PARTIAL COORDINATION: A PRELIMINARY EVALUATION AND FAILURE ANALYSIS

Ajit Kambil
Department of Information Systems
New York University
Leonard N. Stern School of Business
44 West 4th Street, Suite 9-82
New York, NY 10012-1126
(212) 998-0843
fax: (212) 995-4228
akambil@stern.nyu.edu

David Bodoff
Department of Information Systems
Leonard N. Stern School of Business
New York University
44 West 4th Street, Suite 9-181
New York, NY 10012-1126
(212) 998-0822
fax: (212) 995-4228
dbodoff@stern.nyu.edu

Working Paper Series
Stern #IS-97-15

Partial Coordination: A Preliminary Evaluation and Failure Analysis[1]

Ajit Kambil and David Bodoff
Information Systems Department
Stern School of Business
New York University

Abstract: Partial coordination is a new method for cataloging documents for subject access. It is especially designed to enhance the precision of document searches in online environments. This paper reports a preliminary evaluation of partial coordination which shows promising results compared with full text retrieval. We also report the difficulties in empirically evaluating the effectiveness of automatic full-text retrieval in contrast to mixed methods such as partial coordination which combine human cataloging with computerized retrieval. Based on our study we propose research in this area will substantially benefit from a common framework for failure analysis and a common data set. This will allow information retrieval researchers adapting "library style" cataloging to large electronic document collections, as well as those developing automated or mixed methods, to directly compare their proposals for indexing and retrieval. This paper concludes by suggesting guidelines for constructing such a testbed.

---

[1] This work was supported by a grant from Disclosure Inc. Any opinions, findings, conclusions, recommendations, errors expressed in this material are those of the authors' alone and do not reflect the views of Disclosure Inc.

The authors also gratefully acknowledge the help of Kay Teel, Head Cataloger, Ingalls Library, Cleveland Museum of Art, and Sherman Clarke, Head, Original Cataloging, Elmer Holmes Bobst Library, New York.

1.0 Introduction

Partial coordination was proposed as a method for reducing out of context matches to queries submitted to online catalogs and indexes (Bodoff and Kambil 1997). This paper reports results from a preliminary empirical evaluation of partial coordination. We compare Vector Space Model[2] *retrieval* on *full-text-indexed* documents, with partial coordination *retrieval* on *partially coordinated* documents.

There are many differences between these two approaches: Full-text indexing is derived by computer, while partial coordinated indexing is manually assigned. Full-text indexing does not use vocabulary (synonym or hierarchy) control (although information retrieval research offers some experimental techniques for automatic vocabulary control), while partial coordination can use vocabulary control. VSM retrieval is not coordinated, while partial coordination retrieval uses the partial coordination of document indexes. In comparing methods, the many differences between the two approaches makes it difficult to assign credit (blame) for success (failure) to any particular feature of these methods. Indeed the problems of credit/blame assignment to particular system features, and the effort and expense required to compare manually-assigned indexes to full text indexing make such experimental comparisons (e.g (Tenopir 1985)) extremely rare. Indeed, we know of no direct experimental comparison of the retrieval effectiveness of LCSH versus full-text indexing. This is a glaring omission, especially considering the current debates regarding the role of libraries in the digital age, the applicability of LCSH to World Wide Web documents, meta-data guidelines, and the very need for any assigned indexing (cataloging)[3] in the digital era.

---

[2] The reader is assumed to be familiar with the Vector Space Model (Salton 1989).
[3] We use the terms "index" and "catalog" interchangeably to refer to a document's subject heading, according to ease of readability.

Thus a key contribution of this study is the direct comparison between the most common form of automatically derived indexing of electronic documents, i.e. full-text indexing, and a new kind of manually assigned indexing, i.e. partial coordination. Our results are promising, and we hope they will be considered as experimental evidence in support of Lynch's view of the continued need for librarians in the digital era (Lynch 1997). But while our results are promising, there were a number of challenges to creating a "controlled" experiment to compare these -- or any -- very different approaches to information retrieval. Thus, we outline and discuss the specific experimental design choices made to make this study economically feasible, as well as critically evaluate the results of our experiment. Based on this experience, we propose the need for a common data set and framework for comparing -- in terms of retrieval performance -- alternative approaches to both cataloging and retrieval. The framework should allow comparisons of automatic, manual, as well as mixed approaches. This proposal is a second contribution of this study.

This paper is organized as follows. Section two discusses the experimental procedure undertaken to evaluate partial coordination. Section three reports the key results of comparing partial coordination and full text retrieval methods. Section four undertakes a failure analysis and examines our results further to understand factors that reduce or enhance the effectiveness of partial coordination. Section five derives lessons from our study for future research and proposes the requirement for a well defined testbed for comparing cataloging and automatic indexing methods, and also proposes a framework for failure analysis. Section six presents conclusions.

2.0 Evaluating Partial Coordination: A Preliminary Experiment

A full empirical evaluation of the partial coordination method would compare it against pre-coordination cataloging and retrieval as well as against full-text indexing and retrieval methods. Comparison against pre-

coordination would indicate suitability for OPAC environments. Comparisons against full text indexing would indicate suitability for Intranet and WWW environments where full text data is available. In this study we undertook the latter experiment, comparing the performance of partial coordination to the results reported by the Cornell group at the TREC3 conference (Harman 1995), the latest TREC conference for which full results were publicly accessible at the time this study was initiated. This specific comparison was motivated by a previous review of our paper on partial coordination.

2.1 Experimental Design

In an ideal experiment comparing full text and partial coordination we would have a large document set in which every document was cataloged using partial coordination and also subject to full text indexing. In addition we would have a well defined set of queries on the document set for which the relevant documents would have been previously identified by an independent group of judges.

The TREC conferences have utilized the TIPSTER document collection for standardized testing of alternate automatic indexing and retrieval methods. This collection consists of over a million documents with about 3 gigabytes of data (Harman 1995). TREC uses this database and also provides information retrieval researchers with a consistent set of query topics and corresponding subset of relevance judgments. Given that TIPSTER provides a standard for comparing full text and automatic methods for information retrieval we decided to use a subset of TIPSTER as the sample for our evaluation.

In an ideal experiment of this kind, every TIPSTER document would be cataloged with partial coordination, in addition to being available for

automatic full-text indexing. This was impractical, given the expense of manual indexing (approximately $5 per document in our experiment). We therefore selected a subset of TIPSTER documents for partial coordination cataloging. A completely random sampling would not have been very efficient, since most TIPSTER documents are irrelevant to any given query. Our initial approach to sampling was to first randomly select a number of queries, then to choose a biased sample of TIPSTER documents which might be relevant to the query.

The queries were selected from the TREC3 ad hoc queries 151-200, for which we had available the results of Cornell's full-text retrieval. An earlier reviewer of this work suggested tf*idf full-text indexing with cosine ranking as a baseline against which to compare partial coordination. As the Cornell group performed well in TREC3, and since Cornell has been associated with the development of the Vector Space Model and also used a fine-tuned variation of tf*idf indexing with cosine ranking, we selected their results as a baseline against which to judge the potential of partial coordination. The Cornell team's results for the ad hoc query section of the TREC3 conference were downloaded from ftp-nlpir.nist.gov, together with the official TREC3 topics and relevance judgments. The sampling procedure of queries and documents for our comparative evaluation are discussed below.

