PRE-COORDINATION+POST-COORDINATION= THE CASE FOR PARTIAL COORDINATION

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

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

January 1996
Revised July 1997

Working Paper Series
Stern #IS-97-14

Pre-Coordination + Post-Coordination = The Case Partial Coordination[1]

David Bodoff and Ajit Kambil
Information Systems Department
Leonard N Stern School of Business
New York University[2]

Abstract: The introduction of computerized post-coordination has solved many of the problems of pre-coordinated subject access. However, the adoption of computerized post-coordination results in the loss of some pre-coordination benefits. Specifically, the effect of hiding terms within the context of others is lost in post-coordination which gives lead status to every document term. This results in spurious matches of terms out of context. Library patrons and Internet searchers are increasingly dissatisfied with subject access performance, in part because of unmanageably large retrieval sets. The need to enhance precision and limit the size of retrieval sets motivates this work which proposes partial coordination, an approach which incorporates the advantages of computer search with the ability of pre-coordination to limit spurious partial matches and thereby enhance precision.

[1] This work was supported by the National Science Foundation under Grant No. 9319331 and a follow on 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 the National Science Foundation or Disclosure Inc.

[2] The authors wish to 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 University.

1.0 Introduction

In the era of card catalogs, user searches by subject were not as frequent as known item searches (Larson 1991; Markey 1980; Markey 1984), and thus received little academic or practical attention (Bates 1986; Cochrane 1983). However, the introduction of on-line catalogs (OPAC's) renewed users' interest in subject searches (Lipetz and Paulson 1987; Matthews and Lawrence 1984). With the new OPAC's, users searched by subject more often than any other search method (Markey 1985) but were frustrated by the various limitations of the early OPAC's (Besant 1982; Larson and Graham 1983, March; Markey 1984; Matthews and Lawrence 1984) cited in Drabenstott(1994) p. 123.

In response to users' demands for better subject access, researchers have proposed a series of rather elaborate online public access catalog (OPAC) designs, containing, among other things, on-line thesauri, syndetic structures, class schedules, and document clustering (Bates 1986; Cochrane 1985; Croft and Thompson 1986; Drabenstott and Vizine-Goetz 1994; Hildreth 1989; Larson 1989; Markey 1984; Markey 1988; Peters 1991). A promising development is the combination of probabilistic and Boolean retrieval in the Okapi (Robertson 1997) and Cheshire (Larson et al. 1996) systems. Nevertheless, subject searching still leaves much room for improvemens. Many of these proposals cited above remain in an experimental stage (Borgman 1996), and some place new burdens on users who have to master advanced system features for effective use, where it is known that even traditional Boolean queries are difficult for most users (Borgman 1984) cited in (Croft 1986), (Borgman 1996). For these and other reasons, subject searching remains difficult for users. The recent special issue of JASIS dedicated to OPAC's began with Borgman's paper entitled "Why are Online Catalogs *Still* Hard to Use?" (Borgman 1996). The editors of the special issue accepted the premise that "despite the advances in technology, research continues to show that these systems (i.e. OPAC's) are ineffective and hard to use" (Beaulieu and Borgman 1996).

The frequency of subject search failures, i.e. subject queries which return zero hits (Bates 1986; Lynch 1989; Markey 1988) or a very high average number of hits per query (Drabenstott and Weller 1996; Markey 1984; Prabha 1989) are the most frequently cited search failures. Taken together, these results imply that users are either getting few useful results or are overloaded with results in response to a given query.

As the costs of digital publishing fall, inexpensive online and Internet publishing is rapidly expanding the available number and types of searchable documents. However, for a given level of search precision, increasing the size of the searchable world of documents means a corresponding increase in the size of the result set, confronting users with an information overload problem. Users increasingly discover that queries to search engines like Alta Vista that index a large part of the public Internet result in unmanageably large response sets (Lynch 1997). To make matters worse, Internet indexing methods have heavily favored full text over other types of indexing. While there are some results to the contrary, the overwhelming evidence is that full text search results in lower precision ratios than keyword search for large databases (Blair and Maron 1985; Blair and Maron 1990; Sievert and McKinin 1989;

Svenonius 1986; Tenopir 1985)[3]. Moreover, as databases grow the problem of low precision in full-text search is *proportionally* worse (Blair and Maron 1990). Thus we can assume that the increasing size of result sets will render full text search of documents published on the Internet or other very large electronic document collections increasingly impractical and inefficient.

In this introductory paper we propose a new method which we call partial coordination that combines the strengths of two subject cataloging and search methods for documents -- pre-coordination, as in traditional card catalogs, and post-coordination, as in computerized keyword search. We propose this method will result in greater precision for a given level of recall. The proposal is made in the spirit of Lynch's belief that librarians can help computers bring order to the digital world (Lynch 1997). Following this introduction, section 2 critically evaluates the strengths and weaknesses of pre- and post-coordination. Section 3 examine the limits of post-coordination in greater detail classifying different types of out-of-context or imprecise matches. Section 4 introduces partial coordination and provides a detailed illustrative example. Section 5 examines the strengths and weaknesses of partial coordination and section 6 concludes by outlining the implications and future directions for research. Our follow-on paper to this introductory article provides a preliminary empirical evaluation.

[3] Some of these studies focus on reduced recall for free-text searches. Assuming a recall/precision tradeoff, however, we may infer that achieving identical levels of recall would hurt precision in those studies.

2.0 Pre-coordination and Post-coordination Critically Reviewed

Historically, library science focused on providing good subject access to documents using pre-coordination of subject terms. Fixing the citation order of each subject heading e.g., United States -- History -- Civil War in a card catalog system is known as pre-coordination. For a user of the catalog system to successfully find relevant information on the history of the US Civil War, they would have to:

- choose the right terms from the many possible synonyms which refer to a given concept, and
- choose the correct ordering from the many terms in a compound subject (i.e. a subject composed of more than one term). For example, know to look under United States -- History -- Civil War in an alphabetical catalog to find relevant documents.

The latter requirement is obviated by the introduction of computers and OPAC's. Using a keyword search, the computer would find documents labeled with the subject heading United States -- History -- Civil War, even if the user input the query terms in the 'wrong order' -- e.g. a query 'Civil War History United States'. Allowing the terms of a compound subject heading to be effectively re-ordered to match any query is known as post-coordination. In some communities such as the World Wide Web, post coordination (usually in the form of ull text indexing) is generally accepted as more practical and useful, and its costs and effectiveness relative to pre-coordination have been largely unquestioned. However, there are substantial benefits and limitations to both methods which are reviewed below.

In the following, we refer to pre-coordination as a method of indexing *and* retrieval. When discussing pre-coordination, then, we intend a situation in which the catalogers supply terms and an order among them, and this order is then actually used in the process of retrieval. Similarly, when discussing post-coordination we intend a situation in which subject terms are not given an order, nor are users anticipating any such ordering of terms. We focus our analysis at a level that allows meaningful comparisons without getting distracted by the specific implementation of a particular system (e.g the nuances of LCSH rules).

2.1 Pre-Coordination Strengths and Weaknesses

The enhanced precision and recall benefits of pre-coordination arise from the *standardization of term orderings*, and from *the selection of intelligent term orderings.*

Standardization of Order imposes a specific term ordering in the catalog "to ensure that the same composite subject is always treated in the same way, no matter how it may be expressed in natural language" ((Foskett 1977) page 80). This ordering enhances *recall* when the *same idea* can be expressed with different syntaxes using the same natural language subject terms. Standardized term ordering reduces the variability from different syntaxes, and provides a consistent way for users to find documents regardless of variety of ways of combining natural language terms to express the same idea. This particular benefit is also realized in post-coordination, since post-coordination ensure a topic can be repeatedly found regardless of the cataloger's or user's selection of a citation order of subject terms for the topic.

