



## COVER SHEET

---

**Desai, Monica and Spink, Amanda (2005) An algorithm to cluster documents based on relevance. Information Processing and Management 41(5):pp. 1035-1049.**

**Copyright 2005 Elsevier.**

Accessed from: <http://eprints.qut.edu.au/archive/00004753/>

## AN ALGORITHM TO CLUSTER DOCUMENTS BASED ON RELEVANCE

Monica Desai  
Department of Computing Science and Engineering  
The Pennsylvania State University  
220 Pond Laboratories, University Park, PA 16802  
Email: [mdesai@cse.psu.edu](mailto:mdesai@cse.psu.edu)

Amanda Spink  
School of Information Sciences  
University of Pittsburgh  
610 IS Building, 135 N. Bellefield Avenue  
Tel: (412) 624-9454  
Email: [aspink@sis.pitt.edu](mailto:aspink@sis.pitt.edu)

## ABSTRACT

Search engines fail to make a clear distinction between items of varying relevance when presenting search results to users. Instead, they rely on the user of the system to estimate which items are relevant, partially relevant, or not relevant. The user of the system is given the task of distinguishing between documents that are relevant to different degrees. This process often hinders the accessibility of relevant or partially relevant documents, particularly when the results set is large and documents of varying relevance are scattered throughout the set. In this paper, we present a clustering scheme that groups documents within relevant, partially relevant, and not relevant regions for a given search. A clustering algorithm accomplishes the task of clustering documents based on relevance. The clusters were evaluated by end-users issuing categorical, interval, and descriptive relevance judgments for the documents returned from a search. The degree of overlap between users and the system for each of the clustered regions was measured to determine the overall effectiveness of the algorithm. This research showed that clustering documents on the Web by regions of relevance is highly necessary and quite feasible.

## INTRODUCTION

The standard IR system paradigm models relevance as a dichotomous entity. Indeed, the bipolar nature of relevance is incorporated and reflected in a number of areas within IR research. However, relevance has been identified as a non-dichotomous concept (Spink, Greisdorf & Bateman, 1998). Items may be relevant, partially relevant, or not relevant. There exists a conspicuous middle range in between the extreme ends of the relevance spectrum that needs to be accounted for. However, search engines presenting items on the commonly used one-dimensional ranked list do not make the distinction between documents of varying grades of relevance. The user must guess the cutoff point between highly relevant and partially relevant items without any clear indication of how many documents should be examined within the list returned. With search engines commonly returning thousands of results, this task becomes almost impossible. Access to the most highly relevant documents in a ranked list is limited by the lack of clear boundaries delineated for each region of relevance. To solve this problem, the regions of relevance should be clearly delineated on a system level.

This paper presents a clustering scheme to group documents by relevance so that documents within relevant, partially relevant, and not relevant regions are explicitly identified and displayed for a given search. An algorithm was devised to distinguish between the relevance of documents presented in a one-dimensional ranked list and to determine the feasibility of clustering the documents based on relevance. Based on the clustering decisions made by the algorithm and user relevance judgments for each item, the degree of overlap between users and the system was measured.

The search queries were inputted on the World Wide Web and an interface displays the resultant documents grouped in clustered format. To evaluate the clusters, end-users issued categorical, interval, and descriptive relevance judgments for the documents returned from the

search. The degree of overlap between users and the system, along with the overall effectiveness of the algorithm was measured for each of the clustered regions.

## RELATED STUDIES

Few previous studies have developed and tested a clustering scheme based on relevance regions. Previous studies have identified the existence of documents of varying degrees of relevance. Relevance is not a concrete binary concept but a fuzzy concept (Spink & Greisdorf, 1997). Different documents may belong to one set only to a certain degree. However, systems typically collapse results into two sets, in which one set combines partially relevant items with highly relevant items, and the other set consists of non-relevant items (Spink, Greisdorf & Bateman, 1998).

Rees and Shultz (1967) determined that a simple two-point scale is insufficient and inappropriately collapses “a variety of degrees of relevance into yes/no decisions”. Sperber and Wilson (1986) emphasized that relevance should be considered in terms of degrees, since the presence or absence of relevance is not absolute. Certainly, a conspicuous middle region exists in between the two absolutes. Wallis and Thom (1996) stress that it is not simple to convey relevance in terms of degrees but emphasize the need to retrieve material that is both relevant and partially relevant. Binary assessments hide the variability, complexity and continuity of relevance (Schamber, 1994).

Since relevance is a non-dichotomous concept, it is important to reflect this in the system’s ranking and evaluation. Indeed, it is important that highly relevant documents are isolated from those documents that are marginally relevant. This would allow easy access to the documents that are most highly relevant to the information need. Likewise, partially relevant items should be separated from non-relevant items, as partially relevant documents can also play an important role to certain users.

Specifically, partially relevant documents can provide users with new information, shifting them towards new directions (Spink, Greisdorf & Bateman, 1998). Novice users can often utilize information obtained from partially relevant documents to lead them through further stages of the information seeking process toward a possible resolution of their information problem (Spink & Greisdorf, 1997). Partially relevant documents can help users redefine their initial query to obtain the results they are looking for.

In TREC, the need to use a non-binary scale to effectively retrieve documents has come to the forefront. Sormunen introduced a four-point relevance scale to reassess the document pools in TREC-7 and TREC-8 to distinguish between highly relevant documents rich in topical information, and marginally relevant documents poor in topical information (Sormunen, 2002). The study found that about 50% of documents assessed as relevant in TREC were actually marginally relevant and of the remaining half, only 16% of documents deemed relevant were actually highly relevant (Sormunen, 2002). Thus, the study emphasizes that relevant documents can consist of both highly and somewhat relevant.