2.1. Sampling

Defining the Query Set - Seven queries were randomly selected from among the fifty "ad-hoc" queries from TREC3. The query numbers are 153, 159, 173, 177, 187, 190, and 192. For each of these queries, TREC provides expert

relevance assessments for a large sample of the TIPSTER documents most likely to be relevant (see Harman (1995) for details on the pooling method by which documents are selected for expert review).

The TREC queries are themselves structured documents, which describe various aspects of the information need. Most TREC participants apply automatic indexing methods to these queries, but the rules allow for manual selection of query terms on the basis of the provided query description. In the sections below on VSM and partial coordination we discuss how the queries were represented under each system.

Defining the Document Sample - As described in Harman(1995), relevance assessments for a variable number of documents are available for each query in TIPSTER. As it would be prohibitively expensive to manually catalog each document with a relevance assessment in relation to a query, a biased sample of 787 documents was initially selected for cataloging. This sample was constructed as follows.

For a query with N relevant documents according to the TIPSTER relevance assessments, the population MIN(150,N) (i.e the lower of either 150 or N) of Cornell's top-ranked documents were selected. In addition, where N > 150, *all* relevant documents regardless of their Cornell ranking were included in the population of documents. Thus, the only documents *not* included in this population were those (the vast majority) which, for each query, did not rank highly by the Cornell search (within N or top 150) and were also not judged as relevant to the query. The included documents numbered 1446, and were

distributed to catalogers, but only 787 were ultimately cataloged, due to lack of catalogers' time.

Subsequent analysis revealed that this method of sampling could potentially bias results to favor partial coordination over full text indexing. In particular, under some circumstances, inclusion of all relevant documents (in addition to those ranked highly by full-text) could favor the competing method of partial coordination. To overcome this problem and be certain that the sampling bias completely and considerably favored full text indexing, we decided to limit the population of documents to the top MIN(150,N) documents as ranked by Cornell for each query. From the 787 documents that were cataloged, 418 met this new criteria.

The above sampling method assures a very conservative evaluation of the benefit of partial coordination. The method limits itself to the top-ranked Cornell documents for each query. As the definition of "top-ranked" depends on the query, for queries with only a small number of documents, there are a correspondingly smaller number of top-ranked Cornell documents to be considered. *This selection method amounts to limiting the experimental comparison to the documents retrieved in the very lowest levels of recall using full-text retrieval, where full-text retrieval performs best.* If the lower-ranked Cornell documents were included in the population, the relative improvement of partial coordination would in all likelihood be greater -- perhaps significantly greater -- than the results reported in this paper.

2.2 Indexing and Retrieval

Given the sample of documents and queries, indexing and retrieval mechanisms were adopted to enable the comparison of full text and partial coordination methods.

Baseline Full-Text System

To represent the full-text approach, we chose the Cornell team's implementation of the Vector Space Model (VSM) from TREC3. The Cornell system performs full-text indexing of documents, and full-text retrieval for each query. The full results of the Cornell team were downloaded from ftp-nlpir.nist.gov. These results show the retrieval score and rank of the top 1000 documents for each query. We used these published results as the basis of comparison. The exact representation of each document and query according to the Cornell algorithms is not published, and was not used nor necessary for this study. Rather than attempting to re-create and re-run the exact Cornell algorithms, we used their published *results* as a baseline for comparison.

The Partial Coordination System

To implement the partial coordination system we had to specify a process for partially coordinated document indexing, develop a specification of queries for submission to the partial coordination retrieval system, and implement a partial coordination retrieval system.

The partially coordinated indexing was mainly undertaken by four professional catalogers at New York University and the New York Public Library. For queries 187 and 190, a proportion of documents was cataloged by three New York University undergraduate students. All catalogers, both

professional and student, received a one-hour presentation of the motivation and method of partial coordination. All received compensation for their time. A random sample of documents was also cataloged by more than one cataloger in order to evaluate inter-cataloger consistency for a forthcoming study.

The documents were indexed in two groups. The first group consisted of documents for queries 153, 159, 173, 177, and 192. The second group consisted of documents for queries 187 and 190. There were two important differences between these two stages. In the first stage, as previously mentioned, only professional catalogers were used. A second difference is that in the second stage, catalogers were allowed to use an OR operator in their dependencies in addition to the implied AND operator, so that a term could be specified to depend on any arbitrary combination of other terms. For example, the term "proposal" could be specified to depend on either the two terms "nuclear" and power", OR on the single term "pollution". This feature was added after a meeting with the catalogers following the first round of five queries. The catalogers were not aware of the content of the queries against which the documents would be compared.

We used two different methods for query formulation. One method took the full text of the topic description supplied by TREC, omitted stopwords, and included all remaining words as query terms. In the second method, each topic was shown to a group of MBA students, who were asked to read the topic and select query terms for it. In both approaches, each selected query term was treated as an individual term. Even if a student indicated that two words form a phrase, each term was entered individually, as one aim of

partial coordination is to relieve the user of any need to determine whether two terms form a composite term or phrase. Appendix 1 includes a definition of the partial coordination scoring algorithm, and Appendix 2 includes one example query description and one example document subject heading using partial coordination.

A partial coordination retrieval engine was implemented to read all the document indexes and respond to user queries by retrieving and scoring individual documents. This system was implemented in the C programming language. Like Cornell, we used a version of the SMART stemmer for both document and query terms. The score of each document with respect to a query was computed as described in the companion paper (Bodoff and Kambil 1997), with the exception of queries 187 and 190, for which the algorithm was extended to account also for the OR operator as described above.

As described, each query was formulated using two different methods. The second method effectively results in a distinct query formulation for each student. Thus, there were many query formulations for each topic. Furthermore, many documents were cataloged by more than one cataloger. To account for the numerous versions of each document entry and query, the score of each document for a given query was its *average* score across all formulations of the query and across all different corresponding cataloger entries. Separate analyses were performed to differentiate the two query formulation methods and the different document cataloging styles. In those analyses, not reported here, the alternative methods of query formulation were found to have little impact on the final retrieval results. The results

reported here are based on the average scores across all query formulations and document entries.

