PRE-COORDINATION + POST-COORDINATION = PARTIAL COORDINATION

David Bodoff
Department of Information Systems
New York University
Leonard N. Stern School of Business
44 West 4th Street, Suite 9-181
New York, NY 10012-1126
phone: (212) 998-0800
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) 998-0800
fax: (212) 995-4228
akambil@stern.nyu.edu

February 1996

Working Paper Series
Stern #IS-96-1

Pre-Coordination + Post-Coordination = Partial Coordination

David Bodoff and Ajit Kambil

Information Systems Department

Leonard N Stern School of Business

New York University

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.

1.0 Introduction

In the era of card catalogs, subject searches were not as frequent as known item searches (Larson 1991; Markey 1980; Markey 1984), and received a limited amount of practical and academic attention (Bates 1986, November; Cochrane 1983, March). The introduction of on-line catalogs has renewed users' interest in subject searches (Lipetz and Paulson 1987, Spring; Matthews and Lawrence 1984, December). However, while users now turn to subject searches more than any other search type (Markey 1985, January/March), they also complain most about the limitations of subject searches, and ask for improvements in subject access (Larson and Graham 1983, March; Markey 1984 page 84; Matthews and Lawrence 1984, December). The most frequently cited limitation is the frequency of search failures, i.e. subject queries which return zero hits (Markey 1988, September). Other studies have found a very high average number of hits per query (Markey 1984 page 67) Taken together, these results imply that users are either getting no useful results or are overloaded with results in response to a given query. Evaluating complete user sessions, rather than a given query, the Council on Library Resources survey identified that only forty five percent of users found some of, more than, or all of what they were looking for (Bates 1986, November) The users' self-reported success in 'finding what they were looking for' may be greatly inflated due to the particular features of subject heading rules (especially LCSH's rule of specific entry). Success rates were much lower when this latter factor was removed from the equation (Bates 1986, November).

In response to these limitations, researchers have proposed a series of rather elaborate OPAC designs, containing, among other things, on-line thesauri, syndetic structures, class schedules, and document clustering (Bates 1986, November; Cochrane 1985; Hildreth 1989; Larson 1989; Markey 1984; Markey 1988, September; Peters 1991). It is almost impossible, however, to find reports of the success of such projects in actual implementation. One problem would seem to be the extra burden placed upon patrons to use these advanced system features.

The problem of information overload is only going to get worse as publishing costs fall. Inexpensive online and Internet publishing is rapidly expanding the available number and types of documents. Soon, the whole public Internet will become searchable with the submission of one query to one service. 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, thereby exacerbating the problem of information overload. To make matters worse, the computerphilic culture of the Internet has heavily favored full text search over keyword search. 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, March; Blair and Maron 1990; Sievert and McKinin 1989; Svenonius 1986, September; Tenopir 1985)[2]. Furthermore, 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 size of result sets will make infeasible the practical use of full-text search of the whole Internet, even if other technical obstacles are

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

overcome.

Libraries should be especially attentive to these developments. The information explosion and widespread networked computing is daily diminishing the role of libraries as central repositories of information. To remain relevant in this new environment, the library establishment is investing in projects such as INTERCAT to catalog Internet resources. In this model, libraries increase their relevance by uniquely providing subject access to all resources, real and virtual. The advantage libraries have is not in owning any special search capabilities, or even in the superiority of expert keyword indexes (i.e. subject headings) over full text access. Rather, the advantage libraries have is the potential for seamless, one-stop access to all resources, whether real or virtual.

In this environment, the library as an institution depends on the effectiveness of OPAC's in its effort to remain an important information provider in the information age. Improvements in OPAC technology can reduce users' dissatisfaction. Because of the libraries' advantage in providing comprehensive access to both paper and electronic documents, any advance in subject access is a relative gain for libraries over other potential subject access providers.

In this paper we review the strengths and weaknesses of the two methods for subject access to documents -- pre-coordination, as in traditional card catalogs, and post-coordination, as in computerized keyword search. We propose a new method which we call partial coordination, to combine the strengths of pre- and post-coordination. This method should result in greater precision for a

given level of recall, just the sort of improvement which can help libraries stay relevant.