Past research has heavily emphasized the need for clustering within IR systems. Indeed, there has been extensive research on how clustering can be used to improve retrieval. Traditionally, clustering was introduced for efficiency of retrieval since matching a query against a centroid might be more efficient than matching against the entire collection (Hearst, Pedersen, 1996). Many clustering techniques build on the cluster hypothesis, which states that relevant documents tend to be more similar to each other than to non-relevant documents. Statistical clustering algorithms form clusters based on topical similarities between documents among a set of retrieved documents. The Scatter/Gather method clusters documents by topic, providing an alternative to viewing results in traditional ranked lists (Hearst, Pedersen, 1996). In this paradigm, documents are clustered into topically-coherent groups that are displayed through summaries consist of topical terms and titles characterizing the contents of the cluster.

(Hearst, Pedersen, 1996). Past clustering techniques emphasize the matching of the query to cluster centroids (Willett, 1998). Through a hierarchical approach, a query could be compared against each cluster from from the top-down or bottom-up (Willett, 1998). A similarity score generated between the query and the centroid would determine the ranking of the clusters to be displayed.

While most clustering schemes emphaize topicality as a means of grouping documents, our technique strives to group documents according to how relevant they are to the user. Instead of forming cluster centroids and obtaining a similarity score between the query and the centroid, our scheme provides a score for individual documents based on pre-defined document text characteristics that associate a document as being relevant, partially relevant, and non-relevant. These scores are then grouped together according to a cutoff point that determines the score necesseary for a document to be displayed in a certain cluster.

## RESEARCH DESIGN

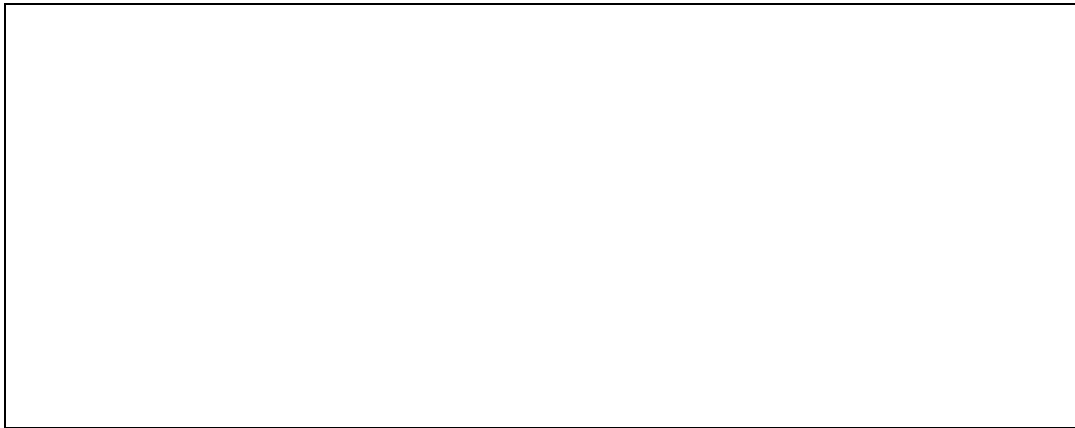
The intent of this study is to incorporate the non-dichotomous nature of relevance within search engines. The fundamental research questions that this study attempts to answer include: Can the middle range of relevance be identified to enhance the output of traditional IR systems? Are there certain characteristics inherent within partially relevant documents that can help identify this middle region? If the partially relevant documents can be extracted, can we also distinguish relevant and non-relevant documents in a results set? How successful are the clusters in grouping results that match user judgments and in guiding users towards uncovering documents that correspond to their needs? To answer these questions, we have designed a system that consists of two main components. The system component produces the clustered results through a series of automated steps, while the user element involves end-users generating relevance judgments and creating clusters through a manual approach. Both of

these clustered results can then be compared to determine how well the algorithm clustered the results.

### Algorithm

The documents returned in a ranked list produced by AlltheWeb.com will be clustered according to the specifications set forth by the algorithm. The algorithm decides which cluster the documents belong in, and makes the distinction between highly relevant, partially relevant, and not relevant documents. The algorithm utilizes similarity measures and ranking heuristics that evaluate the relevance of a page. The high-level workflow of the algorithm is in Figure 1.

Figure 1. Workflow of algorithm.



Each document in the collection is evaluated based on key similarity metrics and ranking heuristics. For each field, a relevancy grade is recorded for the document based on its satisfaction of the criteria listed for the given field and grade. A ranking function produces an overall score that combines the grades with the weight for each field for a given document and is explicitly detailed in a later section. The resultant score determines the cluster the document belongs in. The fields, weights, ranking criteria, and relevancy grades used to deduce a score for each item is illustrated in Table 1.

Table 1 SEQ Table \\* ARABIC 1. Fields, weights, relevancy criteria.