A second, often overlooked advantage of pre-coordination, is that standardization of order enhances *precision* when the same terms can express *different ideas* through different syntaxes. For example (after Foskett), 'Wars (due to) Economic Crises' is different from 'Economic Crises (due to) Wars'. In both cases the same terms are related by a different cause-effect relationship. Theoretically, pre-coordination can define a standardized ordering of concepts when two terms are related in such a semantic relationship. For example, where the terms are related by the "cause-effect" relationship, it may be standardized that the 'effect' term must always precede the 'cause' term. This standardization would allow 'Wars-Economics' to have an entirely different meaning from 'Economics-Wars', and appropriately assigns documents to distant parts of the card catalog, to avoid confusion on the subject[4]. In post-coordination the unordered list of query terms 'war, economics' matches documents on both subjects. Moreover, with post-coordination search a query term can match any subset of the document's subject terms, regardless of those terms' position in the subject heading. An arbitrary subset of the subject heading's terms can represent a very different meaning from the original heading. Thus the *precision enhancing benefit* of pre-coordinated standard orderings is totally lost in post-coordination. This precision enhancing benefit of pre-coordination is generally underemphasized in most comparisons of pre- and post-coordination (e.g. (Drabenstott and Vizine-Goetz 1994) p. 10). All these benefits are somewhat diminished in a dictionary catalog with a rule of specific entry as opposed to

[4] LCSH does not ordinarily rely on the standardization of term order to indicate such meaning. For example, the effects of A on B is distinguished from the effects of B on A not by the ordering of terms, but explicitly as "A--effects of B on" versus " B--effects of A on"

an alphabetico-classed catalog. However, even dictionary catalogs such as LCSH are strongly influenced by the principles of classed catalogs because of "the decided advantage of the alphabetico-classed catalog for grouping related objects together" which has proved too much for the "keepers of LCSH" to resist (Drabenstott and Vizine-Goetz 1994).

In summary, for every relationship among terms -- syntactic or semantic -- a cataloging formalism can be defined to standardize term ordering and thus enhance precision (along with recall). Various pre-coordinate schemes differ in the number and kind of relationships for which they standardize a citation order among the terms. One ambitious scheme of this kind is PRECIS (Dykstra 1987) which we discuss below in section 5.8.

While selection of any standardized citation order can achieve the aforementioned benefits, selection of ***intelligent term orderings*** further enhances the *precision* of pre-coordination. Regarding recall, consider a document on the "religious aspects of dreams". This document should be grouped with other books on dreams rather than with other books on 'religious aspects (of everything)'. This reasonable grouping is achieved by ordering the "dreams" term before the "religious aspects" term. This intelligent decision by a cataloger enhances recall by making it easier for a user to correctly guess the citation order.

Intelligent term orderings also enhance precision by grouping together like documents and un-grouping un-like ones. This means that once a user has found the *complete,* and *correct* subject heading, he or she will find mostly related and relevant documents. A second precision enhancing mechanism of intelligent ordering is the *avoidance of improper partial matches* by clearly defining the ***context*** of each heading term. In pre-coordination a partial match occurs when the user has guessed the first terms of a subject heading in the correct order, but he or she has not guessed all the terms. By not completing the subject heading, the user must now browse the various subdivisions. An intelligent ordering will help ensure that users will not achieve partial matches which require them to browse through mostly irrelevant material. The ordering ensures that if the user matches the subject's outer terms in a partial match, he will be browsing in the right ballpark; at the same time, it ensures that the inner terms are hidden, and cannot be the basis for a partial match.

In the example book on religious aspects of dreams, the enhanced precision is clear. Suppose this book were cataloged under the subject heading Religious Aspects -- Dreams. Then a user interested in this aspect of dreams, who partially matched by looking under 'Religious Aspects' would find himself browsing through numerous documents and subdivisions related to Religious Aspects of Everything, and totally unrelated to his area of interest. In this sense, he would have partially matched the term Religious Aspects out of context. It is better for precision if the term Religious Aspects is only found as a subdivision of the more substantive Dreams term. In that case, the Religious Aspects term would be matched only in the context of Dreams, and an inappropriate partial match would be avoided.

Pre-coordinate schemes use two mechanisms to establish the **context** for a term. Citation order is one mechanism, (inverted) term phrasing is another. For example, the LCSH heading 'Art, Asian' uses a term phrase to ensure that the term Asian matches only in the context of Art. With a compound subject, the cataloger prevents spurious partial matches by either forming a phrase or using citation order between terms.

While standardized and intelligent selection of term orderings provide a number of precision enhancing benefits, there are numerous limitations to traditional pre-coordination. First, there is a limit to the precision with which documents can be cataloged using pre-coordination. Most LCSH authority subject headings contain a single main heading with no topical subdivisions; only "occassionally" is there an authority heading with two or more topical subdivisions (Drabenstott and Vizine-Goetz 1994). This is a necessary -- not incidental -- limitation with pre-coordination, since users of pre-coordinated search are required to correctly guess each term in its correct order, in order to locate the intended subject heading. Only a fairly shallow subject heading has any chance of being properly guessed at by the user. But this limit on the number of subdivisions directly limits the potential precision of search results. Related to this, not only is the *number* of subdivisions limited, but only a limited *kind* of semantic relationships is represented in LCSH subdivisions. There is only a small number of subdivision types, e.g. form, location, sub-topic. And even the topical subdivisions often represent one of a limited number of relationships, often established in a pattern e.g. -- Economic Aspects. Here again, this limitation is not incidental to LCSH, but is a necessary limitation of a pre-coordinated approach[5]. Second, ordinary users

[5] One reason for the limited number of semantic relationships reflected in LCSH topical sub-divisions relates to the process of pre-coordinate search. As described in (Drabenstott and Vizine-Goetz 1994), pre-coordinated search is a two-step process, the first step of which requires the user to locate his intended subject heading. Only in the next step are actual documents retrieved. The subject headings must therefore be transparent in their meaning. An arbitrary implicit relationship between two terms is often undecipherable, hence the limited kinds of relationships reflected in pre-coordinated headings.

cannot be expected to encode their queries according to elaborate rules. Even in the relatively simple LCSH scheme, users find it difficult to guess the citation order of compound subjects (Bates 1977; Markey 1984). Steinberg and Metz (Steinberg and Metz 1984) found that only 28.2% of the users even knew the subject heading needed to be an authorized LC heading. Perhaps for reason of these difficulties, subject searches were less common than known-item searches in the era of pre-coordinated card catalogs (see (Larson 1991), contrary to (Markey 1984)). These results suggest that attempts to improve subject search results should introduce no new burdens on users.

2.2 Post-Coordination: Strengths and Weaknesses

The primary benefit of post coordination is to reduce the users' effort of learning any formalisms, including rules governing citation order, to construct queries. However, the benefits of standard or intelligent orderings are thereby lost.

The most significant loss in performance in contrast to pre-coordination is the loss of *precision* arising from *intelligent* orderings to prevent inappropriate partial matches. This is doubly unfortunate because partial matches are an even more significant issue in the post-coordinate document retrieval. With pre-coordination, a partial match only occurs to the extent the user guesses the first terms in their proper order; the match is then partial with respect to the later terms which he has omitted. In this partial match the user does not immediately retrieve documents but instead, is presented with a pre-coordinated hierarchy of subject headings. The topic and documents the user desires is somewhere below the subject heading fragment which the user has guessed. In post-coordinate systems, on the other hand, a partial match of one or more keywords actually retrieves lists of documents, and furthermore, any document is retrieved if any of its uncoordinated subject terms matches any term in the uncoordinated list of query terms. This large number of inappropriate partial matches may be reflected in lowered precision and large retrieval sets of post-coordinate keyword search.

