

Generality: A New Criterion for Measuring Generality of Documents

Hyun Woong Shin

Computer Science
Department

Integrated Media Systems
Center

University of Southern
California

+1-213-740-3696

hws@usc.edu

Eduard Hovy

Computer Science
Department

Institute of the Information
Science

University of Southern
California

+1-310-448-8731

hovy@isi.edu

Dennis McLeod

Computer Science
Department

Integrated Media Systems
Center

University of Southern
California

+1-213-740-4504

mcleod@usc.edu

Larry Pryor

USC/Annenberg School for
Communication

University of Southern
California

+1-213-740-9083

lpryor@usc.edu

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.

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) [5, 20, 23, 29]. 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 [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.

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 [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.

The proposed concept of generality has been rarely researched to date. The most similar study related to generality is conducted by Resnik [18] 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

$$sim(c_1, c_2) = \max_{c \in S(c_1, c_2)} [-\log p(c)].$$

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 reasoning behind generality quantification. The generality algorithm is presented in Section 3. Section 4 describes the experiments. The experimental results are discussed in Section 5. Section 6 concludes this study.

2. TOWARD QUANTIFYING GENERALITY

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, $S_i = T_i - \bigcup_{j \neq i} T_j$.

Identification of the index terms is one of the core parts of this algorithm. The identification part is a subset of indexing. Indexing is the extraction and weighting of index terms from the text documents in order to represent the document’s content, and has been the subject of much study in IR [21, 23]. The most well established techniques rely on exploiting statistical information about term occurrence. According to the most widely accepted indexing

technique [3], indexing is divided into the following steps.

- term extraction
- stop word removal
- stemming and conflation
- index term weighting

Term extraction consists of breaking the full text of a document up into words (called index terms). The index terms usually include non-informative words. The stop word removal phase removes non-informative words such as prepositions, articles, and conjunctions by utilizing stop word list [26]. Index terms also generally word stems or roots rather than full words, since this means that matches are not missed because of trivial word variations such as with singular/plural forms or verb tenses. Stemming and conflation are carried out by using a standard stemming algorithms and suffix lists. A number of stemming algorithms exist [12, 17, 25]; we chose Porter’s algorithm [17], one of the most widely accepted ones. The final phase is weighting the index terms. The basic idea of weighting index term is discrimination. A good index term should distinguish the particular document from all others. A number of weighting schemes are reported [22, 27]. However, we do not use any weighting scheme in our study.

3. THE GENERALITY ALGORITHM

In Section 2, we defined the basic idea behind quantifying of the degree of generality. For its quantification, we introduced the concept “specific word set” that 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 D_j .

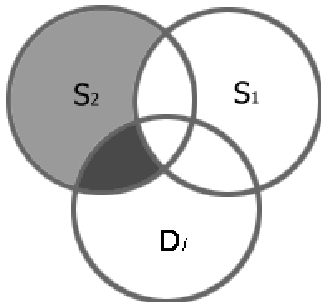


Figure 1. The degree of generality of the document D_j

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 $((S_2 \cap D_j) - S_1)$. 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 $\chi : I \times J \rightarrow \{0, 1\}$

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

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

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

If $D_j \cap S = \emptyset$, 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. 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”.

In many IR systems that utilize ontologies, a user submits a query, and the system determines the proper ontology node(s) that is (are) best matched with a given query. After the appropriate node(s) is (are) determined, the system retrieves the information from the concept hierarchy that is rooted at the chosen node. 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. 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. In Figure 2, “Sports” is a parent node of the child node “Baseball”. If there is a document d_1 belonging to the child node “Baseball” with degree of generality 5, the degree of generality for the document d_1 of the parent node “Sports” should be higher than 5. This assumption we base on the intuition that domain-specific Ontologies will generally lie mostly below the level of Basic Concepts [19], and hence tend to become of more general interest as one moves upward through them.

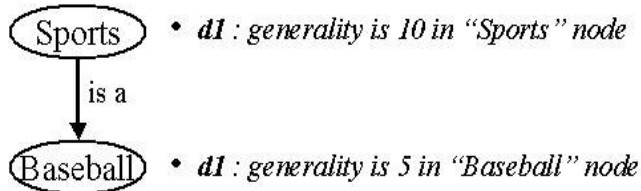


Figure 2. Example of generality in a concept hierarchy

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.