Id (i)	Field Name	Weight	Relevancy Grade ( $\lambda$ )		
			[3]	[2]	[1]
1	Term Frequency	$[W_1]$	$[C_1\theta, \theta]$ query terms each appear with frequency $Y_3$	$[C_1\theta, \theta]$ query terms each appear with frequency $Y_2$	$[C_1\theta, \theta]$ query terms each appear with frequency $Y_1$
2	Title	$[W_1]$	$[C_1\theta, \theta]$ query terms appear	$[C_2\theta, C_1\theta]$ query terms appear	$[0, C_2\theta]$ query terms appear
3	URL	$[W_1]$	$(1, \theta]$ query terms appear	$[1, 1]$ query terms appear	$[0, 0]$ query terms appear
4	Anchor Text	$[W_2]$	$[C_1\theta, \theta]$ query terms appear	$[C_2\theta, C_1\theta]$ query terms appear	$[0, C_2\theta]$ query terms appear
5	Location	$[W_3]$	Max query terms in Top 1/3 region	Max query terms in Mid 1/3 region	Max query terms in Bottom 1/3 region
6	Emphatic Text	$[W_3]$	$[C_1\theta, \theta]$ query terms appear	$[C_2\theta, C_1\theta]$ query terms appear	$[0, C_2\theta]$ query terms appear
7	Headers	$[W_3]$	$[C_1\theta, \theta]$ query terms appear	$[C_2\theta, C_1\theta]$ query terms appear	$[0, C_2\theta]$ query terms appear

In Table 1, key fields within each document are identified and assigned a weight and relevancy grade. The motivation for choosing these fields is further identified in the next section.  $\theta$  represents the maximum number of words in the end-user query minus words that have no meaning. It is assumed that more than one key term is entered so that  $\theta > 1$ .  $C_1$  and  $C_2$  represent constants with the values  $2/3$  and  $1/3$  respectively. Thus,  $C_1\theta$  and  $C_2\theta$  represent fractions of the number of query terms. In all cases, the resulting value is rounded to the nearest whole number. Thus, the notation  $[C_1\theta, \theta]$ , for example, represents the range in the number of query terms that must be included within the specified field.  $Y_3, Y_2,$  and  $Y_1$  represent the number of occurrences of a query term in a document. The values used for this trial were  $Y_3 \geq 3, 2 \leq Y_2 < 3, 0 \leq Y_1 < 2$ . Note that  $W_1, W_2,$  and  $W_3$  are weights used to assign an importance value to each field, and were set to 0.214, 0.142, and 0.072 respectively.

As an example, suppose the end-user query consists of 6 distinct terms. Thus,  $\theta$  is 6,  $C_1\theta$  is 4, and  $C_2\theta$  is 2 since  $C_1$  and  $C_2$  are set to  $2/3$  and  $1/3$  respectively. For the term frequency category, the document returned for the given query must contain between 4 and 6 query terms inclusive with each term occurring with a frequency of at least  $Y_3$  in order to receive a relevant grade. To receive a partially relevant grade, the document must contain between 4 and 6 query terms with each term occurring with a frequency of  $Y_2$ . A grade of not relevant is given as the default case for a document that does not satisfy the relevant or partially relevant categories. Similarly, the document receives a relevant, partially relevant, or not relevant grade for each of the remaining fields. The motivation for using these fields, weights, constants, and criteria are delineated in the sections below.

### Fields

The HTML makeup of page contains key fields that can indicate the importance of the document and improve retrieval. Intuitively, the title, six headings, and emphasized text such as bold, underline, and italic provide useful information about the page (Cutler, Shih & Meng, 1997). Another heuristic that can play a significant role in retrieval effectiveness is location (Notess, 1999). The idea behind location is that a term near the beginning of the page may carry greater significance than terms lower on the page (Notess, 1999). For term frequency, if a term occurs many times in the document, it represents the importance of the term within the page and may symbolize the importance of the term in the document.

### Constants

$C_1$  and  $C_2$  were used to account for variation in the number of query terms to include since queries may contain multiple words. In our scheme, including these constants allows for flexibility in the evaluation of query terms, since the user may repeat or add multiple terms in the query with the same meaning. Specifically, setting  $C_1$  to  $2/3$  and  $C_2$  to  $1/3$  provides three

ranges that can be used to represent regions of relevance within our scoring breakdown, based on ad-hoc common sense.

### Relevancy Grades

The relevance grades used in this study are derived from the notion of multi-graded relevance that is amply evident previous work. For each heuristic, a top grade is given assuming the document satisfies the necessary requirements to the fullest. This stringent criteria for each field is visible down the leftmost column of Table 1 under grade three. Similarly, a satisfactory grade is given when a document only partially satisfies the criteria. The requirements across each field are evident in the middle column of Table one under grade two. Finally, a low grade, which is displayed in Table 1 under grade one, characterizes non-relevant documents that either fail to meet the ranking criteria. Many documents may receive conflicting grades by satisfying relevant criteria in some cases, and partially or non-relevant criteria in other cases. Thus, these scores are aggregated into a weighted ranking function that combines individual scores as a weighted average to predict the most fitting category for the document, based on nature of relevance criteria within the document.

### Weights

Each measure and heuristic is given a weight to assign an appropriate importance value to the field. This value represents how much weight the field carries in assessing the relevance of a document (Cutler, Shih & Meng, 1997) and is included in our clustering scheme. In our algorithm, the assignment of weights to each heuristic is based on tiered-approach, where fields that are equally important are grouped together based on ad-hoc common sense and given a proportional weight in comparison to the other tiers.

### Ranking Criteria

To determine the nature of the criteria in Table 1 that best fits relevant, partially relevant, and non-relevant documents, previous work on document text characteristics was applied.

Since our algorithm attempts to cluster according to regions of relevance, characteristics of relevant, partially relevant, and non-relevant regions serve as decisive factors within our ranking function. Highly relevant pages tend to discuss the topic at length, deal with several aspects of the topic, have many terms that pertain to the requested topic, and have many expressions to refer to the concepts discussed (Sormunen, Kekalainen, Koivisto & Jarvelin, 2001). Indeed, highly relevant documents often answer the user's question, include the user's search terms or concepts, are specific to the user's query, and are authoritative sources (Spink, Greisdorf & Bateman, 1998).

