# Cover Page

# "Assessing the Agreement of Cognitive Space with Information Space"

*A Research Seed Grant Proposal to the*
*UNC-CH Cognitive Science Program*

Submitted by:

Dr. Gregory B. Newby
Assistant Professor
School of Information and Library Science
CB 3360 Manning Hall
Chapel Hill, NC, 27599-3360
email: gbnewby@ils.unc.edu
Telephone: 919-962-8064
Vita: http://ils.unc.edu/gbnewby/vita.html

Submitted to:

Mr. Chuck Wise
Department of Psychology
Davie Hall
CB#3270

**ABSTRACT**

The proposed research will use a psychometric survey approach to map the cognitive spaces of subjects. The domain under study is information retrieval, in which an information space is automatically generated from a collection of text documents, and documents relevant to a particular topic are sought. This basic research will address the extent to which agreement between human cognitive space and a system's information space will produce a greater number of relevant documents, and take steps towards better human-machine synergy.

**INTRODUCTION**

Cognitive space is measurable by various means. The proposed research will make use of multidimensional scaling (MDS) to generate maps of the cognitive spaces of research subjects. The maps will cover a limited domain of concepts related to an ongoing investigation of information retrieval (IR).

Cognitive space is used here to refer to the knowledge, opinions, attitudes and beliefs held by humans. In MDS, cognitive space mapping is limited to concepts (typically, words or phrases) and the degree of dissimilarity among concepts (see Kruskal & Wish, 1984). This is in contrast to maps of cognitive space based on hierarchical, semantic, time-ordered or other more complex types of relationships. The conceptual benefit of MDS is that "similarity" or "dissimilarity" may be considered as fundamental units of measure. In other words, "dissimilarity" may not be broken down into smaller parts or units, as can, for example, measures on most Likert scale surveys or Myers-Briggs personality inventory assessments.

For this study, an additional important benefit of MDS is that it offers measurement of cognitive space in a format comparable to that generated by automatic representation of textual documents. The approach taken by the investigator (Newby, 1998; Newby, 1997) is similar to other research approaches to information retrieval (cf. Salton & McGill, 1983; Rehder et al., 1998). The investigator's approach is based on multidimensional "information space," in which documents and concepts (typically, terms contained within documents) may be located in relation to one another.

The challenge of building information space is to determine appropriate measures of dissimilarity to apply to concept-concept, concept-document, and document-document pairs. Ideally, measures should be derived automatically from the full text of the documents themselves, with little or no human intervention. This enables the representation of millions of documents and thousands of terms in a single information space. The investigator has used term co-occurrence across documents as the basic measure. By generating a term by term co-occurrence matrix for a collection, relations among terms may be identified. Eigenvectors from the co-occurrence matrix are used as coordinates in a multidimensional geometric space, in which each term has a measured distance (dissimilarity) from each other term. Documents are located at the center of the terms they contain.

Within the information space, information retrieval proceeds by locating a query, topic or other statement of information need at the center of its associated terms. Document representations located most closely to the query are then retrieved and presented to an information seeker. Ideally, only documents considered relevant to the query will be retrieved.

The driving force of this proposal is this assumption:

> *An information space which is in agreement with the cognitive space of an information seeker (or group of information seekers) will be more successful at information retrieval than one which is less in agreement.*

This is, in fact, the driving force behind most types of information retrieval systems, such as that used by library catalogers who choose appropriate indexing terms from a controlled vocabulary to assign to documents: by anticipating the cognitive spaces of generic user groups, they hope to generate an information space which will help achieve a high degree of relevance.

The extent to which cognitive space may be accurately measured, and to which these measurements have bearing on the reality of social interaction, language, action, etc. is a topic of debate in the social sciences. For the proposed research, however, the goal is practical: by using an established technique for the measurement of cognitive space, a basis for comparing different approaches to the generation of information space will be generated. This will enable further research on the relationship between information space and cognitive space, and help move towards a cognitively-based understanding of information seeking behavior (see Ingwersen, 1996; Newby 1998).

**METHODOLOGY**

This proposal is to administer paired-comparison surveys to a group of at least 20 subjects. Because of the length of the surveys and need for training in generating paired-comparison judgements, the subjects will be paid. The paired-comparison judgements will be used to form a map of the shared and individual cognitive spaces of the subjects using MDS and related techniques, including Principal Components Analysis (PCA) and eigensystems analysis (also known as spectral analysis).

Paired-comparison surveys take the form of "Given that A and B are X units apart, how far apart are C and D?" For example:

> Given that tasty sauce and good cheese are 100 units apart,
>
> How far apart are spices and fresh vegetables?

If, for example, 10 concepts from the pizza domain were under study, **`[ 10 * (10 - 1) ] / 2`** or 45 separate pairs would need to be examined to build a complete pairwise matrix of scores. The goal in choosing the "ruler" (in this case, tasty sauce and good cheese) is to identify a pair with high agreement among subjects. The effectiveness of the ruler pair, and the extent to which the concepts of the domain under study are understood by subjects, is assessed during pre-testing.

For this research, the domain under study is a series of eight topic statements from the Text REtrieval Conference (TREC, 1991-1997). TREC has generated an extensive test collection of topics and relevance judgements, with over 350 topics and nearly one million documents from different document collections. Document types include patent abstracts, newspaper articles, government information (Federal Register and Congressional Record), magazine articles, and translations of foreign radio broadcasts (Foreign Broadcast Information Service).

TREC participants generate ranked lists of documents they believe are relevant for each topic. Impartial judges then decide whether each submitted document is relevant or non-relevant for that topic. Although the overall goal is to achieve perfect precision (that is, no non-relevant documents) it is also desirable to have high recall (when a high proportion of the known relevant documents are presented, even if some non-relevant documents are also included).

