RAVE Reviews:
Acquiring relevance assessments from multiple users

Richard K. Belew
Cognitive Computer Science Research Group
Computer Science & Engr. Dept. (0114)
Univ. California - San Diego
La Jolla, CA 92093
http://www-cse.ucsd.edu/users/rik

John Hatton
Summer Institute of Linguistics
International Linguistics Center
7500 W. Camp Wisdom Road
Dallas, TX 75236
U.S.A.
john.hatton@sil.org

Abstract

As the use of machine learning techniques in IR increases, the need for a sound empirical methodology for collecting and assessing users' opinions -- "relevance feedback" -- becomes critical to the evaluation of system performance. In IR the typical assessment procedure relies upon the opinion of a single individual, an "expert" in the corpus' domain of discourse. Apart from the logistical difficulties of gathering multiple opinions, whether any one, "omniscent" individual is capable of providing reliable data about the appropriate set of documents to be retrieved remains a foundational issue within IR. This paper responds to such critiques with a new methodology for collecting relevance assessments that combines evidence from mulitple human judges. RAVe is a suite of software routines that allow an IR experimenter to effectively collect large numbers of relevance assessments for an arbitrary document corpus. This paper sketches our assumptions about the cognitive activity of the providing relvance assessments, and the design issues involved in: identifying the documents to be evaluated; allocating subjects' time to provide the most informative assessments; and aggregating multiple users' opinions into a binary predicate of "relevant." Finally, we present preliminary data gathered by RAVe from subjects.

Introduction

The evaluation of information retrieval (IR) system performance has long been recognized as a notoriously difficult feature of research in the area. Chief among the causes is the field's difficulty in adequately defining "relevance," the construct by which retrieved documents are judged as successful responses to a users query. As the use of machine learning techniques in IR increases, particularly those depending on "relevance feedback," the need for a sound empirical basis for collecting and assessing users' opinions becomes even more acute.

Until quite recently, the conventional IR research methodology depended heavily on a small set of corpora for which "relevance assessments" were available. While the procedures by which these assessments were obtained have often been unclear, typical assessment procedures relied upon the opinion of a single individual: an "expert" in the corpus' domain of discourse is identified, presented with a series of query/document pairs, and then asked to determine whether the document was or was not relevant to that query.

Spurred in part by the need for larger test collections, recent years have seen the development of new methodologies for relevance assessment [Harman, 1993]. Because the collections are so large, however, getting even a single relevance assessment has been an expensive and time-consuming activity. Once again, therefore, the tacit assumption has necessarily been that a single expert can be trusted to provide reliable relevance assessments. Apart from the logistical difficulties, whether any one, "omniscent" individual is capable of providing reliable data about the appropriate set of documents to be retrieved

remains a foundational issue within IR. For example, a number of papers in a recent special issue of JASIS devoted to relevance advocated a move towards a more "user-centered," "situational," view of relevance [Froehlich, 1994]. This paper responds to such critiques with a new methodology for collecting relevance assessments that combines evidence from *mulitple* human judges. A new suite of software tools are also presented that facilitate relevance assessment.

The defining characteristic of this methodology is that rather than having relevance be a Boolean determination made by a single, omniscent expert, we will consider it to be a *consentual central tendancy of the searching users' opinions.* The relevance assessments of individual users and the resulting central tendancy of relevance is suggested by the following diagram.

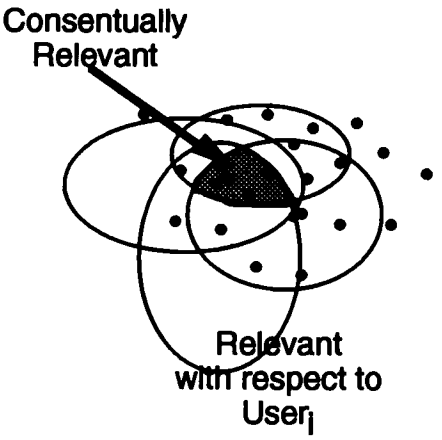

Figure 1: Consentual Relevance

Two features of this definition are significant. First, consentual relevance posits a "consumers" perspective on what will count as IR system success. A document's relevance to a query is not going to be determined by an expert in the topical area, but by the users who are doing the searching. If they find it relevant, it's relevant, whether or not some domain expert thinks the document "should" have been retrieved.