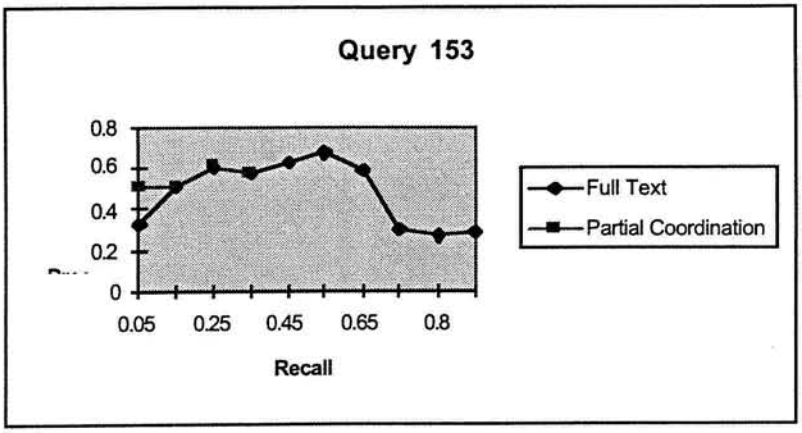
2.3 Experimental Comparison

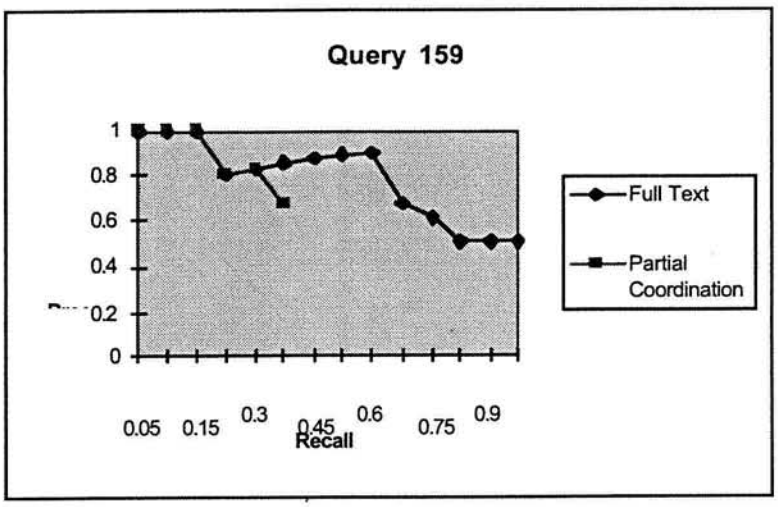
To compare systems, a ranked order of documents was determined for each query on each system. For the full-text system, the original published rankings were used, except they required transformation into a contiguous series 1,2,3,... to "skip over" the documents not selected in the random sample. For partial coordination, the query was run and documents were ranked by the system according to the average score across all versions of the query formulation and document entry.

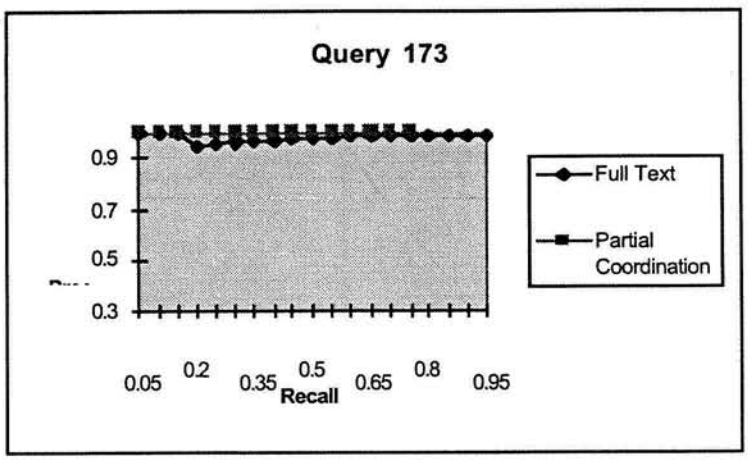
For each query, total recall is defined as the retrieval of all relevant documents from among the 418 documents. Recall and precision curves were then plotted for various levels of recall. These curves presented below were used to compare systems. These curves all represent a sort of magnification under two approaches of the documents represented in the high recall end of the original recall-precision curves achieved by the full-text system over the whole TIPSTER document collection.

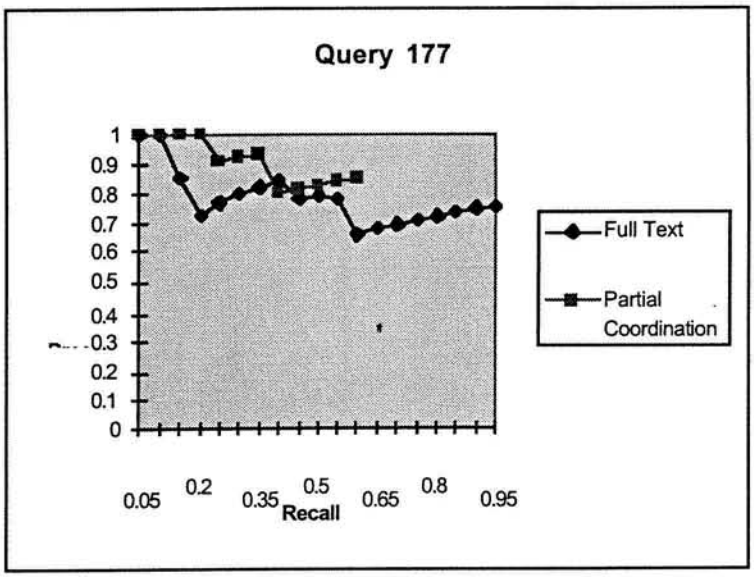
3.0 Results and Analysis

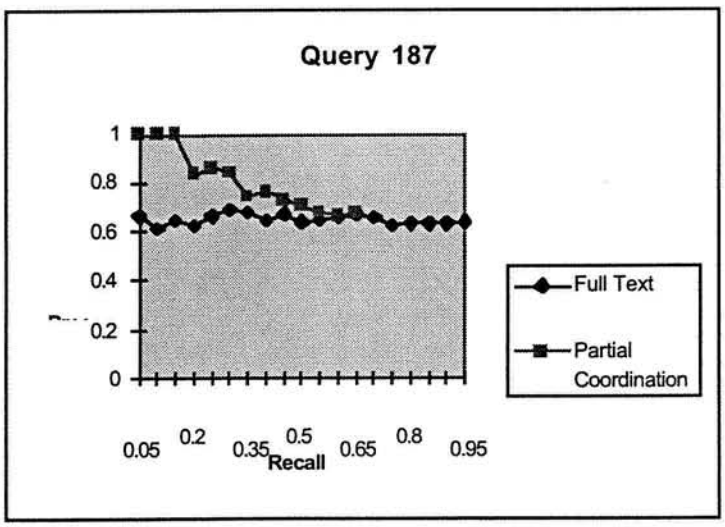
The plots below provide the precision-recall curves that compare the results of full-text and partial coordination for each of the seven queries. These results were positive and encouraging. For nearly every level of recall for each query, partial coordination showed greater precision than full-text retrieval. In some cases, the improvement was very significant.