In contrast, partially relevant items tend to mention the topic only briefly. They contain only a few words matching the topic, and may discuss the topic from alternative viewpoints extending upon the original request (Sormunen, Kekalainen, Koivisto & Jarvelin, 2001). They often deal on partially with the subject, are not specific to the user's query, and contain multiple concepts (Spink, Greisdorf & Bateman, 1998).

Finally, Non-relevant documents are often totally off target (Sormunen, Kekalainen, Koivisto & Jarvelin, 2001). This description of relevant documents, partially relevant, and non-relevant items can be translated into specified criteria that these classes of documents possess, as described in Table 1.

### Ranking Function

Scores for each field in a document are aggregated to achieve a total overall score based on the weight of each field, and satisfaction of the criteria for each field. A weighted average consists of an estimation of the importance of every ranking factor through a weight proportional to the projected value (Rapela, 2001). Thus, our ranking function combines the weights and relevancy grades received by a given document for each factor. The overall score is calculated as:

$$sc(q, d) = \sum_{i=1}^n W_i * \lambda_i \quad (1)$$

where  $n$  represents the total number of ranking factors,  $W$  is the weight of each factor,  $\lambda$  is the relevancy grade received by a document for each factor,  $q$  is the query, and  $d$  represents the document. The score represents the category that best suits the document, and can fall in any one of three possible regions depending upon the characteristics of the document itself. The score category result for the document will fall in one of the following clusters by converting  $sc(q,d)$  to a region of relevance, namely  $Cluster(q,d)$ .

$$Cluster(q, d) = \begin{cases} R & \text{if } sc(q, d) \geq \sigma f_1 \\ PR & \text{if } \sigma f_2 \leq sc(q, d) < \sigma f_1 \\ NR & \text{if } sc(q, d) < \sigma f_2 \end{cases} \quad (2)$$

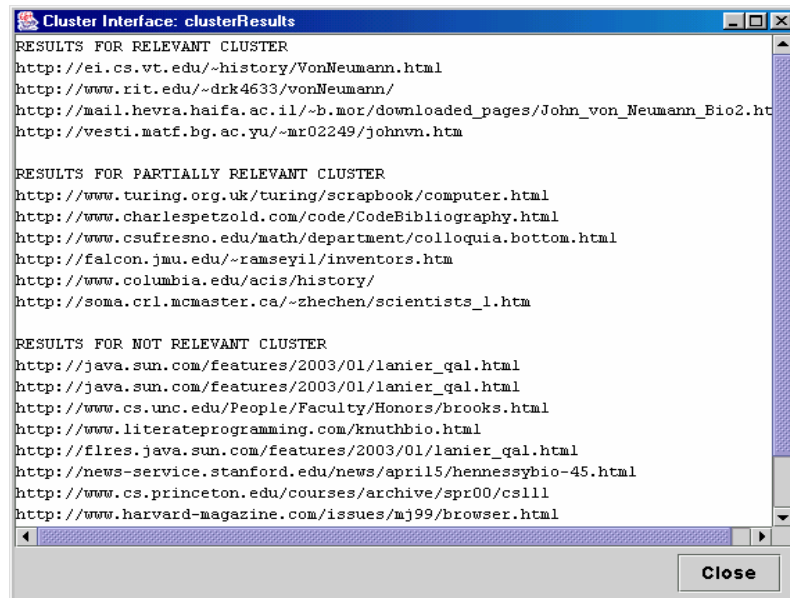
The constant  $\sigma$  represents the maximum score possible from  $sc(q,d)$ ,  $f_1$  represents a constant factor of the maximum, and  $f_2$  represents a second constant factor of the maximum. The settings for the values used in our study allow for equal ranges that the score can fall within for each of the three regions of relevance. The constant values are 3 for  $\sigma$  since the maximum possible score according to  $sc(q,d)$  is 3,  $7/9$  for  $f_1$ , and  $5/9$  for  $f_2$ . The lowest possible score according to  $sc(q,d)$  is 1. As a result of this equation, every document  $d$  for a query  $q$  will be placed in either the relevant, partially relevant, or non-relevant cluster.

### Cluster Interface

Figure 2 displays a sample cluster interface that is returned by the system for a query. This interface illustrates how the document URLs should be displayed to the user once the user submits a query. The clusters delineate the region of relevance each document belongs in. Documents within a specific cluster are not grouped internally according to relevance. Note that this interface represents the output of the system and was not shown to users for evaluation purposes. Instead, users were presented with the original list of results produced by

AlltheWeb.com to make their judgments to ensure that the clusters formed by the algorithm do not influence the judgments of users.

Figure 2. Sample output interface.



## Relevance Data Collection

### Study Participants

Users' relevance judgments provided the basis for determining the success of the algorithm, as the clusters created by the users could be matched with the clusters created by the system. The data analyzed in our study was gathered from five end-user computer science undergraduate and graduate students at the Pennsylvania State University within the department of Computer Science and Engineering during the spring semester 2003. While they did not have any formal training in IR evaluation, they were all Computer Science and Engineering students searching a topic related to Computer Science.

The end-users were provided with a ranked list of search results for a predetermined topic. The topic chosen in this study was exactly stated as "biography of computer science pioneer

John Von Neumann.” This topic was chosen since it is unambiguous, intelligible, and returned a suitable number of results. This topic yielded a total of 98 results. All users evaluated the same set of search results produced by AlltheWeb.com for the same exact query. Search results were saved and compiled on a web page that provided the URLs that link to each resultant page in the order they were produced from AlltheWeb.com. We ran our system to produce the clusters 15 minutes before users conducted the study to maintain consistency.

### Relevance Worksheet

Volunteers were given a worksheet on which they could indicate and describe their relevance judgments for the given topic. This worksheet first developed by Spink, Batemen and Greisdorf (1998) is shown in Figure 3.