Revisiting our previous example, suppose a user is interested in the religious aspects of dreams. He would like to specify the two keywords Dreams and Religious Aspects (as in LCSH subject heading Dreams -- Religious Aspects). To this user, a book on religious aspects of War (or of anything else) is of no interest, while a book about dreams which is not about its religious aspects has some relevance. But the user's query vector (Dreams, Religious Aspects) will retrieve documents of both sorts with an equal partial match[6]. The subject term 'Religious Aspects' from the document about religious aspects of war, will match the Religious Aspects term in the query about dreams. Yet this match is totally out of context, and represents a false drop. In this way, the loss of an intelligent ordering of terms may result in many false drops.

Post-coordinated search may therefore result in lower precision than pre-coordinated search, *given the identical subject terms*. But because the process of keyword selection is different when assigning an authorized pre-coordinated subject heading and when selecting post-coordinated keyword terms, the two retrieval methods typically do *not* operate on the identical subject terms. In the following paragraphs, we indicate that the differences in

[6] Term weighting, which may solve this problem in only this sort of simple linear case, is discussed below in section 5.3.

keyword selection may actually exacerbate the problems of false drops in post-coordinated retrieval.

In contrast to Library of Congress style pre-coordination, the *lack of a theory for selecting postcoordinate keywords* encourages the use of *more subject keyword terms* and *more narrowly specialized keyword terms*. With full-text indexing, these tendencies are automated. For manually assigned subject terms, reports of these tendencies are anecdotal but appear widespread.[7].

In post-coordination it is assumed the user can only benefit from the application of more keyword terms to a document. The user is not penalized in post-coordination for missing the citation ordering or for missing an additional subject term, so the additional terms can only lead to more matches, which is assumed to benefit the user. Adding more keyword terms seems reasonable. But if the additional keyword terms would rightfully be ordered after another term if term coordination were employed, then adding the term without coordination creates more opportunity for an out of context post-coordinate match.

The addition of *narrow* keyword terms can also exacerbate out of context matches. Despite the LCSH rule of specific entry, pre-coordinated subject heading terms describe a category rather than its instances. Fuggman refers to these as general concepts (Fuggman 1985). For example, "Insects" is an authorized subject heading, while the names of particular kinds of insects are not (with a few exceptions). This example was originally introduced by

[7] While we provide an argument for why expanded and narrow keyword selection creates precision problems, the extent of the problem remains open to empirical evaluation.

Fuggman as demonstrating the need for controlled vocabulary to enhance recall through "representational predicatability" (Svenonius 1986). But the use of general concepts in a controlled vocabulary actually enhances precision, especially in a post-coordinated environment. The reason for this is that a specific instance of a category may simultaneously be an instance of many other categories potentially causing false drops. For example, "Tarantula" may be considered by some as a poisonous insect, and by others as (say) a horror-film prop. The categorical subject term "Insect" will not cause an out of context match, but the additional narrower term "Tarantula" may indeed cause out of context matches, as a user may intend biological research and the document may be about B-movies (or vice versa). For example, an Infoseek search produced in the top 10 documents, some which regard tarantulas as pets, and one of a Native American folk story of creation (i.e. the world began with a tarantula). To some authors, a tarantula is an instance of a house pet, while to others it is an instance of a Native American mythical figure. The problem of non-categorical subject terms can be stated generally as follows: Non-categorical terms that are the instance of a specific category can simultaneously be instances of other categories. The addition of such narrow terms can thus further exacerbates the problem of out of context partial matches.

The two described characteristics of post-coordinate keyword selection versus pre-coordinate LC heading term selection -- i.e. more and narrower terms -- are not necessarily misguided. The addition of more and narrower terms for post-coordinated keyword search does indeed have some benefits. But these may be offset by reductions in precision. An empirical study of these tradeoffs is a target for future research. Our point here is not to resolve these issues but to indicate that the problems of false hits in post-coordinated search may be exacerbated by the subject term selection process.

In summary, pre-coordination enhances precision by specifying both *some* order and an *intelligent* order on terms. In contrast, post coordination reduces the user's and cataloger's efforts to select and order terms. However, post coordination does not recognize the benefits of term ordering and thus gives rise to spurious partial matches. As the volume of material increases in a post coordinate system, the number of false drops continues to rise. In an ideal post-coordinate system, users would not receive spurious partial matches, but would still be relieved from the need to specify term orderings, or the need to think about synonym and hierarchy control.

Below we critically examine and further classify the types of out-of-context partial matches. Partial coordination which we propose later in this paper overcomes many types of out-of-context partial matches, and aims to re-introduce the precision-enhancing benefits of term coordination to the post-coordinated online environment of OPAC's.

3.0 Out of Context Matches: Types and Issues

Out of context matches are a primary source of false drops. Understanding the types of out-of-context matching is critical to improving search methods. Our analysis identified five distinct and major types of out-of-context matches[8]. Some of these false drops are particularly problematic in full text search.

[8] We do not claim to exhaustively cover all out of context matches.

Ordered Relationship among Terms: In our previous example on "Dreams -- Religious Aspects" false drops can occur if a query term matches "Religious Aspects". The term "Religious Aspects" has *matched out of context of its relationship to another term in the heading.*

Polysemy[9]: Out of context false drops can also occur when a term in the document index differ in meaning from the same term in the query. Anyone searching the WWW with a full-text search engine such as Lycos or Infoseek will have encountered any number of surprising -- and sometimes very amusing -- false drops which result from a polysemous term. For example anyone searching for articles on the birth of Andromeda (Stoll 1995) may get articles on both the birth of starfish or a galaxy. Unlike the previous out-of-context problem, this occurs because an identical term (i.e. Andromeda) has different meanings in the document and the query. Polysemy is worst in the case of full-text indexing, since cataloging schemes would generally make some effort to avoid including polysemous words through a controlled vocabulary. This is consider a problem of "context" in the sense that a human reader -- or intelligent machine -- could identify the *various meanings from the context in which they appear.*

Out of Phrase Terms: These false drops occur when a query or document phrase is not treated as a single unit. For example, query #159 from the TREC3 conference (Harman 1995) asks about electric cars under development, and the full-text retrieval we studied (see companion paper (Kambil and Bodoff 1997))

[9] A word having multiple meanings

ranked as 36th (out of a million documents) an AP newswire about a couple whose *car* was hit by lightning in an *electrical* storm. In the newswire, the terms electric and car both appear, but not in the phrase 'electric car'. In these types of false drops, an individual term may occur within a phrase in the document heading, and in a very different phrase or as an individual term in the query (or vice versa). A false match occurs when a term matches *out of context of its term phrase*. This problem has been extensively addressed in IR research (Fagan 1987).

***Exhaustivity: Secondary Topic Keywords*:** A fourth sort of out-of-context matches occurs when the matching term in the document and query can be said to have an identical meaning, yet the document term does not represent a primary topic of the document. Manual assignment of subject headings can avoid this problem by not including keywords of secondary importance, but then a certain degree of recall and precision would be lost, as only the broadest topic of each document would be indexed. In the case of full-text indexing, much research effort has gone into this challenge of automatically identifying the 'about-ness' of documents. The goal of these efforts is to rank documents which are very 'about' the query topic, above documents which are less 'about' the query topic. In spite of those efforts, many of the false drops we studied from TREC3 were a result of this sort of error. This problem is essentially one of reduced precision from increased exhaustivity of indexing[10]. This problem may be viewed as a document matching 'out of context' in the

[10] We follow Foskett's definitions of exhaustivity and depth in indexing. According to these definitions, exhaustivity regards the inclusion of the document's less central topics, while depth regards the specificity with which each topic is represented (e.g. Siamese cats versus Cats) (Foskett 1977). This is somewhat contrary to the use in Chan as cited in (Drabenstott and Vizine-Goetz 1994)

sense that the matching terms are assigned unwarranted prominence by an algorithm *which does not understand the primary topic or 'context' of the document*. This problem is also addressed by partial coordination.

Depth: Non-Categorical Terms: As discussed in the earlier review of post coordination, the use of non-categorical terms as subject heading keywords poses problems because a query or document subject term which is an instance of one category (or broader term) can simultaneously be an instance of one or many other categories. This problem is essentially one of reduced precision due to increased depth in indexing. A false drop may occur when the narrower term matches *out of the context of its intended broader category*[11].