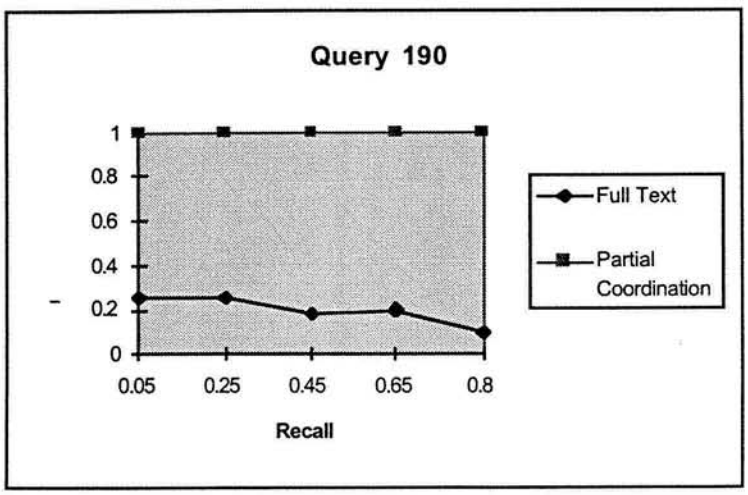


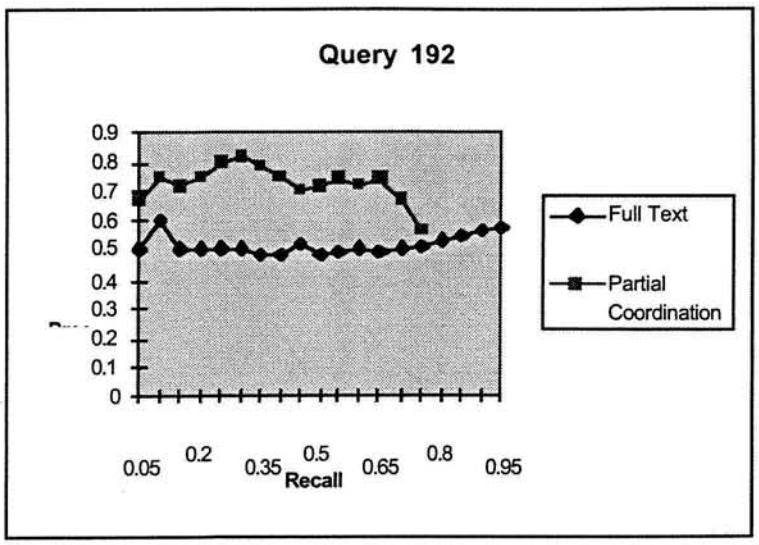












For most queries, the recall-precision curves for partial coordination do not extend all the way to total recall. This is a reflection of the fact that under partial coordination, all subsequent documents were assigned a score of zero and not retrieved to satisfy those levels of recall. This result would appear to affirm the results of Tenopir and others (Tenopir 1985) that full-text produces higher levels of recall. But in the current experiment, this result is merely an artifact of experimental design. In this conservatively constructed experiment, total recall was *effectively defined* as the set of relevant documents retrieved by full-text, so it was inevitable that partial coordination would produce lower recall than the full-text competition. The curves are presented for each query separately, rather than in one overall curve, because of the difficulty of combining them meaningfully where no data is available for one or more of the queries. The separate curves indicate that partial coordination may be better suited to certain types of documents or queries to be determined in future research.

The above comparison results are very conservative estimates of the benefits of partial coordination. To represent the full-text approach, we selected results of an algorithm which benefited from years of refinement, and was *specifically optimized* over three years of TREC conferences for improved performance on the TIPSTER database. Moreover, we selected documents from the best-performing highest part of the full-text recall-precision curve. Against this, we compared the results of our first attempts at partial coordination. We therefore believe the above recall-precision curves and reported results are reliable and conservative indicators of the promise of partial coordination.

Although partially coordinated indexing and retrieval compared favorably with full text indexing and retrieval, it is not clear whether this result is due to the partial coordination of subject terms per se. In order to isolate the advantage of partial coordination as a means of coordinating terms, an experiment is required in which catalogers choose subject terms for traditional post-coordinated retrieval, separately choose subject terms for partial coordinated retrieval, and then compare performance of each retrieval method with its corresponding catalog of document entries. The same approach would be repeated to compare the use of partial coordination to pre-coordination. It is not possible to test performance of the same subject terms with and without pre-, post- or partial coordination, because the choice of terms is affected by the method of coordination, as discussed in the companion paper. The experiment reported here compares two very different approaches to indexing and retrieval, and the use of partial coordination has not been isolated. It may be, for example, that carefully manually selected free text subject terms chosen for post-coordinated retrieval would also out-

perform full text retrieval, even without the use of partial coordination. We know of no such published experimental comparison. We return in section 5 below to the question of isolating features of indexing and retrieval in experimental comparisons between systems.

4.0 Further Analysis of Partial Coordination Results: Failure Analysis

In order to better understand the factors that affect the performance of partial coordination, we undertook a *failure analysis* which examined cases of false hits and misses. The ranked results of partial coordination were analyzed and each failure was assigned by the authors to one possible cause.

Because retrieval results are ranked, it is unclear what constitutes a (false) hit or a (false) miss. Our literature review identified no specific guidelines for failure analysis in the case of ranked results. Thus we defined a false hit as any irrelevant document ranked above at least one relevant document. This definition was motivated by the fact that some "errors" in a ranked retrieval environment are "relative". In contrast, in the absence of ranked results, where a document is either retrieved or not-retrieved, as in (Tenopir 1985), a false hit can only be attributed to improper inclusion of a term in the document index. But in a ranked retrieval environment such as that of partial coordination we often found that the cataloger's choice of terms resulting in a false hit was reasonable, but the false hit occurred because other *relevant* documents should have been indexed to achieve a *relatively higher* score and rank. In a ranked retrieval environment, an irrelevant document with a non-zero score results in an imperfection in recall and precision, only if it results in a *relative* ranking higher than at least one relevant document. Thus if the "error" is a relative one, then the blame for the false hit can be

assigned to either the high-scoring irrelevant document or the low-scoring relevant one. Table 2 below uses fictitious document rankings to illustrate our definitions of false hits and false misses.

Document	Partial Coordination Score	Rank	Relevance Assessment 1=Yes, 0= No	Error Type
WSJ970614-0010	2.3	1	1	No Error
AP974415-1210	2.1	2	0	False Hit
WSJ961212-1902	1.9	3	1	False Miss
WSJ970912-0101	0.5	4	0	False Hit
DOE1-13-173	0	5*	0	No Error
DOE2-12-013	0	5*	1	False Miss

As illustrated by the fourth ranked document in Table 2, an irrelevant document is considered a false hit when it is ranked relatively higher than at least one relevant document. This is effectively a "relative" false hit compared to the sixth document in the table.

Similar to false hits, we defined a false miss as any relevant document which is ranked after at least one non-relevant document. As illustrated by the third ranked document in Table 2, a relevant document is considered a false miss when it is ranked relatively lower than at least one irrelevant document. Thus the third ranked document is also effectively a "relative false miss" compared to the second document.