Figure 3. Worksheet for user relevance judgments.

A	B	C	D	E	F
ITEM#	RELEVANCE	JUDGEMENTS			DESCRIBE
	(place vertical line indicating how relevant this item is)	(check one box only)			
		NR	PR	R	
4	NR  -----  R				
5	NR  -----  R				
6	NR  -----  R				
7	NR  -----  R				
8	NR  -----  R				
9	NR  -----  R				
10	NR  -----  R				
11	NR  -----  R				
12	NR  -----  R				
13	NR  -----  R				
14	NR  -----  R				
15	NR  -----  R				
16	NR  -----  R				
17	NR  -----  R				
18	NR  -----  R				
19	NR  -----  R				
20	NR  -----  R				
21	NR  -----  R				
22	NR  -----  R				
23	NR  -----  R				

The first measure on the worksheet provided an interval measure of the users' relevance judgments on a 77-mm line ranging from not relevant (NR) to relevant (R). The next measure on the worksheet provided a categorical measure of users' relevance judgments and was comprised of three boxes labeled relevant, partially relevant, and not relevant. The third measure

on the form allowed users to explain “why” they made their judgments through a brief description.

## RESULTS

Measuring the degree of overlap between the user-generated clusters and the system-generated clusters serves as a key measure in revealing the algorithm’s effectiveness in classifying the results.

### Relevance Judgments

Our data collection consisted of six sets of clusters, with one system-generated set, and five user-derived sets based on the relevance judgments the end-users made. Evaluating the degree of overlap on a cluster-by-cluster basis provides an underlying measure of how well the algorithm performed. Table 2 displays the percentage of overlap between the system and end-user relevance judgments for each cluster.

Table 2. Percentage overlaps between system-generated clusters and user relevance judgments.

	% Relevant (R) Judgments Issued	% Partially Relevant (PR) Judgments Issued	% Not Relevant (NR) Judgments Issued
System’s ‘Relevant’ Cluster	87%	13%	0%
System’s ‘Partially Relevant’ Cluster	5%	69%	26%
System’s ‘Not Relevant’ Cluster	0%	13%	87%

The relevance judgments made by each user for documents within the system’s relevant, partially relevant, and not relevant clusters represents the variation of the system compared to each individual user. This variation is depicted in Table 3.

Table 3. Variation of user judgments for relevant, partially relevant, and not relevant clusters.

Users	No. of Judgments issued for documents within <i>Relevant Cluster</i>			No. of Judgments issued for documents within <i>Partially Relevant Cluster</i>			No. of Judgments issued for documents within <i>Not Relevant Cluster</i>		
	R	PR	NR	R	PR	NR	R	PR	NR
1	8	1	0	1	15	4	0	10	59
2	8	1	0	0	16	4	0	2	67
3	8	1	0	2	3	15	0	7	62
4	8	1	0	1	17	21	0	17	52
5	7	2	0	1	18	1	0	9	60

#### Overlap: Relevant Cluster

The first region considered here is the relevant cluster. Each URL within the system's relevant cluster is matched with the number of users classifying that URL as relevant, partially relevant, or not relevant. In this case, the system determined that a total of nine documents out of the 98 total returned from the search belonged in the relevant cluster. Thus, 45 collective user judgments made by the five end-users were considered within this cluster. Of these pooled judgments, the data revealed that 39 out of the 45 total relevance judgments were marked as relevant. As a result, 87% of end-user relevance judgments were also relevant in agreement with the system. On the other hand, only six out of 45, or 13% of the judgments, varied from the system-generated results. The extent of agreement between the system and all five end-users for the relevant cluster is depicted in Table 2. This was a binomial distribution with five subjects with the mean number of respondents for whom algorithm the matched correctly being 4.333.

In calculating these statistics, it is assumed that each URL selected and evaluated by a given user is independent from other selections. The variance in the number of users who produced matching results is 0.578 and the standard deviation is 0.767. From the data, it is evident that unanimous agreement between the system and all five end users existed for seven out of the nine total documents. The relevance judgments made by each user for documents

within the system's relevant cluster represents the variation of the system compared to each individual user. This variation is depicted in Table 3.

The six total judgments that differed from the system varied only by a single relevance category. The high proportion of matching end-user judgments combined with the low variance and standard deviation indicate the relative success of the algorithm in correctly clustering relevant results.

#### Overlap: Partially Relevant Cluster

Besides the relevant cluster, the degree of overlap between the system and users within the partially relevant cluster was determined. The system determined that 20 documents out of the 98 total returned from the search were indeed partially relevant, accounting for exactly 100 user judgments made for documents within the cluster. Among the 100 user judgments issued in this cluster, 69 were also found to be partially relevant and overlapped with the system's classification.

Thus, the system overlapped with 69% of end-user relevance judgments within the partially relevant cluster. Only five out of 100 judgments, or 5% of the user judgments were deemed relevant, while 26 judgments, or 26% were classified as not relevant. Thus, 31% of user judgments varied from the system's classification. The overall agreement between the system and end-users for the partially relevant cluster is table 2. The mean number of end-users that matched the output of the system was 3.45 out of five for this cluster. In addition, the variance in the number who produced matching results is 1.067 and the standard deviation is 1.034. Thus, the variance and deviation within these results is higher than that of the relevant cluster. Yet, the heaviest concentration of judgments overwhelmingly remains within the partially relevant region. For this cluster, 16 out of 20 documents were judged partially relevant by the majority of users (three or more). The extent of overlap on a per-user basis within the system's

partially relevant cluster reveals the scatter of relevance judgments surrounding this cluster. Table 3 depicts the distribution of judgments made by each user in this cluster.