Users have two strategies where the document subject terms are non-categorical: To use the appropriate categorical term in the query despite its likely absence from the document headings ("fight 'em"), or to use in the query the narrower terms he expects to find in the document headings ("join 'em"). However both strategies are vulnerable to problems of recall and precision.

These five types of out-of-context matches reduce the effectiveness of post-coordinate searches. They are partly responsible for the phenomenon of WWW search engines returning thousands or even hundreds of thousands of irrelevant documents in response to queries.

[11] Where the narrower term explicitly includes the broader one, as in "Siamese cats", the increased specificity is unlikely to lower precision. But most narrower terms do not explicitly include their broader term e.g. "Beetle" rather than "Beetle Insect", etc.

The partial coordination method which we propose below can alleviate all the above problems except for polysemy. Below we introduce partial coordination and illustrate the method.

4.0 Partial Coordination

We propose partial coordination as a new method to inhibit inappropriate partial matches of keyword searches through the use of context. Partial coordination differs from pre-coordination by replacing *term orderings* with *term dependencies*, and from post coordination by using the dependency information to achieve better in-context matches between queries and documents. Partial coordination improves precision by defining high relevance scores for documents when particular ***combinations of dependent terms*** match in the query and document in contrast to independent matches of individual terms in both the query and document. In this way, a term might contribute to the match in the right context (i.e. the presence of another term or terms), but nothing or little in the absence of that context.

There are a number different approaches to defining term dependencies and document scores. The term dependency information must accompany either the documents' subject headings or the users' queries. A *user* can define the *query* terms, and then further specify term dependencies which determine the score a document has as a function of which query terms it matches. In the extreme case, for n query terms the user specifies a score for each of the possible 2^n-1 partial matches (e.g. if a document contains terms A and B, the score is 3, if terms B and D, score is 0, etc.). Alternatively, a *cataloger* can specify a *document's* keywords, and further specify term dependencies which determine the score this document will have in terms of potential queries. In the extreme case of the latter approach, for n keywords, the cataloger has to specify document scores with respect to each of the 2^n-1 queries which partially match the document's keywords. We adopt the latter approach in the sense that the dependency information accompanies each document's subject heading, rather than each user's query. This approach imposes the least effort on the user. In our approach the cataloger does not need to explicitly define 2^n-1 possible scores. Instead, he/she supplies term dependency information. A scoring function is then applied to each document during retrieval which calculates the score of the document as a function of the particular subject terms included in the user query.

For each partially coordinated document, the cataloger defines a list of index terms[12]. The list of index terms remains similar to pre- or post-coordinate terms. In pre-coordination the cataloger then forms *term phrases* and *term orderings* to prevent spurious partial matches. In contrast, with partial coordination the cataloger specifies *term dependencies* instead of term phrases and term ordering. Specifically, the cataloger defines each subject term in the document subject heading to depend on zero, one, or more other "dependency" terms. If subject term A is specified by the cataloger to depend on dependency term B in a particular document, then term A in this document will match a corresponding term A in a query *only if the query also includes the term B.* Only subject terms can match query terms. Dependency terms are not access points to the document, and cannot match query terms.

[12] We use the terms "index" and "catalog" interchangeably to refer to a document's subject heading, according to ease of readability.

But a term can be used as a dependency term and also as a subject term. For example, suppose the cataloger has specified that subject term A has the single dependency term B. The cataloger can then independently decide whether to include term B as a subject term, and what *its* dependencies should be. If the cataloger in this case decides to include subject term B, and specifies that it depends on term A, then he/she would effectively form a phrase consisting of both terms A and B in no particular order. A cataloger thus has the usual freedom regarding what terms to include as subject terms for a document. In addition, by specifying term dependencies, a document subject term *will not* automatically form a partial match with every query which contains it. Rather, it will form a partial match only with queries which contain that term and also contain that term's dependencies. We illustrate the partial coordination method with the example below:

4.1 A Partial Coordination Example: Dependencies and Scores

Consider a Wall Street Journal article (from the TIPSTER collection, see companion paper (Kambil and Bodoff 1997)) about the government takeover of an oil spill cleanup effort. Suppose we use the post-coordinated subject terms: "government takeover oil spill cleanup". Then, a user submitting the query "government takeover banks", for example, will get a partial match and retrieve this document with a positive score in a post coordinate system. However, the score zero would seem more appropriate in this case as the query term "takeover" ought not to match "takeover" in the document subject term. This is an example of the out-of-context match discussed above, in which an individual term matches out of the context of a topic/sub-topic relationship between the terms "cleanup" and "takeover". Similarly, any query regarding a spill of anything or regarding the oil industry would retrieve this document if each query term is treated independently, unless the user knew to specify that the two terms form the phrase "oil spill". This is an example of two terms which would match out of the context of a phrase.

By partially coordinating the document subject terms "oil" and "spill" to depend on one another; and by partially coordinating the document's subject term "takeover" to depend on the term "cleanup", and the term "government" to depend on both "oil" and "spill", the cataloger can solve the above mentioned problems. With these dependencies, the subject terms "oil" and "spill" are hidden behind one another -- something that is not possible where genuine orderings are required -- so that neither of those document terms will match any query unless the query contains both of those terms. (To simplify the discussion, hereinafter we treat the terms "oil" and "spill" as a single phrase, since the mutual dependency has this result without the user ever having to specify that they form a phrase.) The term "takeover" is hidden behind the term "cleanup". This dependency captures the topic/sub-topic relationship between "cleanup" and "takeover", so that takeover does not match except where the query regards a "cleanup". And the term "government" is hidden behind the phrase "oil spill". This dependency also represents a topic/sub-topic relationship, so that the term "government" does not match out of the context of its dependency -- i.e. oil spill. A user entering the query "government takeovers banks" will **not retrieve** this document at all, as the document terms "government" and "takeover" will not match with respect to this query.

The flexibility of partial coordination can be seen in the nuances of this example. First, the number of coordinated terms is more than could be reflected in the limited number of LCSH subdivisions allowed in an entry. Second, we identified two dependencies as reflecting a topic/sub-topic relationship, yet these are not the sort of topicical subdivisions we would ordinarily expect in LCSH. For example, we would not expect to see a heading such as "Oil Spill--Government" (see section 2.1 including footnote 5 above for an explanation of why the numnber and kind of LCSH sub-divisions is limited.) But with partial coordination, subject terms which would not be expected in a pre-coordinated heading can be included and related to one another.

A partially coordinated subject heading can be expressed by the cataloger using a simple table notation. Table 1 is an example of the subject heading for the example document on government takeovers of oil spill cleanup efforts:

Subject terms for document WSJ990123-1234	Dependency
oil	spill
spill	oil
government	oil, spill, cleanup
takeover	cleanup
cleanup	oil, spill

Table 1

A table such as this is created for each document. The only access points into the document are the terms listed in the first column. These are the document's subject heading terms. Each subject heading term can then be assigned zero, one, or more terms which serve as dependencies to specify the context in which that document term will match a corresponding query term. The document subject term of the first column will match a corrsponding query term only if the query additionally contains that document term's dependencies as specified in the rightmost column. Any term may appear as a subject term only, as a dependency term only, or as both. A subject term cannot depend on itself.

Table 1 shows that the terms "takeover" and "government" each depend on the term "cleanup", while the term "cleanup" depends on the two terms "oil" and "spill". Then, even if a query contains the term "takeover", for example, this document will not match that query term unless the query also contains the term "cleanup". The cataloger has thus indexed the document with a list of terms, and has further specified dependencies among the terms. If a term has two or more dependencies, then it is dependent on all those terms for its context, and matches only if the query contained all those terms. ***Note that the order of terms in a query never matters.*** A dependency requires that the dependent term occur somewhere -- anywhere -- in the query.

