Measuring Generality of Documents

Hyun Woong Shin[1], Eduard Hovy[2], Dennis McLeod[1], and Larry Pryor[3]
[1]Computer Science Department, Integrated Media Systems Center, University of Southern California
941 W. 37th Place, Los Angeles, CA, USA
[2]Computer Science Department, Institute of the Information Science, University of Southern California
4676 Admiralty Way, Marina del Rey, CA, USA
[3]Annenberg School for Communication, University of Southern California
Los Angeles, CA, USA
[1]{hyunshin, mcleod}@cs.usc.edu, [2]hovy@isi.edu, [3]lpryor@usc.edu

Abstract

Most traditional Information Retrieval (IR) systems, including web search engines, operationalize "relevant" as the word frequency in a document of a set of keywords. Because of this limitation, traditional IR systems frequently retrieve irrelevant documents in response to a user's request. In this paper, we propose a new criterion, "generality," that provides an additional basis on which to rank retrieved documents. The generality is a level of abstraction to retrieve results based on desired generality appropriate for a user's knowledge and interests. We compared our generality quantification algorithm with human judges' weighting of values to show that the developed algorithm is significantly correlated.

1. Introduction

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 operationalize "relevant" as the word frequency in a document of a set of keywords (or index terms) [5, 20, 23, 29]. They can therefore only retrieve information that contains the terms that are in the user's request, or terms easily derivable from it. As the TREC results show year after year, even the best IR systems' precision scores never average higher than 0.6. There have been several elaborations of this approach, including clustering [11, 14, 24, 28], topic mining [4, 8, 9], and ontologies [6, 10, 13]. 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.

The degree of generality is 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 that include 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.

In order to define and compute generality, we require a domain dependent ontology. This ontology, which consists of concept nodes and interrelationships [13], 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.

2. THE GENERALITY ALGORITHM

The basic idea behind quantifying of the degree of generality is that the degree of generality can be quantified by the number of index terms in the document that belong to specific word sets. The concept "specific word set" consists of index terms not belonging to any other ontology nodes. Here, generality is quantified by the appearance of specific index terms t_i within document

Proceedings of the 22nd International Conference on Data Engineering Workshops (ICDEW'06)
0-7695-2571-7/06 $20.00 © 2006 IEEE

D_j . The following is the formal definition of the algorithm.

Let $D = \{D_j \mid j \in J\}$ be a document containing a set of words from an index set J , and $S = \{t_i \mid i \in I\}$ be a set of specific terms from an index set I . The index set J is used to differentiate documents and the index set I contains specific words.

Define a characteristic function $\chi : I \times J \to \{0,1\}$

by $\chi(i,j) = \begin{cases} 1, t_i \in D_j \\ 0, t_i \notin D_j \end{cases}$.

By using the characteristic function, we define the generality of a document d_j as follows:

$g_j = \frac{\sum_{i \in I} \chi(i,j)}{|D_j|}$ for the number $|D_j|$ of terms in d_j

If $D_j \cap S = \varnothing$, then $g_j = 0$ as a special case.

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. In a domain dependent ontology, an instance of a child node is also an instance of a parent node. Thus, the parent node might provide its own instances and instances of children. The generality of its own, direct instances can be calculated utilizing the above algorithm. For the other instances, the algorithm needs to adjust the degree of generality based on their position in the concept hierarchy.

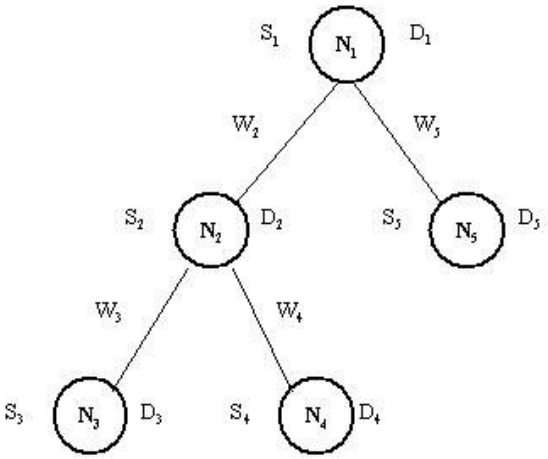

Figure 1. Sample ontology with weights w_i , index term lists D_i , and specific term lists S_i .

Figure 1 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.

The following is the formal definition for the adjustment algorithm:

Suppose that S_k and S_{k+i} are sets of specific terms of a parent node and its children nodes, respectively $for\ i = 1, 2, ..., c$ (number of the children node for parent node N_k . Let $N = \{N_k \mid k \in A\}$ be a set of nodes for an index set A and k is ordering by depth, and $\{g_{i,m} \mid i \in A\ and\ m \in I\}$ be a set of the degree of generality, $g_{i,m}$ for a document d_m in a node N_i . The adjusted generality of a document d_m on an ontology node N_k is defined as follows:

$d(k,m) = g_{k+i,m} + w_{k+i}$

where $w_{k+i} = \frac{|S_k - S_{k+i}|}{|D_k|}$ for the number $|D_k|$ of terms in a node N_k and some child node N_{k+i} containing the document d_m . The node N_k is a parent node of a child node N_{k+i} and the $|D_k|$ is the total number of index terms that are used in the parent node N_k .

3. EXPERIMENTS

Having developed a method to quantify generality, we now determine whether it conforms to human intuitions.