Second, consentual relevance becomes a statistical, aggregate property of multiple users' reactions rather than a discrete feature elicited from an individual. By making relevance a statistical measure, our confidence in the relevance of a document (with respect to a query) increases as more relevance assessment data is collected. This is consistent with the strong link between IR and machine learning research now developing [Lewis, 1994 ; Bartell, 1994] It also anticipates the use of adaptive techniques which transform browsing users' behaviors into changes in the documents' indexed representations [Belew, 1989]. The data presented below is insufficient to allow statistically significant statements, but further data collection is underway.

It seems, however, that our move from omniscent to consentual relevance has only made the problem of evaluation that much more difficult: Test corpora must be large enough to provide robust tests for retrieval methods, and multiple queries are necessary in order to evaluate the overall performance of an IR system. Getting even a single person's opinion about the relevance of a document to a particular query is hard, and we are now interested in getting many!

The rest of this paper reports on RAVe, a Relevance Assessment VEhicle that demonstrates it is possible to operationally define relevance in the manner we suggest. RAVe is a suite of software routines that allow an IR experimenter to effectively collect large numbers of relevance assessments for an arbitrary document corpus. It has been developed as part of an extended investigation into spreading-activation search through associative representations, as well as the use of relevance feedback techniques to adapt such representations over time. The software has been put in the public domain for use by other IR researchers, and is available for FTP file transfer.

This paper sketches our assumptions about the cognitive activity of the providing relvance assessments, and the design issues involved in: identifying the documents to be evaluated; allocating subjects' time to provide the most informative assessments; and aggregating multiple users' opinions into a binary predicate of "relevant." Finally, we present preliminary data gathered by RAVe from more than 40 novice and expert subjects.

Assumed cognitive model

We have designed RAVe and the experimental task making several key assumptions about the cognitive activity of making relevance assessments. An important area for further research is a more detailed cognitive analysis of relevance generally (Sperber and Wilson's book suggest a number of promising leads [Sperber, 1986]) and its assessment in an IR context in particular. Each of these assumptions are a matter of considerable debate [Froehlich, 1994], and should be substantiated in future work. This paper does not attempt to provide evidence for or against any particular cognitive model; here we simply explicate the theoretical basis from which we proceed.

First, we believe the task can best be described as one of *object recognition.* , in the tradition of Rosch and others [Rosch, 1977]. The object to be recognized is an internally represented *prototypic* document satisfying the user's "information need." Then, as a user considers an actual, retrieved document's relevance, he or she evaluates how

well it *matches* the prototype, the model the subject maintains of an ideally relevant document. Barry and others have suggested the many and varied features over which the prototypes can defined can be [Barry, 1994]. Only a small number of these may be revealed by any one of the user's queries, of course.

Since the hypothesized prototypes are interally-represented and may be difficult or impossible to ever inspect directly, the queries become our most important source of scientific evidence. In the experiments reported below we collect data on two distinct classes of queries that both appear to play important roles in real IR system use. The first class might be considered "typical" queries: short lists of query terms that might naturally be expressed by a user. We will treat these as "simple" queries, i.e., ignoring noise words and any Boolean, proximity or other operators sometimes used to structure queries in advanced query languages. Users are instructed to find documents that are "about" these queries. The second class of queries are much larger samples of free text, typically generated by relevance feedback. Users are instructed to find "more documents like" these queries. Since more and more IR systems support this form of querying, and since the frequency distribution of keywords in the query has a significant impact on many weighting schemes, it is important to collect data regarding both "long" (relevance feedback) and "short" (typical) queries.

Next, we assume that the cognitive load (e.g., short- and long-term memory demand) required to read a query specification and build a prototype corresponding to it is comparable to that required to read a document and assess its relevance. We also assume that the load required to do both of these is large, relative to that required to maintain the query's representation in memory and reliably perform a recognition task. Based on these assumptions, our experimental design asks subjects to maintain *several* queries in mind, *simultaneously*. That is, we begin by training subjects to recognize three different queries (two short queries and one paragraph-long relevance feedback query). Then, users are given a document to read and asked to assess its relevance with respect to each of the three queries. Users appear to find it quite easy to keep three distinct queries in mind, and assess a document according to each. While this version of the relevance assessment task may be unrealistic (except as a model for an over-worked reference librarian!), it does allow an important efficiency to our experimental methodology that significantly increases the number of relevance assessments acquired.