Given dependencies, the document scoring for each query is computed from the following rule:

> 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

Table 2 below shows the scores of this document with respect to some of the possible queries containing one or more terms from the document's five subject terms. Unit query and document term weights are assumed.

Query	Traditional Post-Coordination	Partial Coordination
government	1	0
takeover	1	0
cleanup	1	0
oil	1	0
spill	1	0
oil spill	2	2
government takeover	2	0
government cleanup	2	0
takeover cleanup	2	1
government oil spill	3	2
takeover oil spill	3	2
cleanup oil spill	3	3
government takeover cleanup	3	1
government cleanup oil spill	4	4
government takeover cleanup oil spill	5	5

Table 2

Reconsider the query "government takeover banks" to represent an interest in government takeovers of failing banks. The term "government" appears among the document's terms, but its dependent term -- i.e. the phrase "oil spill" -- does not appear in the query. The term "government" is therefore prevented from matching out of context and adds nothing to the score. We go to the next query term, "takeover". This term also appears in the document's index, but again, its dependent term -- i.e. cleanup -- does not, so this query term is also prevented from matching out of context. Lastly, the query term "banks" does not appear in the document index at all. So under partial coordination the score of this document with respect to this query is zero, as it clearly should be.

Dependencies in partial coordination are not transitive. For example if a user interested in a private company's takeover of a toxic waste cleanup program enters the keywords "private sector takeover cleanup toxic waste" the example document receives a score of 1. Two of the query's keywords, "takeover" and "cleanup" appear as subject terms. "Takeover" is considered a match, because both it and its dependent term -- i.e. cleanup -- appear in the query. However, the query term "cleanup" itself is not a match, because although it appears in the document index as a subject term, its dependency term -- i.e. the phrase "oil spill" -- does not. Thus, the term dependencies are not transitive, so while "takeover" depends on "cleanup" and "cleanup" on "oil spill", nevertheless "takeover" does not depend on "oil spill" unless the cataloger expressly adds dependency arrows directly from "takeover" to "oil" and "spill". The reasoning here is that the term "takeover" is not matching out of context, in this cataloger's opinion, as long as both document and query regard takeover of a cleanup effort, and so a point will be added to this document's score for being about the right sort of takeover. But this cleanup is not the same sort of cleanup in the query and document, so the term "cleanup" is not considered to have matched. There are two major benefits to omitting transitivity. First, it helps ensure the scores of *this document with* *respect to various queries* make sense by scoring higher for someone interested in takeovers of a cleanup than for someone interested in another unrelated sort of takeover. Next the *scores of various documents with respect to a given query* also make sense as this document on the takeover of a toxic waste cleanup effort scores higher for this query than documents on takeovers of other kinds.

4.2 Partial Coordination Benefits

Partial coordination increases precision, thereby reducing information overload from spurious partial matches. Most of the thirty-one possible queries using only terms from the document index, result in lower scores under partial coordination compared with post-coordination. For example the query 'government takeover', gives a score of zero under partial coordination, but a score of two under post-coordination. Given a set of index terms, the addition of dependencies can only serve to lower scores. The hope, of course, is that the dependencies will lower scores of relatively less relevant documents, thereby increasing precision. Thus all of these score reductions appear beneficial at first glance.

A possible objection to this supposed improvement is that the cataloger's choice of subject terms 'government takeover' lends itself to out-of-context partial matches. A good cataloger would never include such terms in a post-coordinate environment for fear of such poor matches. This argument actually illuminates an additional benefit of introducing term dependencies. We agree that catalogers choosing keywords for a post-coordinate environment ought to hesitate before including such terms in a document's index. But the document is, after all, primarily about a government takeover of the cleanup effort, not about the cleanup effort itself. Certainly the terms 'government' and 'takeover' ought to be included in the document index to enhance recall, were it not for fear of those terms matching out of context. Partial coordination allows the cataloger to add these terms which may enhance recall, by eliminating the fear that they will decrease precision by matching out of context. Thus, partial coordination should be viewed not only as enhancing precision, but also as enhancing recall by allowing catalogers the freedom to include additional relevant terms without fear of poor precision !

Partial coordination addresses four sorts of out-of-context matches introduced in section 3.0. The meaning of a term dependency is identified with the reason for its use, which is one of the following: an ordered relationship, a phrase, a secondary topic, a narrow term. Partial coordination dependencies effectively substitute for **ordered relationships** such as topic/sub-topic relationships among the index terms, without the user having to guess a correct citation order. This resolves out-of-context matches due to lack of term ordering as in the case of pre-coordination. Partial coordination does not directly address the problem of **polysemy**. However, other sorts of context provided by partial coordination may help ensure that even if an irrelevant document may achieve a non-zero score because of a polysemous word, the document is not likely to achieve a high ranking. Out-of-context matches from a single term **matching out of the context of its phrase** is resolved by establishing term inter-dependencies between the many words of a phrase. Again in this case the user does not have to form specific phrases in his query. Partial coordination also allows the cataloger to include index terms to represent **secondary topics** or other details with greater exhaustivity, without fear that those terms will match out of context. This is simply accomplished by defining dependencies in which the secondary terms depend on the primary ones for their context. Within that context, these secondary terms come into play, to help distinguish the best among the broadly-relevant documents. Finally, partial coordination allows the cataloger to resolve problems caused by non-categorical terms by including **narrow terms** together with the broader terms for increased depth, with the narrow terms depending on the broader terms. For example, if the document is about tarantulas as horror movie props, then the cataloger may have the term "tarantula" depend on a term such as "movie" or "prop"; a different document, in which tarantulas are an example of a poisonous insect, would be cataloged differently, with the term "tarantula" dependent on a term such as "Insect".

We believe the scores in the example of table one, as well as the analysis of out-of-context matches, highlight the potential of partial coordination to shift out the recall/precision curve for post-coordinate keyword indexing. Partial coordination as illustrated above is a powerful way to prevent out of context matches, capturing the ability of pre-coordination to enhance precision without requiring the user to order his query terms (a benefit of post-coordination). Partial coordination also indirectly improves recall. If the fear of matching out of context is removed, catalogers can include more details, i.e. more terms, or narrower terms. Where the query includes both the broader and narrower terms, precision is enhanced by including the narrow term in the document's subject heading. Where the user query included only the narrower term, its inclusion in the document heading enhances recall. Finally, the appropriately lowered scores of irrelevant or partially relevant documents with respect to the various queries will serve to improve the rank ordering of documents.

5.0 Partial Coordination and Other Approaches to Context

Partial coordination differs from pre-coordination and other methods of reducing out-of-context matches. Below we enumerate other methods with some applicability to reducing out of context matches, and indicate the possible advantages of partial coordination over these methods.

5.1 Partial Coordination in Contrast to Pre-Coordination

Partial coordination differs from full pre-coordination in five ways. First, a subject term may be hidden behind another term which is not itself an access point, i.e. it is not itself a subject term. This is accomplished by introducing a dependency term which is not itself a subject term. Second, unlike one linear citation order for all the document terms in pre-coordination, a cataloger can specify a set of individual restrictions on that order. Many complete orderings may satisfy the individual restrictions defined by the cataloger on a document's terms. For example, the cataloger may specify 'A after (i.e. depends on) B, D after E, and no restrictions on B, C, or E' allowing many complete orderings to meet these restrictions.

Third, as described above in section 4.1, partial coordination can reflect any number of arbitrary topic/sub-topic relationships between terms. Pre-coordination can reflect only a limited number and kind of such relationships.

Fourth, like the post-coordinate environment, partial coordination permits partial matches to be recognized, such that the greater the match, the higher the document's score with respect to the query. So if any term, from anywhere in the ordering, matches a query term and fulfills all its dependencies, that term will contribute to a partial match.

