A NEW CRITERION FOR MEASURING GENERALITY OF DOCUMENTS

Hyun Woong Shin
Computer Science Department
Integrated Media Systems Center
University of Southern California

Eduard Hovy
Computer Science Department
Institute of the Information Science
University of Southern California

Dennis McLeod
Computer Science Department
Integrated Media Systems Center
University of Southern California

Larry Pryor
USC/Annenberg School for Communication
University of Southern California

ABSTRACT: *Most information retrieval systems, including Web search engines, use similarity ranking algorithms based on a vector space model to find relevant information in response to a user's request. However, the retrieved information is frequently irrelevant, because most of the current information systems employ index terms or other techniques that are variants of term frequency. In this paper, we propose a new criterion, "generality," that provides an additional basis on which to rank retrieved documents. We compared our generality quantification algorithm with human judges' weighting of values to show that the developed algorithm is significantly correlated.*

Keywords: *Information retrieval, domain dependent ontology, and generality algorithm.*

1. INTRODUCTION

The goal of Information Retrieval (IR) is to provide a facility for a user to easily and efficiently access the information relevant to a user's interest [1]. Most traditional IR systems operationalize "relevant" as the word frequency in a document of a set of keywords (or index terms) [9]. An index term is a word whose semantic reference serves as a mnemonic device for recalling the main themes of a document. The retrieval systems based on index terms are relatively uniform in conception but raise key problems regarding the semantics of the documents and a user's request. The traditional IR systems have simplified these problems to the extent that retrieved documents are frequently irrelevant to a user's request. As the TREC results show year after year, even the best IR systems' precision scores never average higher than 0.6.

The crux of retrieving more-relevant information is better characterizing a user's request. Unfortunately, this is not a simple problem. Most traditional IR systems characterize a user's request as a term frequency. They can therefore only retrieve information that contains the terms that are in the user's request, or terms easily derivable from it. There have been several elaborations of this approach, including clustering [8], topic detection [2], and ontologies [3]. These solutions focus on the semantics of user requests and/or contents. However, there is another aspect to characterizing a user's request: the appropriate level of generality of the retrieved documents.

We use the degree of generality to rerank retrieved documents so that the results displayed to the user are based on not only the index term frequencies, but also the desired generality appropriate for a user's knowledge and interests. Therefore, different users will receive different results, even with the same input query, based on the level of generality appropriate for them. In order to achieve this goal, we create the additional criterion "generality". We hypothesize that retrieval engines including reranking with generality will provide more satisfactory results than those that do not. Before we can test this hypothesis, we have to 1) define generality, 2) quantify the degree of generality as reflected by the position of index terms in the concept hierarchy representing the domain ontology, and 3) confirm that the degree of generality matches

with audience members' intuitive feeling for generality, as determined by human judges. These steps are the goal of this paper.

What is "generality"? In journalism, generality is reflected in both how a story is linearly organized and the status of the audience it is expected to reach. If a story contains several topics, it is called nonlinear and is considered to be a general story; if it focuses on one single or particular topic, journalists consider it to be specialized. Secondly, journalists evaluate the generality of a story on the estimated breadth of the audience that will find the story to be relevant. A general story is thought to be of concern to a relatively large audience that shares a common status and universal interests, and its information potentially applicable or of interest to every member of that audience, such as stories of the Red Sox Curse and its ties to Babe Ruth and how that was overcome in 2004. On the other hand, a specific story has a focus and level of detail that would be of interest to a relatively small niche audience whose members share a common expertise and enthusiasm for the particular topic, such as a story on off-season position player trades, which would interest only Red Sox and baseball fans.