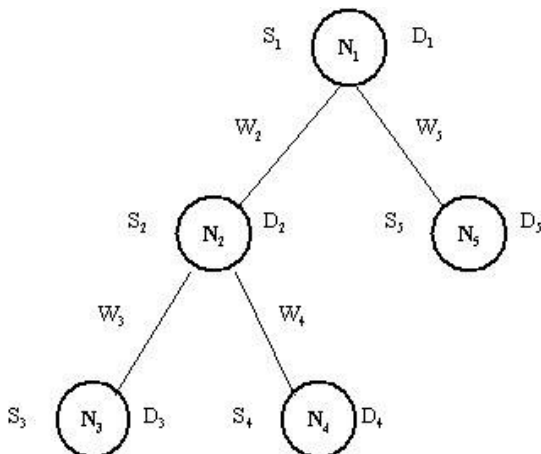


Figure 3. Sample ontology with weights w_i , index term lists D_i , and specific term lists S_i

Figure 3 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 $i=1,2,\dots,c$ (number of the children node for parent node N_k). Let $N = \{N_k | k \in A\}$ be a set of nodes for an index set A and k is ordering by depth, and $\{g_{i,m} | i \in A \text{ and } m \in I\}$ 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 .

4. EXPERIMENTS

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

4.1 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 4.

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. It is safe to say that $|S| \geq |S'|$ between a parent node and children nodes because some specific words $|S|$ of the parent node are general words for children nodes. For example, the parent node "Olympics" contains "Olympiad", "Gold", "Bronze", and some country names that repeat in children nodes. Those specific words, however, are not included in the specific word set for children nodes.

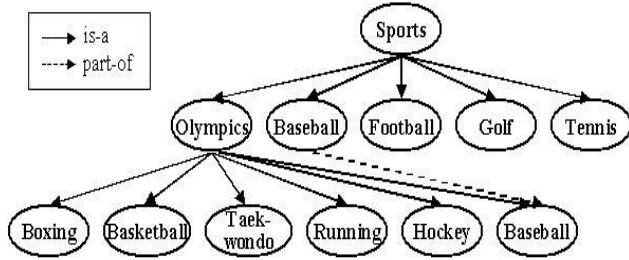


Figure 4. Domain dependent ontology

For the node "Sports", specific terms cannot be determined because this node is the summation of all other nodes. In the node "Olympics", the number of specific terms ($|S| = 976$) 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	

4.2 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 instructed that there are no right or wrong answers.

Since the human judges' values are critical to evaluate the algorithm, the selection of human judges is a huge problem. How many human judges are necessary? How should we adjust for individual variability? To lessen the problem, professional journalists were chosen as human judges. If random people from the general population were included, individual variation would be huge due to their backgrounds and education levels. This variation would increase the standard deviation and bring errors into the study. However, professional journalists are trained to obtain proper writing and reading skills in terms of journalism, so we assume that the individual variation among them in scoring new articles is not too large. The judges are a professor and a graduate student in the school of journalism at USC. We assumed that they already have developed adequate comprehension and writing skills, so no training session was carried out.

As we mentioned before, we measured the relationships between the values obtained from the human judges and from the algorithm, quantified by the correlation ρ (rho) as follows:

$$\rho = \text{average} \left[\frac{(X - \mu_x)(Y - \mu_y)}{\sigma_x \sigma_y} \right]$$

where X is the degree of generality from the experiment,

Y is the degree of generality assigned by each judge.

μ_x is the average of X ,
 μ_y is the average of Y ,
 σ_x is the standard deviation of X , and
 σ_y is the standard deviation of Y .

We used the Pearson correlation coefficients 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.

5. RESULTS AND DISCUSSION

The Pearson correlation coefficient always lies between -1 and +1 $-1 \leq r \leq 1$, 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).

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 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)

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. 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 3 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 5. Figure 6 also shows a similar pattern between the algorithm and the human judges, where the X-axis presents each document and the Y-axis presents the degree of generality. As seen in Figure 6, 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

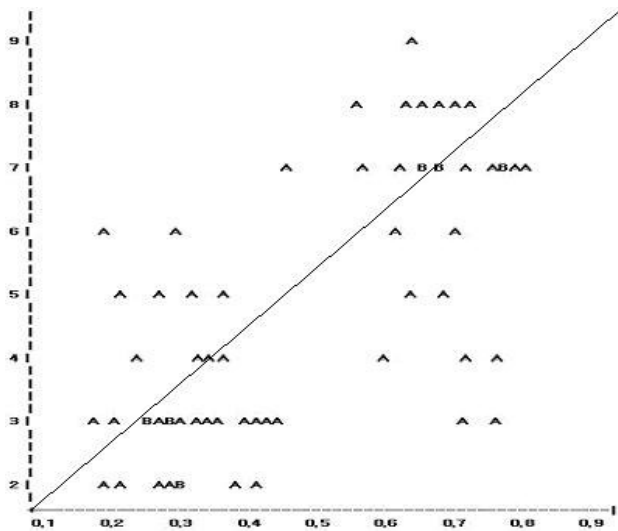


Figure 5. Scatter plot for the degree of generality between human judge and algorithm in Level 1

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.