The eight topics were chosen based on their having many relevance judgements, and being clearly distinguished based on two variables identified as important for retrieval effectiveness in prior TREC research: document length (number of words) and topic difficulty (ratio of relevant to non-relevant documents submitted for judgement).

**TREC Topic numbers under investigation**

| | Hard | Easy |
|---|---|---|
| Long | #116, #172 | #129, #174 |
| Short | #241, #245 | #221, #223 |

All topics are available at http://beryl.ils.unc.edu/is/topic. Across the 8 topics, there are about 300 unique words, some of which are common "stopwords" such as *the*, *and*, and *but*. Paired-comparison surveys will be

constructed for each query, and for overlap among the queries. Pre-testing will determine which terms and phrases are perceived as most important for each topic, instead of choosing all terms. (A full 300-term paired-comparison survey would require almost 45,000 judgments, which is hardly practical.)

A cognitive space map will be built for each of the 8 queries, plus for all queries together ("how far apart are topic 245 and topic 172?").

**ANALYSIS**

The most important outcome of the method described above is something that has never before been created: a series of cognitive space maps directly related to information need topic descriptions. These maps will be used as ideal goals for automatic document processing. Steps to build information spaces which better approximate these cognitive spaces will be sought, and the effectiveness of the MDS-derived cognitive spaces for retrieval will be measured.

Because the cognitive spaces and information spaces are involve many continuous variables, and the topic descriptions and relevance judgements (and qualities of documents judged) are similarly multi-faceted, there is no single definitive approach to the analysis. The fundamental approach will be multivariate statistical analysis, such as discriminant analyses to determine which space (or which dimension in which space) is a better predictor of relevance/non-relevance. Additional approaches will include examining the similarities between the information space and cognitive space (using matrix algebra techniques, such as the difference between two matrices), and looking at relevant sub-spaces within each space which are similar to sub-spaces from other spaces. Finally, visualization of the cognitive and information spaces will be undertaken to determine the extent to which these spaces are navigable or understandable through 3-dimensional navigation (see Newby, 1998).

**CONTINUING RESEARCH**

The investigator is currently involved with two separate proposals to the National Science Foundation due this summer, and has at least two other proposals he will work on this fall. The proposed research, if funded, will provide an excellent basis for the cognitive emphasis of the research. Because of the need to pay subjects and data collectors, it is unlikely the research can be performed without seed money.

The proposed study is basic, new, and investigative: it incorporates a well-known survey technique from psychometrics with well-known approaches to document representation and retrieval from information science. The bond between these areas is a theoretical conjecture about how information spaces might perform if they are derived from or based on cognitive spaces.

In 1975, a well-known information scientist proposed the concept of "exosomatic memory" (Brookes, 1975). His idea was that the ultimate goal for information retrieval systems was to have such a close modeling of the thoughts, goals and knowledge of its users that it could act, in essence, as an extension to human memory. This fanciful notion of how people might interact with information systems has not been forthcoming despite years of work at artificial intelligence, machine learning, information retrieval and other areas. If such a system were to be created, however, it seems obvious that the information space of the system would need to be highly oriented towards the cognitive space of its users or user groups. This research is proposed as a necessary step towards generating an understanding of how such a system would operate, and how a synergy between information systems and human cognitive could be developed.

**BUDGET**

Note: In order to expend funds by the end of the fiscal year as specified in the grant guidelines, all work will be completed during May and June 1998. Analysis by the investigator will continue after data are collected.

| | | |
|---|---|---|
| 1. | Quantity (2) graduate student employees for 20 hours / week for 6 weeks each.<br>- will help pre-test surveys, recruit subjects, train & supervise subjects, enter data, and assist in analysis. Some minor software development for analysis. (Includes insurance.) | $2900. |
| 2. | Quantity (45) subject payments, including pre-tests. Average 2 hours per session at $7/hour. | $ 630. |

SILS departmental resources will be used for data analysis, photocopying and other office supplies.

**REFERENCES**


Brookes, BC. (1975). "The fundamental problem of information science." in V. Horsnell, ed. Informatics 2. London: Aslib.

Ingwersen, P. 1996. Cognitive Perspectives of Information Retrieval Interaction: Elements of a Cognitive IR Theory. Journal of Documentation 52 (1): 3-50.

Kruskal, JB; Wish, M. 1984. Multidimensional Scaling. Beverly Hills: Sage.

Newby, GB. 1997. Metric multidimensional information space. Fifth TREC Proceedings. Gaithersburg, Maryland: NIST.

Newby, GB. 1998. Context-based statistical sub-spaces. Sixth TREC Proceedings. Gaithersburg, Maryland: NIST.

Newby, GB. 1998. The strong cognitive stance as a conceptual basis for the role of information in informatics and information system design. To appear in Proceedings of SCI/ISIS '98, July 12-16. Available: http://ils.unc.edu/gbnewby/papers/cogspace-98.ps

B. Rehder; T.K. Landauer; M.L. Littman; & S. Dumais. 1998. Automatic 3-Language Cross-Language Information Retrieval with Latent Semantic Indexing. Sixth TREC Proceedings. Gaithersburg, Maryland: NIST.

Salton, G; McGill, M. 1983. Introduction to Modern Information Retrieval. New York: McGraw Hill.

TREC (The Text REtrieval Conference). 1991-1997. Online proceedings at http://trec.nist.gov; print proceedings from National Institute of Standards and Technology, Gaithersburg, Maryland.