Fifth, orderings are replaced by existential dependencies. Thus the actual ordering of query terms never matters, and the user does not even have to specify *partial* orderings. What matters is whether a contextual term appears somewhere in the query or is totally absent from it. There is one circumstance in which existential dependencies are not quite as precise as actual orderings. In the case where two or more terms are related in multiple ways (e.g. Wars due to Crises, Crises due to Wars) and a query includes all the terms involved in both documents' dependency relationships, then an irrelevant document may be retrieved. This is because in our scheme, a query with the terms A B C is equivalent to a query with the terms C B A. The small price paid for not requiring the user to order his query terms, is that the scheme does not distinguish between these two queries.

The differences between traditional pre-coordination and partial coordination, interact to create a powerful yet simple cataloging and retrieval mechanism. Each term may contribute to a partial match with respect to the query as in post-coordinate search, but only if that term is requested in the context of other terms.

5.2 Partial Coordination vs. Extended Boolean

If we are willing to shift the burden of specifying context from the cataloger to the user, extended Boolean queries may also able to provide the basis for an alternative context-sensitive model. In traditional Boolean logic, a user my express 'B depends on A' with the Boolean expression 'A AND B'. The problem with this formulation is that, A also depends on B, contrary to the user's intentions. In pre-coordination terms, this is like giving lead status to *neither* term alone. If the user tries '(A AND B) OR A' to indicate that B is of interest only in the context of A, but that A is of interest in any case, the expression simplifies to just 'A'. This simplification occurs due to the binary nature of Boolean logic.

Some proposals extend the definition of Boolean operators to the non-binary case. In order to achieve a ranking of documents, document or query terms must be assigned weights. The definition of Boolean operators is then extended to the non-binary case (Bookstein 1980, July; Salton et al. 1983, November; Waller and Kraft 1979), or a ranking function which can account for term weights is added to a traditional Boolean retrieval (Radecki 1988). Theoretically, it is conceivable that under some definition of AND and OR, a user could express his contextual dependencies, but it is not clear what definition of AND and OR would allow this. Suppose the user is interested in B only in the context of A, and in A in any case. Then, according to a context-sensitive ranking principle, we would like the following ranking:

for a document with A and B as index terms, rank is highest

for a document with only A as an index term, rank is medium

for a document with only B as an index term, rank is lowest

However, it is impossible to achieve these desired rankings for the most common extended Boolean definition of AND as MIN and OR as MAX, even if query term weights are also supplied. In general, it is an open research question to find definitions of extended Boolean operators which would follow an intuitive context-sensitive ranking principle in arbitrarily complex cases, or how complex these user queries would have to become to achieve the desired rankings. Given the difficulty users have with traditional Boolean queries, it seems unreasonable to expect user to formulate complex extended Boolean queries with query term weights to express context and to result in the intended document rankings. Partial coordination in contrast does not encumber the user in this way.

5.3 Partial Coordination and Term Weighting

Weighting of document and/or query terms may enhance precision. But assigning weights to individual terms cannot be used to adequately represent context. Term weights for individual terms represent a linear scheme, which alone cannot achieve the non-linearity of dependencies. Term weights would have to be specified to every pair, triple, etc., of terms, either for each query or for each document in order to achieve the notion of context such that the document score depends on the particular combination of document terms matching the query terms. Thus single term weights alone cannot be used to establish the sort of context achievable with partial coordination.

5.4 Partial Coordination and Full Text Indexing

In recent years full text indexing has dominated information retrieval research. As discussed earlier in this paper, the problem of false drops seems to be greater in the case of full text. But few studies have directly compared full-text to non full-text indexing, and we know of no study which experimentally compares the effectiveness of state of the art full-text indexing and retrieval with manually assigned controlled vocabulary subject headings.

Full text does allow for new techniques which enhance recall and/or precision. These techniques are the focus of much research attention (for example in the TREC conferences). Two methods applicable with full-text indexing that are most closely related to the notion of context, are term phrases and the probabilistic term dependence models which are discussed below[13].

In any case, despite the potential of full-text retrieval, given large paper document collections, no libraries currently provide it, nor are they likely to provide it in the foreseeable future for their paper collections. Partial coordination in contrast is a feasible enhancement to current OPAC's, as it does not require full text availability.

5.5 Partial Coordination and Term Phrases

The extensive research into term phrases is closely related to our proposal[14] (see (Fagan 1987) for a review of phrases in IR). Term phrase construction allows the retrieval engine to account not only for individual terms, but for particular combinations of terms. Experimental results indicate a modest improvement in retrieval results when phrases are manually identified in the query (Croft et al. 1991), but this poses an extra burden on the user. Results

[13] Theoretically these methods are applicable with free text indexing even without full text indexing, but the statistical estimation of parameters requires full text. An exception is the approach taken in (Croft 1986) which is discussed below.

[14] The literature on term phrases is extensive. We refer solely to Fagan's thesis, which reviews much of that literature.

were not encouraging for automatic phrase identification without manual help (a modestly successful experiment conducted with automatic phrase formation relied again on manually formulated query phrases to create a phrase dictionary). Fagan (Fagan 1987) reviews the difficulties of automatic phrase formation. The statistical approaches use co-occurence data as a very imperfect surrogate for identifying phrases (Fagan 1987; Peat and Willett 1991). The syntactic approaches which can theoretically help avoid improper phrases, have generally been even less successful than the statistical ones(Fagan 1987). A second and much more important reason why partial coordination is expected to enhance retrieval beyond the success of phrases, is that term phrase construction is only one special use of partial coordination. Partial coordination can be used to represent arbitrary syntactic and semantic relationships as enumerated above in section three. Indeed Croft concludes "...in Boolean queries, experts often form the AND of two concepts which are not phrase components. This implies a strong relationship between those concepts,...but it is not clear what type of relationship" (Croft et al. 1991). In section 3 of this work we clarified what types of relationship may induce experts to AND two together two terms. We identified a set of contextual relationships which require two or more terms to be treated together. Partial coordination was proposed in section 4.2 as a means of expressing all these contextual relationships. Thus, phrase formation is just one of many uses of partial coordination, which can also express the other relationships hinted at in (Croft et al. 1991) and identified in this work. Moreover, with partial coordination the burden is shifted from the user query to the one-effort of the cataloger.

5.6 Partial Coordination and Generalized Probabilistic Model

The most relevant work to the proposal put forward in this paper is the generalized probabilistic model (Maron 1988; Maron and Kuhns 1960; Rijsbergen 1977; Robertson and Jones 1976; Robertson et al. 1982) which is possible in full-text and non-full-text environments. In this model the probability of relevance to a query is estimated on the basis of the particular *combination* of index terms in each document. The extent to which each document index term indicates probabilistic relevance to a query, is considered to depend on the presence or absence in the document index of every other term in the vocabulary or at least in the query. Thus, for example, joint probability estimates are used to assess the relevance of a document with index terms A and B, separate from the estimation of relevance for the documents indexed with only one or the other term. This approach can theoretically incorporate all notions of context, not only term phrases. although it does so implicitly. In this sense, the probabilistic model is most closely related to our partial coordination. However, this approach theoretically requires relevance data to estimate parameters of the relevant and irrelevant documents separately. Even with feedback data, this general model has been considered impractical because of the exponential number of parameters involved (Salton 1989). The tree term dependence model is a limited version of this model which strictly limits the number of parameters, (Rijsbergen 1977; Salton et al. 1982). In this model, each term is considered to depend on at most one other term in the vocabulary or query. This approach is not only feasible, but has been shown to considerably improve precision for given levels of recall (Harper and Rijsbergen 1978; Salton et al. 1982). Nevertheless this approach still (ideally) requires feedback data for parameter estimation. The favorable experimental results in (Harper and Rijsbergen 1978; Salton et al. 1982) utilized feedback data.

In the absence of feedback data, Croft suggests heuristics for estimating a subset of these parameters (Croft 1986). However, these heuristics require Boolean queries, and were only very marginally effective. Only manual phrase construction by the users was again shown to supply enough information for real improvement in retrieval effectiveness.