Finally, we assume the user is capable of grading the quality of this match. assessment. Our relevance assessment asks subjects to score the quality of relevance match according to a five-point scale shown in Figure 2.

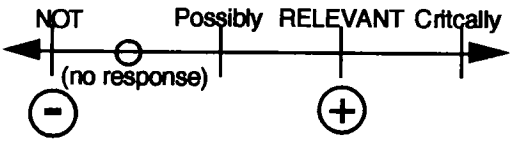

Figure 2: Relevance scale

We view this as an ordered, non-metric scale of increasing relevance-match. In the experiments reported here, we do not support the "not relevant" response; hence users' assesments are restricted to the positive end of the scale. Three grades of positive responses are possible, so that the user can qualify the middle "relevant" response by either weakening it ("possibly relevant") or strengthening it ("critically relevant"). Also note that only these positive assesments require overt action on the part of subjects; "no response" is the default assessment unless the subject makes one of the other, active responses.

RAVeUnion, RAVePlan, Interactive Rave and RAVECompile

It would be most useful if, for every query, the relevance of every document could be assessed. However, the collection of this many s, for a corpus large enough to provide a real retrieval test, quickly becomes much too expensive. If the evaluation goal is relaxed to being the *relative* comparison of one IR system with one or more alternative systems, assesments can be constrained to only those documents retrieved by one of the systems.

We therefore follow the "pooling" procedure used by many other evaluators [Harman, 1993], viz., using the proposed retrieval methods themselves as procedures for identifying documents worth assessing. An under-appreciated consequence of this methodological convenience is that *unassessed documents are assumed to be irrelevant.* This creates an unfortunate dependence on the retrieval methods used to nominate documents, which we can expect to be most pronounced when the methods are similar to one another. For example, if the alternative retrieved sets are the result of manipulating single parameters of the same basic retrieval procedure , the resulting assesments may have overlap with, and hence be useless for comparison of, methods producing significantly different retrieval sets. For the TREC collection, this problem was handled by drawing the top 200 documents from a wide range of 25 methods which had little overlap ..

Similarly, first step in constructing a RAVe experiment is to combine the ranked retrieval lists of the retrieval

methods. to be compared[1] creating a single list of documents ordered according to how interested we are in having them assessed by a subject. We call this function RAVeUnion, and it can be complex in several respects. First, the assessment of a document whose ranked order is highly correlated across retrieval methods provides little information about differences between the methods. Said another way, we can potentially learn most from those documents whose rank order is most different, and hence a measure of the *difference* in ranked orders of a particular document might be used to favor "controversial" documents. This factor has the unfortunate consequence, however, of being sensitive to what we would expect to be the least germane documents, those documents ranked low by any of the methods under consideration. A second factor that could be considered is a "sanity check," including near the top of our list a *random* sample. While we might learn a great deal from these if users agree that these randomly selected documents are in fact relevant, we expect that in general the retrieval performance of the systems should not depend on random documents .

Consequently the current implementation of RAVeUnion produces the most straight-forward "zipper" merge of the lists, beginning with the most highly ranked and alternating. The output of RAVeUnion is a file of (query, document) pairs along with a field which indicates if the pair was uniquely suggested by only one of the methods.

A second challenge in preparing a RAVe experiment is achieving the desired *density* or redundancy of sample points. That is, for each document that we believe may be relevant to a query, how many subjects should evaluate it? The answer will vary depending on such factors as the number of participants, their expertise, their motivation to produce quality judgments, how long each will spend rating documents, etc. A higher density means that less documents will be evaluated, but also that the cross-subject, cumulative assesment is likely to be more statistically stable. This can be especially important with an adaptive retrieval system in which relevance feedback is to be used to change the system over time.