4.1 False Hits and Misses

From a review of failures, we identified possible proximate causes of false hits and possible proximate causes of false misses. Proximate causes for false hits were:

1. The cataloger defined document index terms inappropriate for the document,
2. A document index term was appropriate for the document contents, but only within a specific context not specified by the cataloger
3. Document terms were appropriate and within context, but other, relevant documents were not cataloged with greater exhaustiveness and depth to give them relatively higher scores and ranking

The latter two categories are sometimes hard to distinguish, as the following example shows. Query 177 is about legislative proposals to make English the official language of (or in any part of) the United States. Document AP880516-0234 is about efforts to promote adoption of Filipino as the official language of the Philippines. According to the expert relevance assessments of TREC, this document is not relevant to Query 177. The cataloger of this document included the term "official" with "language" as a dependency, and "language" as a term with "official" as a dependency, to form the phrase "official language". This gave two points for the document with respect to any query containing the two words "official" and "language", and zero points to any query containing one or neither of these terms. While the cataloger included other terms in the index, they are not relevant to the current discussion. As all formulations of the query used the terms "official and "language", this non-relevant document on the Philippines matched the query with a score of at least two. A relative false hit was thereby created, as other documents

which actually were about proposals for an official language in the US also scored no more than two. An analysis of this false hit identifies that causes two and three are possible causes of this false hit, but deciding between these two related causes may be difficult. Should we assign the failure to reason two, and say that the term "official language" in this document should have been further limited to the context of the Philippines? Perhaps, but then this document would achieve a zero score for someone interested in efforts to adopt official languages in general. Thus we should probably say, that inclusion of the term "official language" was not an error, and that the absence of further term dependencies was also not an error, but that documents which are about the adoption of *English* as an official language in the *United States*, ought to have been cataloged with sufficient depth so that they would rank *higher* than the Philippines document in question. This case is therefore best considered a ***relative*** false hit, and blame is assigned to the lack of exhaustivity and depth of ***other more*** relevant documents. Thus causes two and three are related and separation of these categories is often difficult. When in doubt, the assignment of failure was split between these two cause categories which, although logically intertwined, were nevertheless considered useful to enumerate separately.

Proximate causes for false misses were:

1. A missing term from the cataloger's index representing a lack of ***exhaustive*** indexing.
2. A missing term from the cataloger's index representing a lack of ***depth*** in indexing.
3. A term was missing in the index but its synonym was included.

4. A term was included but inappropriately hidden by a specified dependency.
5. A term was included and appropriately hidden by a dependency, but the query used a synonym for the dependency
6. Data entry error, including typo's and errors in following data entry syntax specified in cataloging instructions issued to catalogers
7. Word stemming inaccuracies

We adopt Foskett's definition of the terms "exhaustiveness" and "depth" in indexing (see companion paper (Bodoff and Kambil 1997) section 3.0 footnote). Document AP881102-0249 illustrates a false miss of the second type, in response to Query 177. This document reports "a brief look at major statewide and local ballot issues...". One sentence in the document covers ballot proposals regarding English as an official language. The cataloger included general terms such as "1988 election", but did not include terms to represent the substance of each of the specific proposals. This is a problem of not indexing the document with sufficient depth. In contrast, full-text indexing does not suffer from this problem. The first three categories of false misses are self-explanatory.

The fourth type of a false miss represents a "backfiring" of the use of dependencies to specify context. In document AP881011-0169 for query 177, the cataloger defined the term "official language" to depend on the two terms "English" and "Florida". The content of this document was about a Florida ballot initiative to adopt English as the state's official language. Including "Florida" as a dependency had the effect of preventing the document from matching any query which did not specifically include that contextual term.

No formulations of query 177 specifically refer to Florida, so the document was inappropriately prevented from matching. This example, illustrates the flip side of our previous discussion regarding various forms of false hits. The cataloger who specifies a narrow context runs the risk of a false miss, as in the current example, while the cataloger who omits the narrow context depends on exhaustive and deep indexing of other documents to prevent a relative false hit.

In partial coordination, the problem of synonyms applies to dependency terms as well as to subject heading terms. Category number five represents cases where the cataloger specified a dependency term for context, and the user indeed intended that context, but used a synonym to represent it.

Table 3 show the relative percentages of the causes of false hits for each query, while table 4 show the relative percentages of causes of false misses.

Table 3 False Hits

	Inappropriate Term	Missing Context	More Relevant Doc's Undistinguished	Other
153		25%	75%	
159	50%	25%	25%	
173				
177		63%	37%	
187		30%	55%	15%
190				
192		16%	84%	

Table 4 False Misses

	Term Missing: Synonym	Term Missing: Lack of Exhaustivity	Term Missing: Lack of Depth	Typo	Term Inappropriate Hidden by Context	Term Hidden by Context Synonym	Word Stemming Problem
153	49%	10%		17%	5%	19%	
159	79%	12.5%	8.5%				
173	30%	25.5%	2.5%		42%		
177		69.2%			30.8%		
187	15%						85%
190							
192		11%		89%			

Two of the main sources of failure -- the third cause in table 3 and the second cause in table 4 -- result from insufficient exhaustivity of indexing. The other

primary source of false hits is a lack of context. The other primary source of false misses is synonyms.

The primary focus of the first part of this paper is on the notion of context in indexing, and how partial coordination may be used to indicate context and thereby enhance recall and precision. The experimental results summarized in table 3 show that the problem of context is present even when catalogers were given the tool of partial coordination. The source of false hits labeled "lack of context" includes all five sorts of out of context matches outlined in the companion paper -- i.e. the missing contexts were needed to establish one of the following: phrases, word meaning for polysemous words, broader context for narrower terms used in deep indexing, broader context for secondary terms used in exhaustive indexing, and term orderings for topic/sub-topic relationships. The greatest number of false hits due to lack of context was due to a failure to establish topic/sub-topic relationships between subject heading terms. Partial coordination appeared to be used successfully to establish phrases and to establish a broader context for narrow terms used in deep indexing. There were few cases of secondary topics matching out of the context of a broader topic, but this is perhaps due to the fact that indexing was generally not exhaustive. These results empirically confirm the potential problem of context as a source of false hits. They also indicate that the tool of partial coordination was not utilized to its full potential in establishing context between terms.

The insufficiency of exhaustivenes is evident in the third column of false hits in table 3, as well as the second column of false misses in table 4. The documents in the TIPSTER collection are "newsy", as they report many facts

in few words. Greater exhaustiveness in indexing would reflect the particular facts reported in each document, in addition to the broad topic, and thereby help differentiate relevant from irrelevant documents. In the companion paper, the availability of term dependencies was proposed as enabling greater exhaustiveness by reducing the fear of false hits. The catalogers in this experiment do not appear to have taken advantage of the possibility of greater exhaustiveness.

It is possible that more complete training of catalogers would be needed before they develop the cataloging habits which would take greatest advantage of the new method. Perhaps even an in-depth review of a few of the failures would convey more clearly to catalogers the best use of term dependencies. Greater exhaustiveness in indexing is a cataloging habit which would have to be learned, especially by catalogers whose professional work involves LCSH in which more than one topical subdivision is unusual. One wonders whether manual indexing is best suited to dense, "newsy" documents whose contents are not easily summarized. And better use of partial coordination to establish context apparently also requires additional training.

In fairness to the catalogers, and as additional evidence of the conservative nature of the results of this study, we note that at the time the catalogers received their one-hour presentation of partial coordination, we had not yet completed our analysis of the various meanings of "context" and the corresponding various uses of partial coordination. The catalogers were therefore instructed only on the technical meaning of partial coordination, and were given no additional guidance regarding its use.