In contrast to these approaches, partial coordination aims to increase precision even in the first iteration of search -- prior to feedback -- so users are not initially discouraged by a large number of (false) hits. Our approach also relieves the user of any responsibility for phrase formation.

5.7 Partial Coordination and "Mixed Approaches"

Our literature review identified only one attempt to *directly* combine the strengths of pre- and post-coordination within the traditional non-full-text environment. Gary Lawrence reviews the limitations of both pre- and post-coordination, and follows with a brief section entitled Mixed Approaches (Lawrence 1985). One approach, attributed to Mischo (Mischo 1981), is to "selectively manipulate subject headings and titles to present important words and word pairs at the beginning of the index entry in a heading-based (i.e. pre-coordinated) retrieval system" (Lawrence 1985). In other words, he suggests retaining the basic pre-coordinated environment, but including index entries for many possible citation orderings, depending on "the contents of defined subfields" (Mischo 1981).[15] In this way, the precision of

[15] Lawrence cites Mischo to whom this idea is attributed. In all of Mischo's relevant writings ((Mischo 1979; Mischo 1980; Mischo 1981)), however, we have not found the suggestion that the particular contents of either the document itself or of the 'defined subfields' be used to determine whether to include a particular ordering. In

pre-coordination is not lost, while not requiring that the user to guess the one correct citation order.

While cross-references already provide indirect access through multiple citation orderings, Lawrence suggests providing actual index entries instead of just cross-references. But it is unclear how rotations are selected. And his method of selectively-rotated subject headings does not include the ability of post-coordination to provide partial matches. His proposal is appropriate for addressing the well-known problem of which and how many ordering permutations to include in a card catalog.

5.8 PRECIS

Related to Lawrence's ideas on intelligent rotations, is PRECIS, a cataloging system developed for the paper card catalog environment. While PRECIS is concerned with selectively-rotated entries, and also includes a notion of context, it nevertheless offers no help in avoiding out-of-context partial matches in the online environment.

PRECIS's primary goals were (Dykstra 1987):

1. to allow the cataloger to enter one encoded string to represent each document; the string includes document subject terms and PRECIS codes
2. to automatically produce catalog entries for a number of possible rotations of the document terms. A 'heading' is the access point for each

Mischo's proposal, the determination of which orderings to include is based exclusively on structural features of the subject heading, not its content.

automatically produced entry. A heading contains one or more of the document's subject terms

3. to ensure that entries are produced only for headings which "make sense", that all headings which make sense have an entry, and to ensure that the presentation of each entry makes clear the meaning of the whole subject heading assigned to the document.

A concept of "context dependency" was developed to allow the PRECIS system to produce any and only meaningful entries, and to ensure that the presentation of each entry makes clear the meaning of the whole subject heading, "This principle requires that...each term in the entry is related to the one immediately preceding and the one immediately following it. Each term sets the next term into its obvious context." PRECIS codes, which represent a limited number of facets (such as form, location, etc.) are embedded by the cataloger into each subject heading, so that each term is explicitly assigned to one role. Using this role information, and following the principle of context dependency, the PRECIS algorithm determines which catalog entries to generate, and how each should be presented. The meaning of the whole heading is made clear through adherence to the principle of context dependency.

The importance of automatically creating rotated subject headings with transparent meaning is particular to the mechanics of pre-coordinate search. On the one hand, contrary to post-coordinate search, additional access points are necessary through the use of rotations; on the other hand, the process of pre-coordinate search is a two-step process as described in (Drabenstott and Vizine-Goetz 1994), the first step of which requires the user to understand and identify the intended subject heading. In this way, the philospohy of PRECIS assumes a pre-coordinated environment, and facilitates selection of meaningful rotations.

PRECIS requires catalogers to specify the role played by each term in a heading. This explicit role information *may* be used in an algorithm to help avoid bad partial post-coordinate matches, but we know of no such algorithm, and PRECIS itself does not include such an algorithm. We know of no discussion regarding the application of PRECIS in a post-coordinate environment to limit bad partial matches. Indeed, there exists an online version of PRECIS called COMPASS (Wilson 1991), and in this system, the role operators have been eliminated in favor of traditional post-coordinate term matching (Trotter 1996)[16]. Why not use all the valuable PRECIS role information to limit bad partial matches in the post-coordinated version COMPASS? Presumably, an algorithm to do so has not yet been developed, and promises to be highly complex. We agree with earlier reviews of this work, that the PRECIS role operators would seem to provide valuable information which might be somehow used to prevent false drops. This possibility has not been addressed by the developers of PRECIS or COMPASS. We, too, leave this for future work, and view the utilization of PRECIS codes as a possible means of formalizing the process by which catalogers specify term dependencies. As PRECIS and COMPASS are being phased out of use, such future work ought to focus more generally on exploiting available role information in whatever form it takes. In summary, PRECIS as it has been

[16] In the words of its chief architect, when COMPASS was developed "we did away with the role operators" and allowed traditional post-coordinate term matching (Trotter 1996).

developed to date does not help solve the problem of spurious partial matches in an online post-coordinate retrieval environment.

5.9 Feasibility

Partial coordination is proposed as an extension to non full-text OPAC's, as well as to full-text available databases such as the WWW. In the case of OPAC's which rely on LC or other pre-coordinated subject headings, the additional manual effort required to provide partially coordinated keywords, seems very small compared to the effort already required to produce the LC heading. Moreover, catalogers are already asked to include additional non-controlled subject terms (MARC 653 field). The incremental effort required to instead include partially coordinated terms is small.

If an OPAC were extended to rank documents using partially coordinated subject terms, existing MARC records with subject headings would not require revision, because a document with no term dependencies is just a special case of a partially coordinated heading. The ordering of terms in the LC heading would just be ignored as in traditional keyword search. One issue which does require resolution, however, is that the rank order of documents for a query would need to fairly integrate the scores of partially coordinated documents with those unrevisied records which lacked partial coordination.

In the full text environment, the additional effort required to assign partially coordinated subject terms is great, since the full text indexing requires no manual effort at all. We envision our proposal as relating to full text indexing in two ways: First, automatic techniques may be pursued which will automatically create some term dependencies. Second, the labor of adding additional subject terms with partial coordination may be distributed among the documents' readers. We are currently working on both these proposals.

6.0 Conclusions

The partial coordination proposal in this paper combines the advantages of intelligent pre-coordination -- i.e. greater precision -- with the chief advantages of post-coordination -- i.e. the user is relieved of the burden of learning cataloging rules such as citation order, and partial matches are supported. The technique we propose concentrates on the benefits of *intelligent* ordering, which can be realized in an OPAC environment without any additional effort on the part of the user; the user enters his keywords as before, but the intelligent coordination of documents' keywords prevents inappropriate partial matches. Our critical analysis of out-of-context matches and review of partial coordination in the context of alternate techniques suggests that this is a promising technique for improving precision and recall in OPAC and other emerging information retrieval contexts such as the WWW.

The follow-on paper to this introductory article provides a preliminary evaluation of this technique.

References

Bates, M. J. (1977). "Factors Affecting Subject Catalog Search Success." *Journal of the American Society for Information Science*, *28*, 161-169.

Bates, M. J. (1986). "Subject Access in Online Catalogs: A Design Model." *Journal of the American Society for Information Science*, *37*, 357-376.

Beaulieu, M., and Borgman, C. L. (1996). "A New Era for OPAC Research: Introduction to Special Topic Issue on Current Research in Online Public Access Systems." *Journal of the American Society for Information Science*, *47*(7), 491-492.

Besant, L. (1982). "Early Survey Findings: Users of Public Online Catalogs Want Sophisticated Subject Access." *American Libraries*, *13*(3).

Blair, D. C., and Maron, M. E. (1985). "An Evaluation of Retrieval Effectiveness For a Full-Text Document-Retrieval System." *Communications of the ACM*, *28*(3), 289-299.