The trade-off between the most important of these factors is captured in the following formula:

$$x = \frac{NR}{STQ}$$

where:

x	= number of documents to be evaluated for each query
N	= number of subjects
R	= expected subject efficiency (votes/user/time)
T	= time spent by subjects
S	= desired density (votes/document)
Q	= number of queries to be evaluated

Note that this formula ignores the overlap between queries that occurs when the user sees a document that may be relevant to two or more of the queries in the user's list. Care must be taken, therefore, to minize expected overlap between the topical areas of the queries. We have also found that the assessment densitities constructed using this formula to be unfortunately uneven. The main source of these is variability in R, the rate at which subjects are able to produce relevance assesments. This rate can only be estimated, at least until some pre-test experience with the population is available. Data below will show there to be high variability (exceeding 500%) demonstrated by the subjects in our experiments.

RAVePLAN takes as input a list of Q query specifications, a list of N subject logins, the desired density S, and the number of documents $R*T$ that should be allocated to each subject. The query specifications indicate which queries can go in which fields, and which queries should not be shown together. This allows us to limit possible interactions between queries about similar topics.

Having made these decisions, we are now ready to present queries and documents to users for evaluation. The interactive facility to do this (written in TCL/TK) is shown in Figure 3. The top of RAVe's window displays the three queries against which the subject is to judge each document. Two queries are short sentences or phrases, like "APPLICATIONS OF AI TO EDUCATION", and the third is a scrolling pane containing the text of the long, relevance-feedback document. While the subject must judge the documents shown to him or her for being "about" the two short queries, the task associated with the query-document is to find "documents like this."Below each query the RAVe window contains four radio-buttons labeled "Not (relevant)", "Possibly (relevant)", "Relevant", and "Critically (relevant)". Since we asked our subjects to spend two hours each, but could not assume their participation would necessarily be continuous, there is a "QUIT" button which allows the subject to suspend the session; when the subject launches RAVe again, the session will be resumed where he or she left off. The

[1] Our methodology will be descibed in terms of two particular IR systems (see below) but can be easily generalized to comparison among more alternatives.