Synonym control was also responsible for a large number of false misses. In future studies we intend to utilize automatic synonym control.

Care should be taken, however, in interpreting the results of this failure analysis. There is no guarantee that the overall retrieval effectiveness would improve if a specific source of failure were removed. For example, if synonym control or greater exhaustiveness were introduced, *new* false hits could arise.

4.2 Example Successes of Partial Coordination

As hypothesized, we observed many cases where partial coordination was more effective than full-text retrieval at improving precision, due to specification of dependencies. For example, Query 190 regarded the use of electronic computers to perpetrate fraud. Partial coordination performed many times beter than full-text retrieval for this query. Full-text retrieved any document about any one of electronics, computers or fraud. Many of the false hits were technical Department of Energy documents in which the term "computer" or "electronic" appeared at least once. DOE1-11-0130 begins with the sentence "The technique of the numerical simulation of plasmas can be readily applied to problems in accelerator physics." Subsequent sentences in this short document use the word "computation" and "electron", which match the terms "computer" and "electronic" after word-stemming. The partially coordinated catalog entry also included the term "computer", but with a dependency of the term "plasmas", preventing the document from matching computer-related queries which were not also about plasmas. In this example, deep indexing was used to specify that the document is about a particular aspect of plasmas -- i.e. a computational aspect which is applicable

to other fields. This deeper indexing was facilitated by partial coordination which allowed the cataloger to include the term "computer" to specify the aspect of plasmas being reported, without fear that the term "computer" would match by itself, out of context, or in the context of a different broader term.

As another example, document AP891025-0131 regards a stamp fraud ring that cost the post office $16 million. The term "fraud" caused a false hit in full-text retrieval. In the cataloger entry, the term "fraud" was made dependent on the term "stamps" and the false hit prevented. This term dependency may be viewed as establishing the topic/sub-topic relationship stamps--fraud. On the other hand, had the query been about attempts to defraud the government in general, this specific dependency would have been too restrictive. A more effective cataloging practice would use a number of broader terms joined with the OR operator to define possible alternate dependencies (e.g. "fraud" depends on "stamps" OR "government").

Query 187 regarded the demise of independent publishing. Document AP891101-0291 regarded the announced retirement of Robert Bernstein, CEO of Random House Inc. As an example of the successful use of context dependencies with the OR operator, the partially coordinated entry included the term "publishing" with the dependency "resignation" OR "retirement", so that the document would not match just any unrelated query on publishing.

Manual indexing also highlighted other benefits and difficulties. Manual indexers can, in the presence of any ambiguity, define where a document

begins and *ends*. Librarians also control the *selection* of the documents to be acquired and cataloged. These functions are much more difficult to perform automatically. For example, many Wall Street Journal articles in TIPSTER are reproductions of the "What's New" column on the paper's front page. This column consists of numerous unrelated brief reports, but is defined as one document in TIPSTER. In TREC, the choice was made to consider these as atomic documents, in order to test the full-text retrieval algorithms under "realistic" conditions of an undifferentiated, continuous stream of electronic text (Harman 1996). Our catalogers were confused over how to index such a document, since it obviously covered many different topics, and would normally have been divided into individual segments by any human cataloger. In most cases, our catalogers just gave up, and rather than indexing all the various topics, indexed none of them, or the topmost one. In order to be able to directly compare results of partial coordination to the full-text results, we had to play by the same rules as TREC, so these "multi-documents" remained undivided, and catalogers simply struggled with them. But these rules of the game artificially suppress a very important role that human catalogers could play in indexing online documents, namely, catalogers could identify where a document begins and ends, a task which is far from obvious in the electronic context. This advantage of humans is in addition to the widely recognized advantage of humans in the editorial process of selecting documents worthy of indexing. These advantages can and should be studied and quantified in experimental settings which capture the ability of different indexing and retrieval systems to properly divide the electronic stream into coherent document chunks, and to select better documents in the first place.

5.0 Implications for Future Research

The emergence of the digital age has raised fundamental questions about the role of traditional cataloging methods. Where past decades of research into OPACS largely assumed a traditional form of cataloging, current research and standards workgroups are re-thinking the applicability of those methods in the digital age. Considering the importance of this research for the practical and academic future of library science, it is important that empirical research be facilitated, so that various indexing and retrieval schemes can be directly compared, and so that research results are comparable across studies to allow the field to quickly accumulate knowledge. In particular, it is important that empirical research establishes (renounces?) the value of manual subject cataloging in the electronic environment where full text data is available.

The results of our preliminary evaluation of partial coordination were encouraging. These results suggest the need for further evaluation of partial coordination in comparison to traditional pre-coordinate and post coordinate indexing and retrieval mechanisms. More generally, these results are evidence of the potential advantages of manual indexing. Such results should encourage further research into questions such as these: How can manually assigned subject cataloging be improved to work with online retrieval? Under what conditions does manual indexing out-perform automatic indexing? How can manually assigned subject cataloging be made more efficient, so that it is feasible for the plethora of electronic documents ? Under what conditions is the improved performance of manually assigned indexes worth the additional cost? Can document authors or document users be effectively trained to catalog documents? Can collaborative processes be effectively implemented to build quality archives? Extending research to account for the

economic value of different methods and cataloger and user processes for increasing retrieval and indexing effectiveness, will guide the implementation of such systems in real settings.

But it is costly to execute empirical research comparing manual versus automatic or mixed mixed model approaches

In this study, we directly compared a fully automated approach to full text indexing and retrieval against the mixed model approach of partial coordination which involves manual indexing and automatic retrieval. We encountered first-hand the many logistical and theoretical difficulties which attend this sort of experiment. These difficulties included significant sampling issues, issues regarding query construction, and the difficulty of ensuring that the two methods were executed with comparable quality. These difficulties -- and the investment required to overcome them -- must certainly help explain the scarcity of empirical studies which directly experimentally compare alternative theoretical approaches to subject cataloging. This section discusses the feasible steps required to facilitate this research to experimentally compare different indexing and retrieval methods, including approaches which are manual, automatic, or mixed.

5.1 Need for a Common Testbed for Research

The sub-field of computer science dealing with full-text indexing and retrieval has benefited very substantially from the TREC conferences. The accumulation of knowledge in this field over the past five years is already tangible, as best practices form each conference are adopted by other TREC participants in subsequent years. But the TREC conference and the TIPSTER

dataset are geared towards testing alternative full-text indexing and retrieval algorithms. We argue that a testbed is necessary for the general field of library and information science where fundamental questions about methods of indexing and retrieval can be empirically investigated.

In the current study we had to use the TIPSTER database and TREC queries as it was the only dataset widely available with a) a defined body of documents, b) queries on the documents, c) expert relevance assessments of documents, and d) results of other indexing and retrieval engines. However, TIPSTER was designed for comparing automatic indexing and retrieval on large textual document sets. Given its large size it is infeasible to fully catalog this document set, hence the biased sampling and selection of documents in the current study.