Blair, D. C., and Maron, M. E. (1990). "Full-Text Information Retrieval: Further Analysis and Clarification." *Information Processing and Managament*, *26*(3), 437-447.

Bookstein, A. (1980, July). "Fuzzy Requests: An approach to Weighted Boolean Searches." *Journal of the American Society for Information Science*, *31*(4), 240-247.

Borgman, C. L. (1984). "The User's Mental Model of an Information Retrieval System: Effects on Performance," PhD Thesis, Stanford University.

Borgman, C. L. (1996). "Why Are Online Catalogs Still Hard to Use?" *Journal of the American Society for Information Science*, *47*(6), 493-503.

Cochrane, P. A. (1983). "A Paradigm Shift in Library Science." *Information Technology and Libraries*, *2*(1), 3-4.

Cochrane, P. A. (1985). *Redesign of Catalogs and Indexes for Improved Online Subject Access*, Oryx Press, Phoenix, Arizona.

Croft, W. B. (1986). "Boolean Queries and Term Dependencies in Probabilistic Retrieval Models." *Journal of the American Society for Information Science*, *37*(2), 71-77.

Croft, W. B., and Thompson, R. H. (1986). "IR: A New Approach to the Design of Document Retrieval Systems." *Journal of the American Society for Information Science*, *37*(6), 389-404.

Croft, W. B., Turtle, H., and Lewis, D. (1991)."The Use of Phrases and Structured Queries in Information Retrieval." *Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval*, (32-45).Chicago, IL.,

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.

Dykstra, M. (1987). *PRECIS: A Primer*, The Scarecrow Press, Metuchen, New Jersey.

Fagan, J. L. (1987). "Experiments in Automatic Phrase Indexing For Document Retrievel: A Comparison of Syntactic and Non-Syntactic Methods," Ph.D. Thesis, Cornell University.

Foskett, A. C. (1977). *The Subject Approach to Information*, Clive Bingley, London.

Fuggman, R. (1985). The Complementarity of Natural and Indexing Languages. L. M. Chan, P. Richmond, and E. Svenonius, eds., *Theory of Subject Analysis: A Sourcebook*, (392-402), Littleton, Colorado, Libraries Unlimited, Inc.

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

Harper, D. J., and Rijsbergen, C. J. V. (1978). "An Evaluation of Feedback In Document Retrieval Using Co-Occurrence Data." *Journal of Documentation*, *Vol. 34*(Number 3), 189-216.

Hildreth, C. R. (1989). "The Online Catalog." , Library Association Publishing Ltd, London, 212.

Kambil, A., and Bodoff, D. (1997). "Partial Coordination: A Preliminary Evaluation and Failure Analysis." Center for Research in Information Systems Working Paper *IS-97-15*, Stern School of Business, New York University.

Larson, R. R. (1989)."Managing Information Overload in Online Catalog Subject Searching." *Proceedings of 52nd ASIS Annual Meeting*, (129-135).Washington, D.C.,

Larson, R. R. (1991). "The Decline of Subject Searching: Long-Term Trends and Patterns of Index Use in an Online Catalog." *Journal of the American Society for Information Science*, *42*(3), 197-215.

Larson, R. R., and Graham, V. (1983, March). "Monitoring and Evaluating MELVYL." *Information Technology and Libraries*, *2*(1), 93-104.

Larson, R. R., McDonough, J., O'Leary, P., and Kuntz, L. (1996). "Cheshire II: Designing a Next-Generation Online Catalog." *Journal of the American Society for Information Science*, *47*(7), 555-567.

Lawrence, G. S. (1985). "System Features for Subject Access in the Online Catalog." *Library Resources & Technical Services*, *29*(1), 16-33.

Lipetz, B.-A., and Paulson, P. J. (1987). "A Study of the Impact of Introducing an Online Subject Catalog at the New York State Library." *Library Trends*, *35*, 597-617.

Lynch. (1989). Large Database and Multiple Database Problems in Online Catalogs. *OPAC's and Beyond: Proceedings of a Joint Meeting of the British Library, DBMIST, and OCLC*, (51-56), Dublin, Ohio, OCLC Online Computer Library Center, Inc.

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

Markey, K. (1980). "Analytical Review of Catalog Use Studies." *ED 186 041*, OCLC.

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.

Maron, M. E. (1988). "Probabilistic Design Principles for Conventional and Full-Text Retrieval Systems." *Information Processing and Management*, *24*(3), 249-256.

Maron, M. E., and Kuhns, J. L. (1960). "On Relevance, Probabilistic Indexing and Information Retrieval." *Journal of the ACM*, *7*, 216-244.

Matthews, J. R., and Lawrence, G. S. (1984). "Further Analysis of the CLR Online Catalog Project." *Information Technology and Libraries*, *3*, 354-376.

Mischo, W. H. (1979). "Expanded Subject Access to Reference Collection Materials." *Journal of Library Automation*, *12*(4), 338-354.

Mischo, W. H. (1980)."Expanded Subject Access to Library Collections Using Computer-Assisted Indexing Techniques." *Proceedings of 43rd ASIS Annual Meeting*, (155-157).Anaheim, Calilfornia,

Mischo, W. H. (1981). "Technical Report on a Subject Retrieval for the Online Union Catalog." *OCLC/DD/TR-81/4*, OCLC.

Peat, H. J., and Willett, P. (1991). "The Limitations of Term Co-Occurence Data for Query Expansion in Document Retrieval Systems." *Journal of the American Society for Information Science*, *42*(5), 378-383.

Peters, T. A. (1991). *The Online Catalog: A Critical Examination of Public Use*, McFarland & Co., Jefferson, N.C.

Prabha, C. (1989). Managing Large Retrievals: A Problem of the 1990s? *OPAC's and Beyond: Proceedings of a Joint Meeting of the British Library, DBMIST, and OCLC*, (33-38), Dublin, Ohio, OCLC Online Computer Library Center, Inc.

Radecki, T. (1988). "Probabilistic Methods for Ranking Outout Documents in Conventional Boolean Retrieval Systems." *Information Processing and Management*, *24*(3), 281-302.

Rijsbergen, C. J. V. (1977). "A Theoretical Basis for the Use of Co-Occurrence Data in Information Retrieval." *Journal of Documentation*, *Vol. 33*(No.2), 106-119.

Robertson, S. E. (1997). "Overview of the Okapi Projects." *Journal of Documentation*, *53*(1), 3-7.

Robertson, S. E., and Jones, K. S. (1976). "Relevance Weighting of Search Terms." *Journal of the American Society for Information Science*, *27*(3), 129-146.

Robertson, S. E., Maron, M. E., and Cooper, W. S. (1982). "Probability of Relevance: A Unification of Two Competing Models For Document Retrieval." *Information Technology and Libraries*, 1-21.

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

Salton, G., Buckley, C., and Yu, C. - T. (1982). An Evaluation of Term Dependence Models in Information Retrieval. G. Salton and H.-J. Schneider, eds., *Lecture Notes in Computer Science*, (151-173), Berlin, Springer-Verlag.

Salton, G., Fox, E. A., and Wu, H. (1983, November). "Extended Boolean Information Retrieval." *Communications of the ACM, Vol. 26*(Number 11), 1021-1036.

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.,

Steinberg, D., and Metz, P. (1984). "User Response to and Knowledge about an Online Catalog." *College & Research Libraries*(January 1984), 66-70.

Stoll, C. (1995). *Silicon Snake Oil: Second thoughts on the information highway*, Doubleday, New York.

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.

Trotter, R. (1996). Personal Communication. FAX June 13, 1996

Waller, W. G., and Kraft, D. H. (1979). "A Mathematical Model of a Weighted Boolean Retrieval System." *Information Processing and Management*, *15*(5), 235-245.

Wilson, N. (1991). "COMPASS: News from the Front." *SELECT*, *4*.