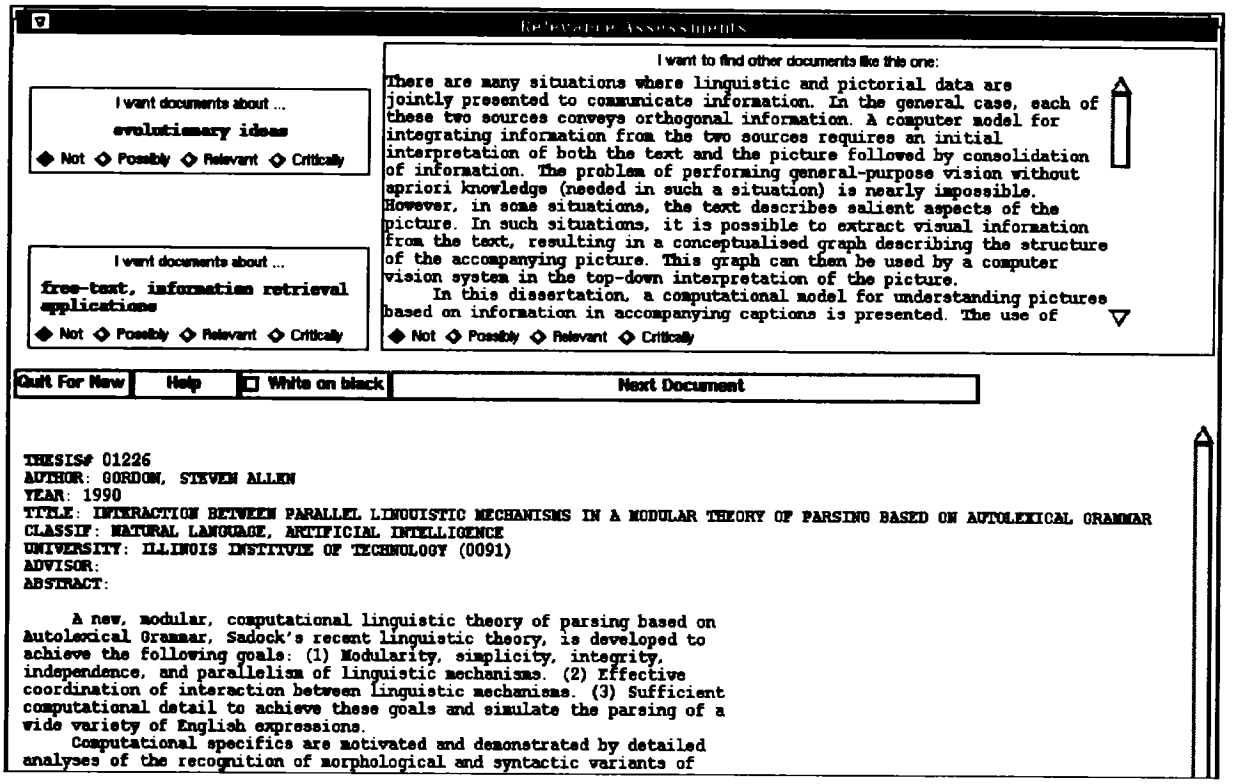

Figure 3: RAVeInteractive Window

"white on black" toggle button allows subjects to choose the more comfortable choice between white text on black background or vice versa. Finally, the "Next" button is pressed after the subject has read and recorded his or her relevance assessments for a document.

Once all the data has been collected in this fashion, the final step is to transform the data about the *distribution* of users' assessments into more reduced statistics. Mode, mean, variance of this distribution are all of interest, but here we restrict ourselves to the Boolean relevant/non-relevant discriminations typically used in IR evaluation are an extreme reduction. RAVeCompile accomplishes this by collating all users assessing the same query/document pairs and then mapping the set of four-valued relevance assessments into a binary value. RAVeCompile lets the experimenter configure a simple predicate of the following form: .

$$s_{q,d} = a * PossVote + b * RelVote + c * CritVote$$

$$Rel?_{q,d} = \begin{cases} (VoteCount_{q,d} \geq Quorum) \wedge \\ ((VoteCount_{q,d} * s_{q,d} \geq MinSum) \vee \\ (s_{q,d} \geq MinAvg)) \end{cases}$$

where

a = weight assigned to votes of 'possibly-relevant'

b = weight assigned to votes of 'relevant'

c = weight assigned to votes of 'critically-relevant'

$s_{q,d}$ = weighted aggregate score across relevance levels

$VoteCount_{q,d}$ = total number of active votes collected for (q,d) pair

$Quorum$ = minumum number of votes required for (q,d) to be considered "relevant"

$MinSum$ = threshold for cumulative assessment scores

$MinAvg$ = relevance criterion

In the data presented below, we give examples of two predicates constructed in this fashion. These are:

Permissive: if (two or more POSSIBLE votes) or (at least one RELEVANT vote)

Stringent: if (two or more RELEVANT votes) or (at least one CRITICAL vote)

The AIT Experiment

To give a concrete example of how RAVe can be used to evaluate the relative performance of two (or more) IR systems, this section will describe its use as part of an evaluation of ToAir, the most recent implementation of class of IR systems based on associative representations and using spreading-activation search [Belew, 1986 ; Belew, 1989 ; Rose, 1994]. SMART [Salton, 1971 ; Buckley, 1985] is used to provide comparison against a strong, well-known and understood standard. The corpus used is the Artificial Intelligence Thesis (AIT), a set of approximately 2300 thesis abstracts in the area of artificial intelligence. The average document is two-three paragraphs long (about 2000 characters on average), and the entire corpus is approximately two megabytes. These thesis abstracts have many interesting characteristics, some of which have been explored in other work by our group [Steier, 1994], but will be viewed as simple textual samples in our experiments here. We have also developed a set of eleven varied and representative queries; these are listed in Appendix 1. The subjects used for this experiment were 25 "experts" in AI: faculty, post-docs and advanced graduate students working in the area of AI at UCSD. Each subject was assigned three queries (two short and one long, relevance-feedback) and then spent two hours using the RaveInteractive system to read and evaluate AIT documents.

Since this was our first experience using RAVE with real subjects, our first question was just how quickly they would be able to collect relevance assessment. We found a wide variation in the length of time each subject spent reading each document, as shown in Figure 4. In future experiments, we intend to provide more clear instructions about how evaluate each document. In debriefing, subjects asked questions such as: "Is it OK to go on to the next document if I can tell at a glance that this document is not relevant?". Further, we may want to distinguish between topicality and pertinence (situational relevance), among other categories of relevance [Park, 1994].