In addition to the size of the document base, there are other features of a testbed which make it suitable for testing fundamentally different approaches to indexing and retrieval with varying degrees of human intervention. For example, as previously indicated, the testbed would need to define measures of success in a manner which accounts for all the possible advantages of human intervention, including editorial policy and the segmenting of knowledge into coherent "chunks". In addition to varying degrees of human intervention, the various approaches to be tested could differ along any of the previously studied dimensions such as vocabulary control, exhaustiveness of indexing, type of term coordination, etc., as well as along new and still undefined or unimagined dimensions such as inclusion of various kinds of metadata, etc. Just as the testbed must allow varying degress of human intervention, it must allow each of these other features to be leveraged to

varying degrees by the competing systems. The testbed must define a dataset and procedures which are not biased in favor of any of these features.

Additional complications arise when comparing two methods which differ in their approach to *both* indexing and retrieval. For example, to properly compare LCSH with full-text indexing, we would want to compare LCSH indexing as it would be used in a pre-coordinate search, with full-text indexing as it would be used with post-coordinated search. It is difficult to even measure the success of a two-step pre-coordinate search -- are we looking for the proportion of subject headings which match the user query exactly? partly? first terms? the proportion of bibliographic entries under those headings? etc. -- and still more difficult to formulate a fair direct comparison of retrieval effectiveness between the pre-coordinate retrieval of LCSH headings and the post-coordinated retrieval of a full text index.

We do not provide here a comprehensive list of a the features of the proposed testbed which would render it suitable for directly comparing such fundamentally differing approaches to indexing and retrieval. We have raised a few examples of the criteria for such a testbed, including the manageable size of the document base and the inclusive measures of effectiveness. In general, the testbed must provide an unbiased environment in which the costs and benefits of any approach -- manual or automatic -- to indexing and retrieval can be evaluated.

While other experiments on cataloging have been conducted and document sets developed for research, no single testbed has been adopted or even proposed as meeting all the criteria necessary for direct comparison of all

indexing and retrieval methods. Testbeds in the six NSF digital libraries projects are presently focused mainly on automatic indexing and retrieval of multimedia documents, and efforts such as the Dublin CORE for specifying metadata standards do not provide a standardized testbed for *empirical testing* of the retrieval performance actually obtained by using the proposed metadata. Many of the NSF initiatives have limitations that restrict the use of the data to specific universities or only after a specific time. Other testbeds such as Harvard Business Review used in (Tenopir 1985) might additionally pose access restriction problems.

A testbed is only a first step to advancing research into cataloging methods. A second requirement is a generalized framework for failure analysis.

5.2 A Common Failure Analysis Framework

In a simple world, a controlled experiment could empirically test the benefits of an individual feature of a cataloging method -- e.g. pre-coordination, controlled vocabulary, etc. This could be done for each individual feature in turn, and the best features would be quickly discovered. But the world is not this simple. Features of indexing interact with features of retrieval, with one another, and with the environment (e.g. the document collection, the user population). Furthermore, some configurations of indexing and retrieval features tend to be intertwined in theory and in practice -- e.g. free text and full text indexing. Because any experiment involves more than one isolated variable, detailed analysis is required to shed light on the *reasons* for any failures of the multi-faceted indexing and retrieval approach being examined. This analysis is referred to as "failure analysis".

Failure analysis can be the basis for identifying specific strengths and weaknesses of an indexing and retrieval system. It can take many different forms. Some studies are specifically designed to compare the effect of a specific variable/factor on indexing and retrieval. Svenonius (Svenonius 1986), for example, provides a critical review of studies designed to test the effects of vocabulary control; Sievert (Sievert and McKinin 1989) reports one such experiment. In these studies, the blame for failures is implicitly attributed to the one variable being tested. Other studies compare two systems with many different features. In these cases, it is difficult to assign credit or blame to a specific feature difference, so these studies primarily emphasize performance measures rather than a failure analysis (e.g (Tenopir 1985)). In still other studies, the emphasis is not on comparing alternative systems or in testing a particular feature, but on exploring the use of one system, and categorizing the types and causes of failure. Markey's well-known work (Drabenstott and Vizine-Goetz 1994; Drabenstott and Weller 1996; Markey 1984; Markey 1985; Markey 1988) is a good example of this approach. Still more fundamentally, the literature pays insufficient attention to even clarifying whether a particular experimental comparison holds constant the indexing method in order to compare retrieval methods (as in (Drabenstott and Weller 1996) which compares retrieval methods assuming MARC records with LC subjects) , holds constant the retrieval method to compare indexing approaches (as in (Tenopir 1985) which uses post-coordinate keyword search to compare the use of full text, abstract free text and controlled vocabulary assigned indexes), or compares two methods which differ in both their indexing and retrieval, as in the current study.

Failure analyses also differs on the basis of the unit of analysis. For example, Markey's work typically treats failed *searches*, in which a whole query is described in terms of a query-wide failure -- e.g. a large retrieval set, zero hits, lack of user perseverance. In contrast Tenopir and Sievert treat each document-query pair to indicate the reason a particular document was (not) retrieved for a particular query, then generalize findings across all these pairs, suggesting why full-text results in false misses (Sievert and McKinin 1989) or higher recall (Tenopir 1985).

A final very important difference between failure analyses, is the extent to which failures are attributed to a specific and well-understood attribute of the indexing or retrieval. For example, a failure due to lack of synonym control, apparently indicates a need for that feature in document indexing and query formulation. Even here, because of possible interaction effects, we may be able to advocate that feature (i.e. controlled vocabulary) only in the presence of system features and an outside environment which are identical to the experimental conditions. Other sorts of failure analysis, such as a report of large retrieval sets for some queries in a given system (Drabenstott and Weller 1996), are merely descriptions of the failure, and do not indicate its reasons or its possible solutions in terms of desirable system features. In summary, it is not even clear whether a failure analysis requires only a characterization of failure types, or proposed reasons for the failures. As indicated, both approaches appear in the literature.

Thus, failure analysis in studies to date is highly varied, with many failures not attributed to specific system features. This again limits the ability to

develop suitable comparisons across systems and to develop a cumulative tradition of best practices.

The failure analysis reported above in section 4.1 is another reasonable effort, but one which does not put forward a universally applicable framework for such analyses. Our failure analysis examined each incorrect match of a document to a query and these failures were assigned "proximate causes". Regarding units of analysis, we did not analyze the query level, but rather the level of document-query pairs, as in (Tenopir 1985) and (Sievert and McKinin 1989). Regarding the attribution of failures to a specific system feature, each of our "proximate causes" is more than a mere characterization of the failure, and does define a reason for the failure, but these reasons are *not* directly associated with an underlying and well understood feature of indexing or retrieval. For example, many false misses were due to lack of exhaustivity in indexing; but what feature of our partial coordination system resulted in less-than-adequate exhaustiveness?

In studies which compare two or more systems with many differing system features of indexing and retrieval, a mapping of proximate causes to underlying causes is highly complex due to interactions in system features. Such a mapping was beyond the scope of this paper. A review of the literature comparing alternate systems with many different features, found no studies which provided such a mapping to trace back each failure type to a specific system feature.