3.1 Corpus Analysis

Before starting work on the system, we collected and analyzed terms from a corpus to empirically guide the design of generality and generation of the domain dependent ontology. We studied the terms and conceptual hierarchy used to convey information from multiple sources including Associated Press, ESPN, and current newswire. We created a hierarchy of 12 nodes and three levels based on 62 articles. The domain ontology is delineated in Figure 2.

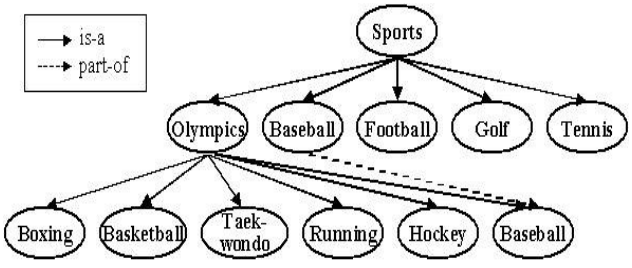

Figure 2. Domain dependent ontology

In order to generate the proper degree of generality, we analyzed the underlying corpus. Table 1 depicts the total number of index terms and specific words in the corpus.

In the table, $|D|$ indicates the total number of index terms at or below a node. The total number of specific terms for each node alone is in $|S|$. $|S'|$ is a summation of children node's specific terms

For the node "Sports", specific terms cannot be determined because this node is the summation of all other nodes. In the node, the number of specific terms for "Olympics" ($|S|$ = 976) is more than the summation of children's specific terms ($|S'|$ = 819). According to our assumption, it indicates that the node "Olympics" may contain more topics than the summation of topics in children nodes.

Table 1. Index terms and specific words

	$\|D\|$	$\|S\|$	$\|S'\|$
Sports	3148		2222
Olympics	1727	976	819
Baseball	757	303	
Football	960	420	
Golf	616	248	
Tennis	700	275	
Olympics-boxing	469	147	
Olympics-Basketball	424	111	
Olympics-Taekwondo	249	78	
Olympics-Running	597	222	
Olympics-Hockey	475	165	
Olympics-Baseball	381	96	

3.2 Evaluation Plan

Our verification of the measure of generality is performed between a domain dependent ontology 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 (as a continuous value) and assigned a score for each document based on their observation of the degree of generality. The judges were two graduate students in the school of journalism at USC and they were instructed that there are no right or wrong answers.

4. Experiment Results

The Pearson correlation coefficient always lies between -1 and +1($-1 \le r \le 1$), 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. Also, the significance of a correlation coefficient is examined by a t-test (n-2 degree of freedom).

Table 2. Pearson correlation coefficient between two human judges1

	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)

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 3. Pearson correlation coefficient1

	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)

Table 2 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 3 shows the Pearson coefficients and the corresponding p-values from the t-test between human judges and the algorithm. Figure 3 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 3, 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 4. As seen in Figure 4, the degree of generality between human judge and the algorithm is positively correlated except for some extreme cases.

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

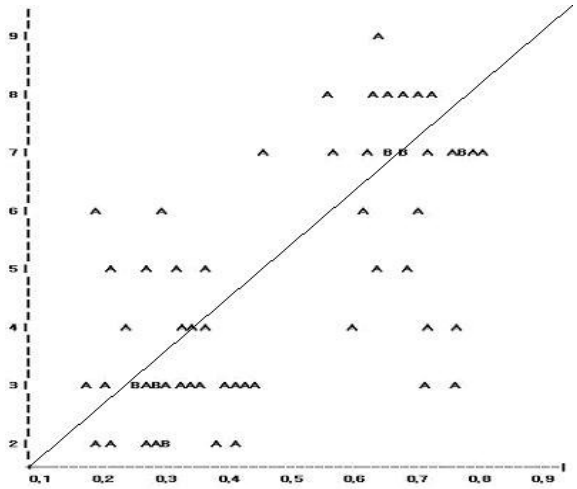

Figure 3. Scatter plot for the degree of generality between human judge and algorithm in Level 1

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.

5. Conclusion

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 have 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 3, 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.

6. References

[1] 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.

[2] Buckley, C. and Walz, J. SMART in TREC 8. Proc. Eighth Text Retrieval Conf., 577-582, November 1999.

[3] Chung, S., and McLeod, D. Dynamic topic mining form news stream data. In Proceedings of the 2nd International Conference on Ontologies, Databases, and Application of Semantics for Large Scale Information Systems, 2003.

[4] Chung, S., Jun, J., and McLeod, D. Incremental Mining from News Streams. Encyclopedia of Data Warehousing and Mining, Idea Group Inc. 2004.

[5] Hatzivassiloglou, V., Gravano, L., and Maganti, A. An investigation of linguistic features and clustering algorithms for topical document clustering. In Proceedings of the 23rd International ACM SIGIR International Conference on Research and Development in Information Retrieval, 2000.

[6] 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.

[7] Liu, X., Gong, Y., Xu, W., and Zhu, S. Document clustering with cluster refinement and model selection capabilities. In Proceedings of the 25th Annual International Conference on Research and Development in Information Retrieval, 2002.

[8] Salton, G., and Buckley, C. Term-weighting approaches in automatic text retrieval. Information Processing and Management, 24(5), 513-523, 1988.

[9] Salton, G., Automatic Text Processing – the Transformation, Analysis and Retrieval of Information by Computer. Addison-Wesley Publishing Co., Reading, MA, 1989.

[10] 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.

[11] Zhao, Y., and Karypis, G. Evaluations of hierarchical clustering algorithms for document datasets. In Proceedings of the 11th International ACM International Conference on Information and Knowledge Management, 2002.

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