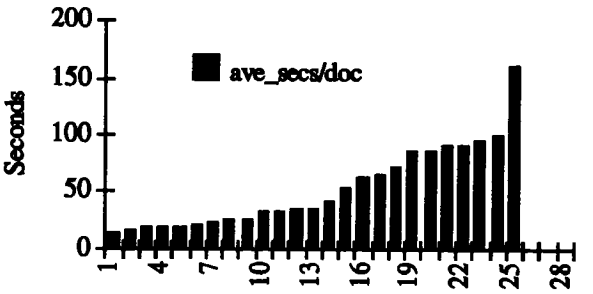

Fig. 4: Avg Assessment Time by Subject

A second question concerns inter-subject variability in the baselines against which each subject assesses relevance. One reflection of such variability is the average relevance assessment assigned by each subject, where the average is formed across all documents viewed by that subject. In Figure 5, a score of 1.0 corresponds to "Somewhat relevant," and we can see first that in general subjects found the (zippered-merge of SMART and ToAIR) retrievals set to in general contain few relevant documents: the average vote for a (query, doc) candidate was half-way between "non-relevant and somewhat-relevant." A second observation is that, with the exception of a few especially tolerant subjects, there was not a great deal of variance across the subject pool in this average.

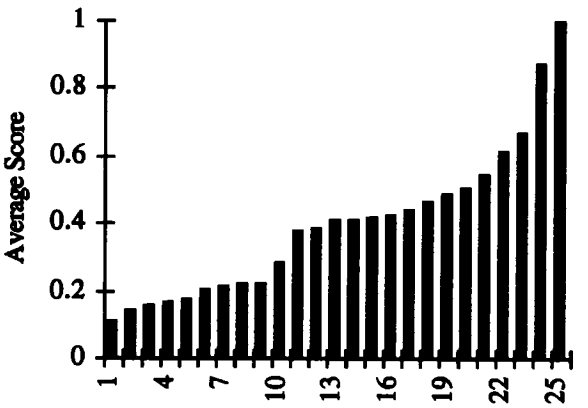

Fig. 5: Avg. Relevance Score by Subject

Figure 6 demonstrates two other important dimensions of the relevance assessment task. First, the results of using the alternative "Permissive" and "Stringent" predicates for resolving the users' multi-valued assessments into the conventional binary, non/relevant distinction is shown for each query. Perhaps not surprisingly, the two measures appear to be highly correlated, with the more stringent criterion reducing the number of "relevant" documents by approximately one half. The second interesting dimension

is the difference between users' assessments of standard, short queries as compared to the longer, relevance feedback queries (9,10,11). For relevance-feedback queries, it appears that users consider, if anything, a larger number of documents at least somewhat relevant, but a significantly smaller number to be relevant according to the more stringent criterion.

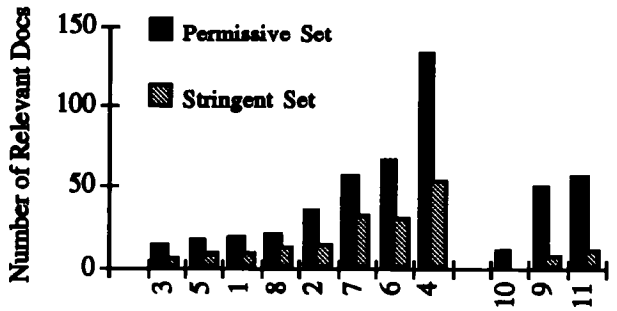

Figure 6: Number of Relevant Docs by Query & Predicate

Conclusion

We have presented an argument that the conventional characterization of "relevance" in IR has been skewed towards an unrealistic assumption that a single, "omniscent" domain expert can be charged with the task of finding documents that should appropriately be retrieved. In part this fallacy has been perpetuated by the methodological difficulty of collecting relevance assessments. RAVe has been presented as a system for efficiently collecting this important data, not from a single purported expert but from large numbers of IR system users who must ultimately be the consumers of this service. The resulting notion of relevance therefore becomes a consentual, statistical one.

This conception of "relevance" brings with it a number of new issues, including:

- an appropriate cognitive model of the relevance assessment process;
- an appropriate sampling of retrieval results from the mulitple IR systems to be evaluated;
- the appropriate density of relevance assessment across the corpus;
- the appropriate criterion to be used to translate rich user reactions into binary non/relevant classifications most typical to IR system evaluation.