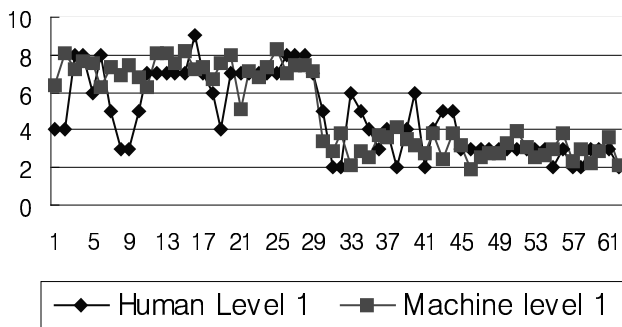


Figure 6. The degree of generality between the human judge and algorithm in Level 1

6. 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 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.

7. 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.

8. REFERENCES

- [1] Baeza-Yates, R., and Ribeiro-Neto, B. *Modern Information Retrieval*. Addison Wesley Longman, 1999.
- [2] Belkin, N.J., and Croft, W.B. *Information Filtering and Information Retrieval: Two Sides of the Same Coin*. Commun. ACM 35, 29-38, December 1992.
- [3] Boyce, B., Meadow, C., and Kraft, C. *Measurement in Information Science*. New York: Academic Press, 1994
- [4] 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.
- [5] Buckley, C. and Walz, J. *SMART in TREC 8*. Proc. Eighth Text Retrieval Conf., 577-582, November 1999.
- [6] Bunge, M. *Treatise on basic Philosophy, Ontology I: The Furniture of the World*. Reidel Publishing Co., vol. 3, 1977.
- [7] Cha-Hwa, L., and McLeod, D., *Exploiting and Learning Human Temperaments for Customized*

- Information Recommendation*. Proceedings of the 6th IASTED International Conference on Internet and Multimedia Systems and Applications, Kauai, Hawaii, August 2002.
- [8] Chung, S., and McLeod, D. *Dynamic topic mining from news stream data*. In Proceedings of the 2nd International Conference on Ontologies, Databases, and Application of Semantics for Large Scale Information Systems, 2003.
- [9] Chung, S., Jun, J., and McLeod, D. *Incremental Mining from News Streams*. Encyclopedia of Data Warehousing and Mining, Idea Group Inc. 2004.
- [10] Gruber T.R. *Toward Principles for the design of Ontologies used for Knowledge Sharing*. In Proceedings of the International Workshop on Formal Ontology, 1993.
- [11] 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.
- [12] Jacquemin, C. *Guessing morphology from terms and corpora*. In: Actes, 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR97), 156-167, 1997.
- [13] 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.
- [14] 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.
- [15] Magnini, B., and Strapparava, C. *Improving user modeling with content-based techniques*. In Bauer, M, Vassileva, J, and Gmytrasiewicz, P. (Eds). User Modeling Eighth International Conference UM2001, 74-83, Berlin, Springer, 2001.
- [16] Pagano, M. and Gauvreau, K. Principles of Biostatistics. 2nd ed. Duxbury, 2000.
- [17] Porter, M. *An Algorithm for suffix stripping*. Program, 14(3), 130-137, 1980.
- [18] 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.
- [19] Rosch, E, Mervis C, Gray W, Johnson D, and Boyes-Braem P, *Basic objects in natural categories*. Cognitive Psychology vol. 8, 382-349, 1976.
- [20] Salton, G., and Buckley, C. *Term-weighting approaches in automatic text retrieval*. Information Processing and Management, 24(5), 513-523, 1988.
- [21] Salton, G., and McGill, M.J. *Introduction to Modern Information Retrieval*. Mcgraw Hill, 1983.
- [22] Salton, G., and Wong, A. *A Vector Space Model for Automatic Indexing*. Communications of the ACM 18, 1975.
- [23] Salton, G., *Automatic Text Processing – the Transformation, Analysis and Retrieval of Information by Computer*. Addison-Wesley Publishing Co., Reading, MA, 1989.
- [24] 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.
- [25] Xu, J., and Croft, B.W. *Corpus-based stemming using co-occurrence of word variants*. ACM Transactions on Information Systems, 16(1), 61-81, 1998.
- [26] Yang, Y., Chute, C.G. *An example-based mapping method for text categorization and retrieval*. ACM Transaction on Information Systems (TOIS 94), 253-277, 1994.
- [27] Yu, C.T., Lam, K., and Salton, G. *Term Weighting in Information Retrieval Using the Term Precision Model*. Journal of the Association for Computing Machinery, Vol 29, No I, January 1982.
- [28] 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.

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