To allow accumulation of knowledge regarding best practices in indexing and retrieval, it is not enough that each proposed system be evaluated with a

common testbed. A common testbed allows us to learn which systems outperform others, without knowing the reasons for the superior performance. To facilitate accumulation of knowledge about the benefits of particular system features, the failure analyses for each system must also be done within a common framework, *and that framework must help us assign credit (blame) for system successes (failures) in terms of an underlying feature of indexing or retrieval.* This requires a one-time intellectual effort to provide a characterization of possible failures, each of which is ultimately related to underlying features of indexing and retrieval methods. Because the mapping from proximate to underlying causes depends on the particular *combinations* of indexing and retrieval features, the effort in building this framework will be significant, but it will be a one-time effort subject only to ongoing maintenance as new methods of indexing and retrieval are introduced. A first step in the direction of such a mapping is presented in the following paragraphs. The adopted framework would enable accrual of experimental results even across studies of very different systems. This is especially important to the field of library and information science which entertains the very different possibilities of automatic, manual, and mixed approaches to indexing and retrieval.

In order to relate a retrieval failure to a specific feature of indexing and retrieval, we first need a more complete model of the indexing and retrieval process. One possibility is Fuhr's (Fuhr and Buckley 1991) model of information retrieval. This process model identifies three distinct components:

- The translation of documents to their representation for query processing. We subdivide this process into these components: The

indexing language, the indexing rules, and application of those rules to the index.

- The translation of queries into a representation for query processing. We subdivide this process into the following components: The query language, and its actual application in a query.
- The query processing itself. We subdivide this process into the following components: The matching algorithm, and other dynamics of query submission and reporting of results.

The failure analysis framework would then list the possible kinds of failure for each specific subprocess. These more specific failures can then be traced to specific features of that subprocess. These may include the presence or absence of various forms of vocabulary control in document indexing or query formulation, manual versus assigned indexing, ranked versus non-ranked retrieval, full-text versus non-full-text indexing, pre-, post-, or partial coordination of terms, the use of query operators (e.g. Boolean, proximity), etc. The use of a complete process model for indexing and retrieval helps trace failures to specific subsystem features.

In summary, we believe that a common testbed would be significantly augmented by a failure analysis framework to trace failures to particular system features. We suggest that a complete model of the indexing (cataloging) and retrieval process will facilitate the tracing of failures to specific features of indexing and retrieval.

6.0 Conclusions

Our preliminary evaluation of partial coordination shows promising results and merits further empirical evaluation, including comparisons with standard LCSH document cataloging. More generally, this study reports promising results of a method for manually assigned indexing over the state of the art automatically derived indexes. As a result, the study provides support for further research on mixed (human and automatic) models of indexing and retrieval for emerging electronic archives such as documents on Intranets and Internet web sites. However, cumulative results in this area will best accrue when the indexing and research community develops a common document testbed for research and a standard framework to guide failure analysis, enabling meaningful comparison of results across systems and lowering the costs of research. We hope this study stimulates such an effort.

Appendix 1: Partial Coordination Scoring Algorithm with AND operator only

(see Bodoff and Kambil 1997)

> For each query term q, if q appears in the document's subject terms, and if ***all*** that term's dependent terms as specified for that document appear somewhere in the query, then q matches, and we add one point (or some function of query or document term weights for term q) to the score; otherwise, q is no match and we go on to the next query term

To account for the OR operator, each subject term is first normalized into disjunctive normal form, so that there may be many instances of a subject term, but each instance depends as usual on the conjunction of all term dependencies.

For example,
A depends on B OR (C AND D)
becomes
A depends on B (instance 1)
A depends on C AND D (instance 2)
Then, the above scoring algorithm is applied, slightly modified:

> For each query term q,
>
> > for each instance of q in the document's subject terms, if ***all*** that term instance's dependent terms as specified for that document appear somewhere in the query, then q matches, and we add one point (or some function of query or document term weights for term q) to the score; otherwise, go to the next instance of q in the document's subject terms;

Appendix 2: An Example TREC Query:

Topic: Instances of Fraud Involving the Use of a Computer

Description: Document will report instances of fraud accomplished anywhere in the world through the use of an electronic computer.

Narrative: To be relevant, document will describe an example or examples of the use of an electronic computer to gain an unfair or dishonest advantage against any entity (government, business, individual) anywhere in the world.

Example of Partially Coordinated Document Subject Heading

Document DOE2-09-0598

Subject Term	Dependencies
Computers	
Programming	
Works	Computers
Use	Computers

Note two rather abstract terms -- "works" and "use" -- are included in the document catalog. Normally, such non-specific terms would not be included in a subject heading because they could match an unlimited number of queries that are unrelated to this document. However, as partial coordination allows specification of the term "computers" as a dependency, this document will match a query with the term "use" or "works" only if the query is also about computers. In this way, the additional terms "use" and "works" can be included in the subject heading to differentiate this document from other computer documents which are not about how computers work or how they are used; at the same time, these additional terms will not cause false drops.

References

Bodoff, D., and Kambil, A. (1997). "Pre-Coordination + Post-Coordination = The Case for Partial Coordination." Center for Research in Information Systems Working Paper *IS-97-14*, New York University.

Drabenstott, K. M., and Vizine-Goetz, D. (1994). *Using Subject Headings for Online Retrieval*, Academic Press, San Diego.

Drabenstott, K. M., and Weller, M. S. (1996). "Failure Analysis of Subject Searches in a Test of New Design for Subject Access to Online Catalogs." *Journal of the American Society for Information Science, 47*(7), 519-537.

Fuhr, N., and Buckley, C. (1991). "A Probabilistic Learning Approach for Document Indexing." *ACM Transactions on Information Systems, 9*(3), 223-248.

Harman, D. (1996). Personal Communication. Electronic mail September, 1996

Harman, D. K. (1995)."Overview of the Third Text Retrieval Conference (TREC-3)." *Proceedings of Third Text Retrieval Conference*).Gaithersburg, MD,

Lynch, C. (1997). "Searching the Internet." *Scientific American*(March), 51-56.

Markey, K. (1984). *Subject Searching in Library Catalogs: Before and After the Introduction of Online Catalogs*, OCLC Online Computer Library Center, Dublin, Ohio.

Markey, K. (1985). "Subject-Searching Experiences and Needs of Online Catalog Users: Implications for Library Classification." *Library Resources & Technical Services, 29*(1), 34-51.

Markey, K. (1988). "Integrating the Machine-Readable LCSH into Online Catalogs." *Information Technology and Libraries, 7*(3), 297-312.

Salton, G. (1989). *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*, Addison-Wesley.

Sievert, M., and McKinin, E. J. (1989)."Why Full-Text Misses Some Relevant Documents: An Analysis of Documents Not Retrieved By CCML or MEDIS." *Proceedings of 52nd ASIS Annual Meeting*, (34-39).Washington, D.C.,

Svenonius, E. (1986). "Unanswered Questions in the Design of Controlled Vocabularies." *Journal of the American Society for Information Science, 37*(5), 331-340.

Tenopir, C. (1985). "Full text database retrieval performance." *Online Review, Vol. 9*(Number 2), 149-164.