In order to define and compute generality, we require a domain dependent ontology. This ontology, which consists of concept nodes and interrelationships [4], models the user's knowledge and represents the connections between the user's goals. The desired generality, which captures the focus and direction of the user's attention, we then represent as a real number between 1 (specific) and 10 (general). The generality appropriate for the user is determined for documents based on nodes in the domain dependent ontology. More exactly, we define *generality* as measuring the specificity of words (subset of index terms) in a document. We define *specific words* as those that do not belong to (resort under) multiple ontology nodes. We assume that a topic can have a certain amount of specific words. Therefore, if a document contains many specific and unrelated nouns, the document probably contains several topics and is general in nature.

The proposed concept of generality has been rarely researched to date. The most similar study related to generality is conducted by Resnik [6] in Natural Language Processing (NLP). The goal of his work is to measure semantic similarity. The information shared by two concepts is indicated by the information content of the concepts that subsume them in an ontology, using the formal definition $sim(c_1, c_2) = \max_{c \in S(c1,c2)} [-\log p(c)]$. The core difference between his work and ours is that his work measures the specificity of concepts while we measure the specificity of documents. In other words, his work can measure the semantic relatedness of concepts, but it is difficult to measure the semantic relatedness of documents. By contrast, our study measures topical differences among documents containing particular concepts.

The remainder of this paper is organized as follows. In Section 2 we present the generality algorithm. The experimental results are discussed in Section 3. Section 4 concludes this study.

2. THE GENERALITY ALGORITHM

In traditional information retrieval systems, index terms are used to index and retrieve documents. An index term is a keyword (or a group of related words) whose semantic reference serves as a mnemonic device for recalling the main themes of documents. Thus, an *index term set* is simply the set of keywords that appear in the text of documents in some collection. We represent each collection by a node in the ontology, attaching to the node its corresponding index term set.

We define the *specific word set* is a subset of the index term set that does not appear in other word sets. Then the degree of generality can be quantified by the number of index terms in the document that belong to specific word sets. For example, assume a collection of documents C_i has the index term set T_i. The specific term set S_i is a set of index terms that do not belong to index term sets in other collections. That is, $S_i = T_i - \bigcup_{j \neq i} T_j$.

Here, generality is quantified by the appearance of specific index terms t_i within document D_j.

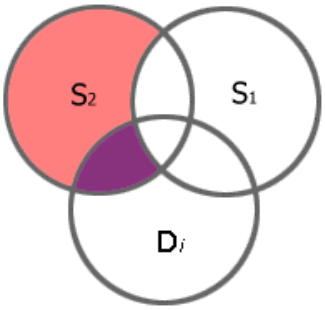

Figure 1. The degree of generality of the document D_j

Figure 1 illustrates the degree of generality of document D_j. In the figure, S_1 and S_2 are specific word sets of two different ontology nodes, and document D_j contains some of their associated terms. The degree of generality of D_j is presented by the dark gray portion, namely $((S_2 \cap D_j) - S_1)$.

Once the degree of generality is determined for each document, we adjust the degree of generality based on the concept hierarchy. The concept hierarchy is a hierarchical structure of related concepts. For example, "Sports" may have a child node, the Olympic-Sports (which we abbreviate to "Olympics"). In addition, the node "Olympics" may have the children nodes "Boxing" and "Taekwondo." In this case, there is a conceptual hierarchy starting from "Sports" to "Olympics" to "Boxing" and "Taekwondo". This adjustment is necessary for two reasons.

First, a child node represents more specific information than a parent node does. In other words, the degree of generality for all documents in a parent node must be assigned a higher value (be more general) than the documents in the child node. This assumption we base on the intuition that domain-specific Ontologies will generally lie mostly below the level of Basic Concepts [7], and hence tend to become of more general interest as one moves upward through them.

Second, documents at the top level of a domain dependent ontology are in practice the most general stories, using the journalism definition. These documents should contain the largest number and variety of specific words. However, it is very unlikely that a document will contain all specific words. To overcome the problem, an adjusted value needs to be added to the degree of generality for each document in child nodes.

