1], these measures can be used to provide a measure of the relatedness of two terms.

Figure 4 shows the (asymmetric) relatedness of several pairs of words obtained with this method. See [1] for more discussion on the interpretation of these values.

Word1	Word2	Relatedness	
		W1-W2	W2-W1
truck	automobile	0.0740	0.0776
gun	shoot	0.1097	0.1917
flower	beautiful	0.1242	0.0481
mouse	keyboard	0.2542	0.2960

Fig. 4. Sample co-occurrence relatedness values for word pairs (from [1]).

We define A as the number of hits for word1, B for word2, and C for word1 AND word2. The relatedness between word1 and word2 can be easily calculated

as $\frac{C}{A}$ or $\frac{C}{B}$ or other variant expressions. A joint entropy measure (Equation 1) was most successful.

$$Rel\ Joint\ Entropy = -\left(\frac{C}{A} \text{Log} \frac{C}{A}\right) - \left(\frac{C}{B} \text{Log} \frac{C}{B}\right) \quad (1)$$

We wrote a C-language program that used the UNIX utility “expect” for managing the interactive telnet with the Google research server, extracted the frequency counts from the return file, and stored the results using the database utility “gdb”.

4.2 Predicting the Generality of Words

Figure 5 shows examples of words from our reference list of 64 terms that demonstrated high-generality and low-generality.

High-Generality Terms	Low-Generality (Concrete) Terms
allow	cognition
take	hawk
approach	hat
nature	atom
change	fluoride
process	minnow

Fig. 5. Examples of high-generality and low-generality terms.

The joint entropy measure was used to confirm that general terms were more related to each other than were concrete terms related to each other. We computed the relatedness of the 32 general terms with themselves, the 32 concrete terms with themselves, and the general terms with the concrete terms (see Figure 6). The difference between 0.168 ($\sigma = 0.017$) and 0.137 ($\sigma = 0.025$) is significant, $t = 10.48, p < 0.01, df = 31$ (two-tailed test). The difference between 0.137 and 0.110 ($\sigma = 0.028$) is also significant, $t = 4.98, p < 0.01, df = 31$.

	General	Concrete
General	0.168 (hi)	0.137 (mid)
Concrete	0.137 (mid)	0.110 (lo)

Fig. 6. Average relatedness of the two sub-lists. Predictions are shown in parentheses.

4.3 Validation With WordNet

As another validation of our approach, we hypothesized that general words would be closer to the root of the WordNet hierarchies [3]. We counted the levels between a word and its root in the WordNet hierarchy. If a word has more than one meanings in WordNet, we take the most common one. The average level of the top 32 general words versus the 32 concrete words is 3.30 ($\sigma = 1.31$) and 6.96 ($\sigma = 2.51$) respectively ($t = 8.87$, *two-tailed*, $p < 0.01$, $df = 31$).

4.4 Relationship Between Term Generality and Document Generality

We took all the terms for each document and filtered them through a stop list of 300 common words. We then took words that were repeated at least twice in the remaining word lists. The generality of each word was computed as its relatedness to the set of 64 reference words using the joint entropy measure. We took the mean of word generality to obtain the generality of the entire document. Finally, we ranked the six documents according to their generality and compared them with the averaged human ratings of the documents' generality. A linear trend test was performed with ANOVA [4]. A specific comparison for the linear component was a significant effect, $F(1, 42) = 6.24$, $p < 0.05$. This confirmed that the generality of the terms was an effective prediction of the generality of the texts.

4.5 Word Frequency

Word frequency has been proposed as a measure of document readability, although more common measures of readability (e.g., [6]) are based on the number of syllables and the length of words. General words tend to be of higher frequency. The correlation of generality and word frequency was $r = 0.62$, $df = 5$, *n.s.*

We believe that our generality measure has more face validity than word frequency for predicting the generality of concepts in the articles. For instance, note that there are several low-frequency terms with high generality such as “approach” and “handle”. There are also high-frequency terms with low-generality. Terms such as “hat” and “game”. We found that the generality measure predicted mean ratings better than the word frequencies using a stepwise multiple linear regression, although the difference of the measures was not significant.

4.6 Hierarchy and Categories

Because of the relationship between these measures and the WordNet hierarchy reported above, we explored whether these measures fully predicted hierarchical relationships. We found that in the hierarchies such as ANIMAL >> MAMMAL >> DOG that mammal was less frequent than required to support the hierarchical order. Rather, the results seem to support the view that DOG is a basic-level category [5], while MAMMAL is not.

5 CONCLUSION

5.1 Applications and Implications

In the traditional approach to document searching, a search return list simply presents the documents ranked according to how well they match a retrieval algorithm. We go further and propose that the notion of relevance ranking should be replaced by indicators of document generality. Indeed, we would go even further and suggest that information retrieval systems should increasingly attempt to educate their users to understand the concepts used in the documents in the return list. This may mean that general documents returned from a search should be presented before concrete documents. More ambitiously, the search return lists could adopt the principles of adaptive hypermedia [2].

5.2 Other Properties of Documents

We have now explored the notions of the “scope” of a collection and the “generality” of a document. There are several other related concepts that we would like to operationalize. These include the “depth” and “coverage” of a document. However, our initial attempts to do this have encountered difficulties. For instance, we would like to measure the “number” of concepts in a document. However, thus far we have had difficulty in defining an unambiguous measure for identifying concepts. Nonetheless, we believe that eventually a full range of metrics will be worked out and that these will help both human beings and machines interact with those documents.

6 Acknowledgement

We thank Google for access to the research port on their service during the Spring of 2002. We also thank encyclopedia.com for the use of their articles.

References

1. ALLEN, R. B., AND WU, Y. Measuring the scope of collections, in preparation.
2. BRUSILOVSKY, P. Adaptive hypermedia. *User Modeling and User Adapted Interaction 11* (2001), 87–110.
3. FELLBAUM, C., Ed. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge MA, 1998.
4. KEPPEL, G. *Design and analysis: A researcher's handbook*, 3rd ed. Prentice Hall, Englewood Cliffs NJ, 1991.
5. ROSCH, E. Principles of categorization. In *Cognition and Categorization*, E. Rosch and B. B. Lloyd, Eds. Erlbaums, Hillsdale NJ, 1978, pp. 27–78.
6. SHERMAN, L. A. *Analytics of Literature: A Manual for the Objective Study of English Prose and Poetry*. Ginn and Co., Boston, 1983.