#### Overlap: Not Relevant Cluster

The extent of overlap between the system's not relevant cluster and user derived not relevant clusters was also measured. The system determined that 69 documents out of the 98 documents returned from the original AlltheWeb.com search were not relevant. The data shows that the majority of users (three or more) also classified 61 out of these 69 documents as not relevant.

Since all five users made judgments for each URL in this cluster, the total number of relevance judgments to consider within this cluster is 345. Out of the 345 relevance judgments issued for these documents in the not relevant cluster, 300 user judgments agreed with the system and were not relevant, producing an 87% overlap. Not a single user classified these documents as relevant. Only 45 out of 345, or 13% of judgments varied from that of the system and were classified as partially relevant.

The system's not relevant cluster extensively matched the judgments of users, the majority of whom considered 61 of the 69 total documents to be not relevant. The agreement between the system and end-users for the not relevant cluster is displayed in Table 2. The mean number of end-users who matched the clustering of the system was 4.35 out of five users. The variance in the number of users who produced matching results is 0.567 and the standard deviation is 0.752. Thus, the mean, variance, and standard deviation of these results are nearly equal to that of the relevant cluster. Table 3 displays the distribution of judgments within the not relevant cluster for each user.

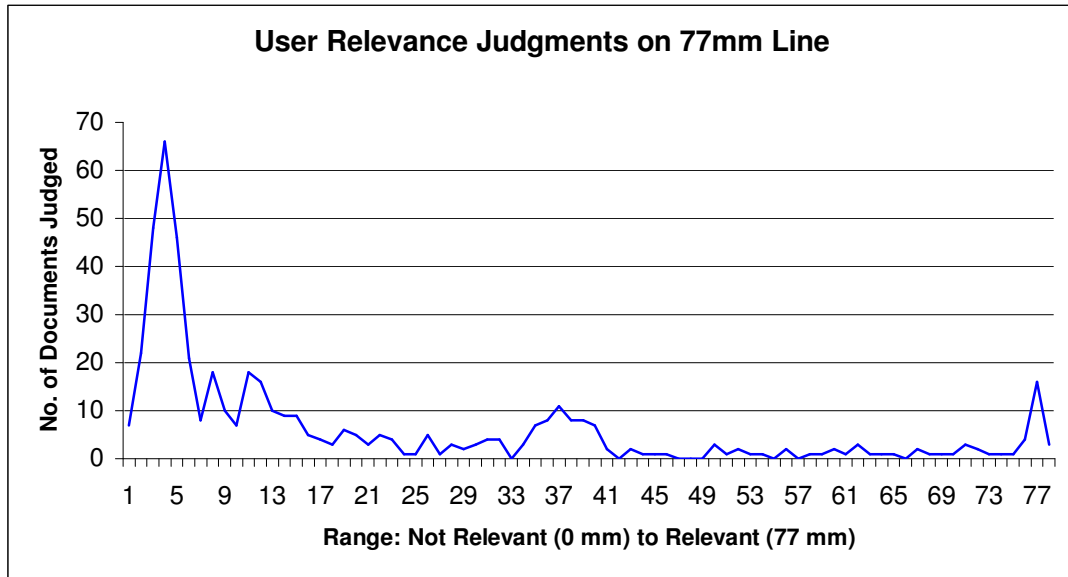
### Comparison Across Clusters

Overall, comparisons can be made regarding the system's ability to cluster across the relevant, partially relevant and not relevant clusters. The percentage match between the system and users for the relevant and not relevant regions is nearly identical at 87%. The partially relevant region, however, showed more variance, with an overlap of only 69%. This data reveals that our system was better able to pull out and cluster documents that were either relevant or non-relevant, as opposed to the partially relevant region. Such a difference may exist since deciding if a document is relevant or not relevant is intrinsically a simpler decision in which the document either fits the criteria or does not. However, partially relevant documents tend to exhibit features that characterize relevant along with non-relevant documents.

### Overlap on a Continuous Scale

The judgments made on the interval scale corresponding to each categorical judgment further exposed the range spanned by each cluster. Figure 10 illustrates the distribution of end-user relevance judgments in our study on a 77 mm interval (Greisdorf & Spink, 2001).

Figure 10. Range of relevance judgments on an interval scale.



From our data, the range of relevance judgments exhibited a pattern of a high peak at the tail end, a relatively flat middle region, and a sharp upswing near the head end, conforming to the distribution pattern exposed by Spink and Greisdorf (2001). This distribution shows the striking number of non-relevant documents that are returned to users from search engines, emphasizing the need for clustering documents returned from search engines on the Web.

#### Descriptive Characteristics Identifying each Region

On a descriptive scale, users identified the specific criteria they used to rank documents as relevant, partially relevant, or not relevant. This compiled set of descriptions is available in Table 4.

Table 4. End-user criteria for judging each region of relevance.

<b>CRITERIA FOR 'RELEVANT' ITEMS</b>	<b>CRITERIA FOR 'PARTIALLY RELEVANT' ITEMS</b>	<b>CRITERIA FOR 'NOT RELEVANT' ITEMS</b>
It was related	Some good information	Wrong context
A great source of information	Has a good link	Not related at all
On topic	Wrong topic but with good link	Wrong topic
Focuses on the topic	Gives partial information	Not useful
It was very useful	Gives good references	Totally unrelated
Exactly on topic	Good information on related topic	Does not mention the query at all
Clear document	Contains new angle on information	Mentions unrelated concept
Contains lots of words from topic	Partially related	Wrong interpretation
Good description	Somewhat useful	Way off
Entire document focuses on topic	Mentions new related concepts	Obviously wrong
	Partly on topic	Possibly good link
	Not 'relevant' & not 'non-relevant'	Duplicate
	Briefly discusses topic	Page not found
	Not totally about topic but good	No relation
	Contains small section on topic	Very few words on topic
	Lot of references	Focuses on something else
	Decent description	Information too hard to find