This paper has only begun to scratch the surface of these issues, and the RAVE software implements what must be considered preliminary solutions to some of them. We have designed the RAVe programs to allow independent investigation of each of these lines of enquiry, and encourage other investigators to contact us to use these tools. The core of our own investigation is the statistical distribution of users' assessments, in particular whether they have the central tendancy hypothesized above.

Acknolwedgements

This research was supported in part by NSF Grant #IRI-9221276 and Apple Computer. Useful comments on early drafts of this paper by Gary Cottrell and Brian Bartell are gratefully acknowledged.

References

Barry, C. L. (1994). User-defined relevance criteria: An exploratory study. JASIS, 45(3), 149-159.

Bartell, B. T., Cottrell, G. W., & Belew, R. K. (1994). Automatic combination of multiple ranked retrieval systems. In W. B. Croft & C. J. v. Rijsbergen (Eds.), Proc. SIGIR 1994, (pp. 173-181). Dublin: Springer-Verlag.

Belew, R. K. (1986). Adaptive information retrieval: machine learning in associative networks. Ph.D. Thesis, Univ. Michigan, Ann Arbor.

Belew, R. K. (1989). Adaptive information retrieval: Using a connectionist representation to retrieve and learn about documents. In Proc. SIGIR 1989, (pp. 11-20). Cambridge, MA:

Buckley, C. (1985). Implemetation of the {SMART} Information Retrieval Retrieval System Cornell Computer Science Dept. Tech. Report No. 85-686).

Froehlich, T. J. (1994). Relevance Reconsidered -- Towards an Agenda for 21st Centry: Introduction. JASIS, 45(3), 124-134.

Harman, D. (1993). Overview of the first TREC conference. In R. Korfhage, E. Rasmussen, & P. Willett (Eds.), Proc. SIGIR 1993, (pp. 36-47). Pittsburg: ACM Press.

Lewis, D. D., & Gale, W. A. (1994). A sequential algorithm for training text classifiers. In W. B. Croft & C. J. v. Rijsbergen (Eds.), Proc. SIGIR 1994, (pp. 3-12). Dublin: Springer-Verlag.

Park, T. K. (1994). Toward a theory of user-based relevance: A call for a new paradigm of inquiry. JASIS, 45(3), 135-141.

Rosch, E. (1977). Classification of real-world objects: origins and representations in cognition. in Thinking:

Readings in cognitive science P.N. Johnson-Laird, P. C. Wason, ed. Cambridge Univ. Press, Cambridge.

Rose, D. E. (1994). A Symbolic and Connectionist Approach to Legal Information Retrieval. Hillsdale, NJ: L. Erlbaum.

Salton, G. (1971). The {SMART} retrieval system - experiments in automatic document processing. Englewood Cliffs, NJ: Prentiss-Hall.

Sperber, D., & Wilson, D. (1986). Relevance : Communication and cognition. Cambridge, MA: Harvard Univ. Press.

Steier, A. M., & Belew, R. K. (1994). Talking about AI: Socially-defined linguistic sub-contexts in AI. In AAAI94, . Seattle, WA.

Appendix 1: AIT queries

Query#1 legal applications

Query#2 free-text, information retrieval applications

Query#3 evolutionary ideas

Query#4 connectionism or neural nets

Query#5 neural network research that is biologically plausible

Query#6 reasoning about uncertainty

Query#7 educational applications

Query#8 genetic algorithms

Query#9 This thesis is a study of the computational complexity of machine learning from examples in the distribution-free model introduced by L. G. Valiant (V84). In the distribution-free model, a learning algorithm receives positive and negative examples of an unknown target set (or concept) that is chosen ....

Query#10 In this research we will develop a framework for applying some abstract heuristic search (AHS) methods to a well known NP-Complete problem: the graph partitioning problem (GPP). The uniform graph partitioning problem ....

Query#11 There are many situations where linguistic and pictorial data are jointly presented to communicate information. In the general case, each of these two sources conveys orthogonal information. A computer model for integrating information from the two sources requires an initial interpretation of both the text and the picture followed by consolidation of information. The problem of performing general-purpose vision...