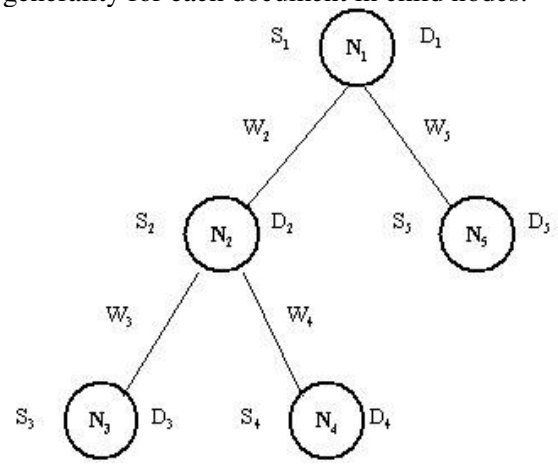

Figure 2. Sample ontology with weights w_i, index term lists D_i, and specific term lists S_i

Figure 2 depicts a sample ontology graph with weights (for adjusting generality), index term lists, and specific term lists. The basic idea behind the degree of generality reflects the differences of specific word sets within the concept hierarchy.

3. RESULTS AND DISCUSSION

Our verification of the measure of generality is performed between a generality algorithm and human judges. Given documents, human judges were asked to mark the degree of generality for each document. The judges used a ten-point scale and assigned a score for each document based on their observation of the degree of generality. The judges were instructed that there are no right or wrong answers.

We used the Pearson correlation coefficients [5] to evaluate the relationship between the scores from two human judges as well as from the human judges and from the algorithm. Also, we used a t-test (t distribution and n-2 degrees of freedom) to determine whether these relationships are statistically significant. If a p-value from the t-test is less than 0.05, we conclude there is a statistically significant correlation between the judges and the system.

The Pearson correlation coefficient always lies between-1 and $+1 -1 \le r \le$, and the values $r = 1$ and $r = -1$ mean that there is an exact linear relationship between the two values. Over 70% is generally considered a good correlation. We first test the generality between two judges' value to show that there is a common generality between human judges. This evaluation assures us that there is a phenomenon to be modeled and computationalized.

Table 1. Pearson correlation coefficient between two human judges[1]

	Level 0 (n=62)	Level 1 (n=62)	Level 2 (n=29)
Pearson correlation coefficient (r)	0.84	0.81	0.81
p-value from the t-test	< .0001 (df=60)	< .0001 (df=60)	<.0001 (df=27)

Table 1 shows the Pearson coefficients and the corresponding p-values from the t-test between the two human judges. This result shows that their evaluations are statistically significantly (more than 80%, p<.001), in spite of individual variability. Level 0 is a parent node of Level 1, Level 1 is a parent node of Level 2, and so on. Although the human judges and the algorithm assign scores for each document in an ontology node, the correlation should be tested among siblings (i.e. same level nodes) because a correlation of each mode cannot provide the correlation in general.

Table 2. Pearson correlation coefficient[1]

	Level 0 (n=62)	Level 1 (n=62)	Level 2 (n=29)
Pearson correlation coefficient (r)	-0.22	0.73	0.68
p-value from the t-test	0.075 (df=60)	<.0001 (df=60)	<.0001 (df=27)

[1] n= number of articles used for the test, df= degrees of freedom

Table 2 shows the Pearson coefficients and the corresponding p-values from the t-test between human judges and the algorithm. Figure 5 shows the corresponding scatter plots for Level 1. The X-axis represents scores from the algorithm and the Y-axis represents human judges' scores. In Figure 5, letters represent the number of observations for scores of human judges and the algorithm ('A': 1 observation, 'B': 2 observations, and so on). For example, if a judge's score is 6, the algorithm's score is 0.6, and it is observed once, the mark 'A' is positioned at the intersection of 6 and 0.6. The linear relationship between human judges' values and the algorithm's values is shown along the line in Figure 3.