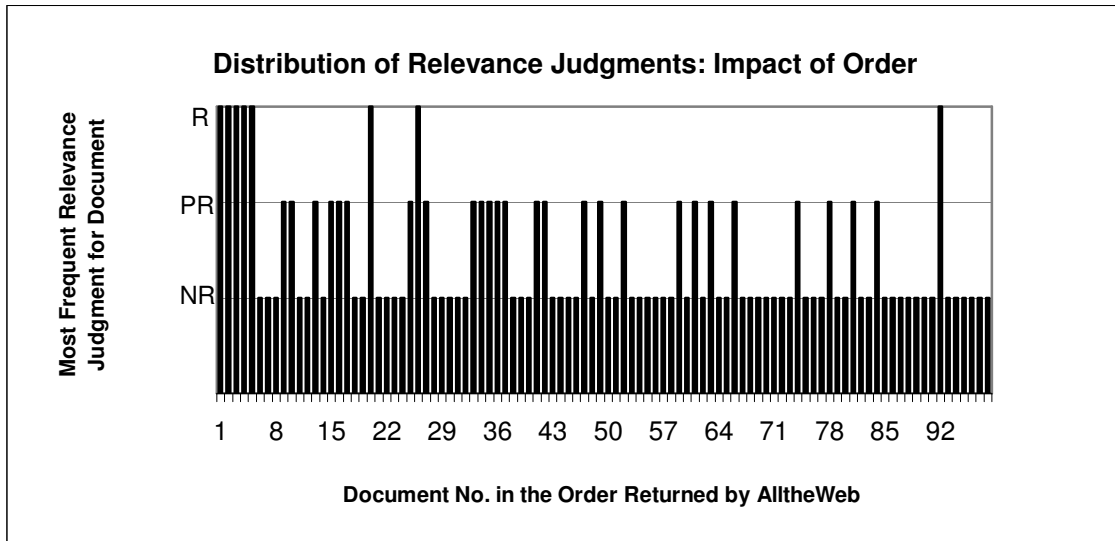
From the compiled descriptions, the criteria for establishing a document as relevant seem very evident and unambiguous. The relevant documents were those that focused on the topic and served as valuable sources of information. The non-relevant documents were often unrelated, out of context, and completely off topic. Partially relevant documents partly discussed the topic, provided satisfactory links to other documents, and often contained minor tidbits of information about the topic at hand.

### Order and Clustering

The motivation to cluster by relevance is evident through the impact of order within search results. Documents are often scattered throughout the result set in a one-dimensional ranked list. The most highly relevant results may not always be listed at the top of the ranked list and partially relevant results can be scattered throughout the set. There exist no clear indication about where the relevant results end and non-relevant results start. Figure 11 displays the most

frequent user relevance judgment for each of the 98 documents in the result set preserved in the order retrieved from AlltheWeb.com.

Figure 11. Relevance judgments for AlltheWeb.com results in the original order retrieved.



Indeed, the distribution of user judgments in our data set for documents returned from AlltheWeb.com reveals that documents of varying relevance can be scattered. For instance, the 92<sup>nd</sup> document was considered relevant while the 6<sup>th</sup> document was non-relevant. Although the general trend does reveal an ordering from high to low relevance, useful documents may be skipped just because of the order in which they appear.

### LIMITATIONS

This research is not without limitations, which are recognized here. The user aspects of the study were limited with a total of 5 end-users participants. This small sampling of users prevents achieving an optimal number of user judgments to produce the most accurate test data. Since this system is intended to run in the real-world environment of the Web, conducting the study with only five participants limits the extent of evaluation. Also, our user group consisted entirely of undergraduate and graduate Computer Science and Engineering students.

However, the user pool could be diversified to include a more assorted group of users with various backgrounds and experiences.

Additionally, our implementation ran one search query returning 98 results. This study could be expanded to include data from multiple search queries returning varied results to provide an even more extensive set of results for the evaluation purposes. Queries should be of a varying nature along with the relevance of the documents returned for each query. Indeed queries retrieving a different mixture of relevant, partially relevant, and not relevant documents should be tested and evaluated.

The query used in our study was intelligible, well-defined and static. However, in a real-world environment, queries may be less clearly defined. In a real-world scenario, the query would shift to account for interactive information seeking. This study did not account for this variation in the types and makeup of queries issued since only a single query was tested. Different queries might yield different results and this disparity should be accounted.

Another limitation of this study is that the documents to be evaluated by users were presented in the original order produced by the search engine. This may have an effect of swaying the user towards a grouping the document in a specific cluster based on the position of the document in the original ranked list. The URLs could be permuted when presented to users to avoid any subjectivity in the order of results presented. Moreover, the constants and criteria used in the algorithm for this study could be further tested to find the combination yielding optimal results.

## DISCUSSION

Despite these limitations, this research provides key evidence that documents can effectively be clustered based upon regions of relevance. This study considered the non-binary nature of relevance on the system level and tested it with a small pool of users. Based on the

data retrieved from our user group, we can conclude that our clustering scheme directs highly knowledgeable users that are certain about their search requirements directly towards the relevant cluster so that they can efficiently access the types of documents they seek. Likewise, the system provides quick and easy access to partially relevant results that novice users uncertain about their search goals may require.

In our test run, the system's clustering overlapped significantly with the clusters formulated by the users. Although the extent of overlap might change with future test runs accounting for a greater number of user participants, a larger variation in the types of users, and a wider range of queries issued, our test data shows a conspicuous overlap between the algorithm's decisions and user classification. While the system successfully clustered documents within all the regions, certain regions were more accurate than others based upon the user judgments provided.

The system's equally high success in clustering relevant and non-relevant documents provides some key implications. Both relevant and non-relevant documents were seen as highly distinguishable. Both of these regions have defining characteristics that accounts for the systems near identical performance in classifying documents within these clusters. In comparison to the relevant and not relevant clusters, the system's partially relevant cluster overlapped with users to a lesser extent. The boundary between them partially relevant and not relevant regions had noticeable overlap.

### Role of Order

Based on user judgments, it was found that the order of documents returned by search engines with one-dimensional lists does not always conform to the assumption that results are ranked from high to low relevance. For instance, it was shown that out of 98 possible positions in the ranked list, documents in the 20<sup>th</sup>, 26<sup>th</sup>, and 92<sup>nd</sup> position were judged as relevant while documents in the 6<sup>th</sup>, 7<sup>th</sup> and 8<sup>th</sup> positions were deemed as non-relevant. With search engines

often returning thousands of results, detecting all of the relevant, partially relevant, or not relevant documents in the pool is virtually impossible, unless each result is examined.

## CONCLUSION

This research showed that clustering documents on the Web by their regions of relevance is not only feasible, but also quite successful. Our clustering scheme offers an accessible, systematic, and versatile approach towards retrieving and organizing search results to enhance the way in which users of all domains meet their information seeking goals. Since partially relevant documents are useful for novice users at the beginning stages of their search, these documents are now clearly identified and grouped together. Likewise, expert users that have a clear idea of what they are seeking, can efficiently access the documents within the relevant cluster with our scheme.

Indeed, for a given information problem, individual users vary significantly in their levels of expertise, knowledge, certainty, and progression through the search process. Some users have a clearly defined notion of what they are looking for, while other users have only a loosely formed idea of the information they are seeking. Some users may have high knowledge about the search topic while others are learning about the topic for the first time. Certain users are at the initial stages of their search when they are still defining their search goals, while others are near the final stages of their search. Thus, a vast disparity exists among all classes of users, and this difference needs to be accounted for within IR systems.

## FUTURE RESEARCH

The research presented in this study can be extended in numerous directions.

- The algorithm can be embedded directly within a major Web search engine clustering scheme so that it can be fully operable on of the Web.

- Within each cluster, results can be ordered so that end-users could more selectively target potentially useful documents within each cluster.
- The number of clusters could be expanded to create an even more fine-grained clustering system.
- An optimal interface to present the clusters to users can be discovered.
- Variables within the algorithm can be tuned to find an optimal combination.
- The effects of order can be measured against the relevance of results.
- The system's performance can be further correlated with user behavior and search patterns.

## REFERENCES

Cutler, M., Shih Y., & Meng W. (1997). Using the structure of HTML documents to improve retrieval. *Proceedings of the USENIX Symposium on Internet Technologies and Systems* (pp. 241-251).

Greisdorf, H., & Spink, A. (2001). Median measure: An approach to IR systems evaluation. *Information Processing and Management*, 37, 843-857.

Hearst, M.A. & Pedersen, J.O. (1996). Reexamining the Cluster Hypothesis: Scatter/Gather on Retrieval Results. *Proceedings of the Nineteenth Annual International ACM SIGIR Conference*, 76-84.

Kekalainen, J., & Jarvelin, K. (2002). Using graded relevance assessments in IR evaluation. *Journal of the American Society for Information Science and Technology*, 53(13), 1120-1129.

Notess, G.R. (1999) Rising Relevance in Search Engines. *Online* 23(3), pp.84-86.

Rapela, J. (2001). Automatically combining ranking heuristics for HTML documents. *Proceedings of the Third International Workshop on Web Information and Data Management ACM Press* (pp. 61-67).

Rees, A.M., & Schultz, D.G. (1967). *A Field Experimental Approach to the Study of Relevance Assessments in Relation to Document Searching*. NSF Report.

Schamber, L. (1994) Relevance and information behavior. *Annual Review of Information Science and Technology*, 29, 3-48.

Sormunen, E. (2002). Liberal Relevance Criteria of TREC - Counting on Negligible Documents? *Proceedings of the Twenty-Fifth Annual ACM SIGIR Conference on Research and Development in Information Retrieval* 36, 324-330.

Sperber, D., & Wilson, D. (1986). *Relevance: Communication and Cognition*.

Spink, A., Greisdorf, H. (1997). Users' partial relevance judgments during online searching. *Online & CD-ROM Review* 21(5), 271-280.

Spink, A., & Greisdorf, H. (2001). Regions and levels: Measuring and mapping users' relevance judgments. *Journal of the American Society for Information Science and Technology* 52(2), 161-173.

Spink, A., Greisdorf, H., & Bateman, J. (1998). From highly relevant to not relevant: Examining different regions of relevance. *Information Processing and Management*, 34, 599-622.

Spink, A., Jansen, B.J., Wolfram, D., Saracevic, T. (2002) From E-Sex to E-Commerce: Web Search Changes. *IEEE Computer* 35(3), pp.107-109.

Sormunen, E., Kekalainen, J., Koivisto, J., & Jarvelin, K. (2001) Document text characteristics affect the ranking of the most relevant documents by expanded structure queries. *Journal of Documentation*, 57(3), 358-374.

Wallis, P., & Thom, J. A. (1996). Relevance judgments for assessing recall. *Information Processing and Management*, 32, 273-286.

Willett, P. (1998) Recent trends in hierarchical document clustering: A critical review. *Information Processing & Management*, 24(5), 577-597.