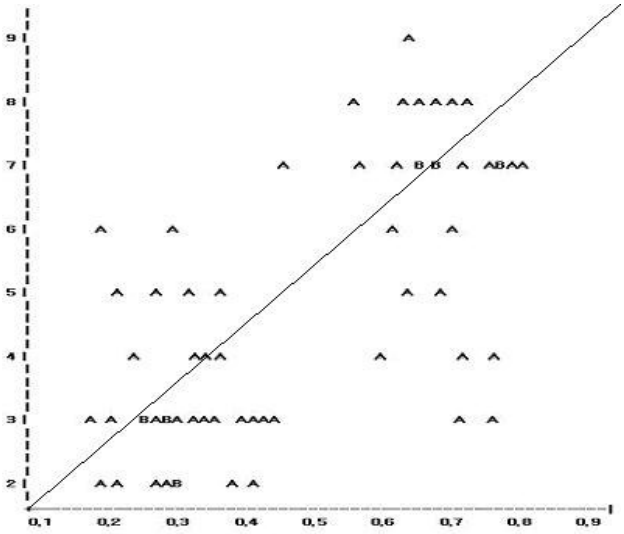

Figure 3. Scatter plot for the degree of generality between human judge and algorithm in Level 1

The results show that there are 73% and 68% correlations between the two at Level 1 and Level 2, respectively, and these relationships are statistically significant (p<0.0001). The scores from the human judges are competitive with those from our algorithm. At the top level, however, the correlation between human judges and the algorithm is very low because no matter what the algorithm calculates as the degree of generality, the judges determine it as 10.

4. CONCLUDING REMARKS

In this paper, we first defined the notion of generality, which is used to indicate how general or specific a document is. A basic idea for

quantification of generality and the algorithm has been devised and developed. We employed Pearson's correlation coefficient to evaluate the relationship between the degrees of generality of the human judge and the algorithm. The experimental results show these relationships are statistically significant ($p < .0001$). As seen in Table 2, the Pearson correlation coefficients are 73% and 68% for Level 1 and Level 2, respectively.

The major contribution of this paper is to propose, devise, and develop a new criterion, generality, for information retrieval society to provide a new facility for capturing user intent and retrieving more "relevant" information in response to the user's request. We investigated the mathematical model of a degree of generality so as to establish a theoretical background.

Our next work will focus on implementing this model in an IR system and testing the results in realistic IR tasks.

ACKNOWLEDGMENTS

This research was supported in part by the USC Integrated Media Systems Center, a National Science Foundation Engineering Center, cooperative agreement No. EEC-9529152.

References

[1] Baeza-Yates, R., and Ribeiro-Neto, B. *Modern Information Retrieval*. Addison Wesley Longman, 1999.

[2] Brants, T., Chen, F., and Farahat, A. *A system for new event detection.* In Proceedings of the 26th International ACM SIGIR International Conference on Research and Development in Information Retrieval, 2003.

[3] Gruber T.R. *Toward Principles for the design of Ontologies used for Knowledge Sharing.* In Proceedings of the International Workshop on Formal Ontology, 1993.

[4] Khan, L., McLeod, D., and Hovy, E.H. *Retrieval effectiveness of an ontology-based model for information selection.* The VLDB Journal, 13(1), 71-85, 2004.

[5] Pagano, M. and Gauvreau, K. Principles of Biostatistics. 2nd ed. Duxbury, 2000.

[6] Resnik, P. *Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language*. Journal of Artificial Intelligence Research (JAIR), 11, 95-130, 1999.

[7] Rosch, E, Mervis C, Gray W, Johnson D, and Boyes-Braem P, *Basic objects in natural categories*. Cognitive Psychology vol. 8, 382-349, 1976.

[8] Schutze, H., and Silverstein, H. *Projections for efficient document clustering.* In Proceedings of the 20th International ACM SIGIR International Conference on Research and Development in Information Retrieval, 1997.

[9] Zobel, J., and Moffat, A. *Exploring the Similarity Space. Proc.* ACM SIGIR Forum, vol. 32, 18-34, Spring 1998.