Prior research into document retrieval effectiveness is unfortunately split into two sorts. The first sort, preferred by library scientists, studies OPAC performance in terms of system features, their use, and patrons' self-reported success or failure. The other approach, taken by computer science information retrieval researchers, studies retrieval techniques rather than system features, and relies on estimates of recall and precision to test the worthiness of a given technique. This paper adopts the latter information-retrieval approach because it introduces a new technique, which is most easily presented in isolation from (other) system features.

The proposal in this paper can be viewed by library scientists as a combination of the strengths of pre- and post-coordination, and by information retrieval scientists as an extension to the Vector Space Model (Salton 1989). Both views are presented here.

This paper is organized as follows. The historical progression from pre- to post-coordination is briefly reviewed in section two. Section three describes the pros and cons of each. Section four reviews previous attempts to achieve the benefits of both methods of coordination. Sections five and six present an example use of partial coordination as a method to combine those strengths. Section seven is a more formal view of the proposal. Section eight further discusses the strengths of partial coordination in light of the example. Section nine analyzes the mathematical property which facilitates the strengths of partial coordination. Section ten argues that keyword selection is positively

impacted by the method of coordination. Section eleven discusses the practical feasibility of partial coordination in OPAC's and for the Internet. A summary concludes the paper and points toward the next step in this research.

2.0 Background

Historically, library science was concerned with providing good subject access to documents using pre-coordination of subject terms. Among the obstacles to successfully finding relevant documents in a card catalog were the problems of choosing:

- the right terms from the many possible synonyms which refer to a given concept, and
- choosing the correct ordering from the many terms in a compound subject (i.e. a subject composed of more than one term). For example, knowing to look under Business -- Chinese -- Ancient in an alphabetical catalog to find relevant documents. This ordering is known as the citation order. Fixing the citation order for each subject heading is known as pre-coordination.

With the introduction of computers and on-line catalogs (OPAC's), the latter obstacle had been removed. The computer would find documents labeled with the subject heading Business--Chinese--Ancient, even if these words were input as query terms in the 'wrong order' -- e.g. a query 'Ancient Chinese Business'. Allowing the terms of a compound subject heading to be effectively re-ordered to match any query is known as post-coordination. The benefits of removing this age-old barrier to effective retrieval was largely unquestioned.

This paper proposes partial coordination which can incorporate the benefits

of pre-coordination into today's OPAC's. It is argued that partial coordination of subject headings will push out the recall-precision curve and allow more effective retrieval.

3 Pre- versus Post-Coordination Reviewed

This section reviews pre- and post- coordination.

3.1 Benefits of Pre-Coordination

The benefits of pre-coordination arise from the standardization of term orderings, and from the intelligence of the particular choice of orderings. The mechanisms by which each of these aspects enhances recall and precision are discussed in turn.

3.1.1 Standardization of Order

A specific citation order for each subject in a traditional pre-coordinated system was specified "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). The same idea can be expressed with many syntaxes using the same terms in a natural language. Pre-coordinating the terms eliminates this variability. The purpose of a citation order can be viewed as enhancing recall. It works both syntactically and semantically. Syntactically, it standardizes the translation from syntactical relationships among terms, such as adjective-noun (e.g. Chinese Business), to an ordered list of those terms; semantically, it standardizes the translation from a semantic

relationship among terms -- i.e. an idea -- into an ordered list of terms. It does both of these, so that the consistency allows us to find documents on a given topic, regardless of the variety of our natural language. However, today computers and post-coordination makes it certain that a topic can be repeatedly found regardless of whether the cataloger, user, various users -- or the same user on different days -- uses a different citation order for that topic.

Discussions of pre-coordinate systems such as LCSH and post-coordination, underemphasize the flip side to the above notion that one idea can be expressed in many syntaxes using the same words. The flip side is that the same words can also be formed in many syntaxes to represent *different* ideas. For example (after Foskett), 'Wars (due to) Economic Crises' is different from 'Economic Crises (due to) Wars'; these terms are related in each case by a cause-effect relationship. Or for another example, 'The case for free trade in a world of multinational corporations' is different from 'The advantages of multinationals in a world of free trade'. Thus the terms in this example relate in each case by a 'position-argued-for and argument' relationship.

When two terms are related in such a semantic relationship, pre-coordination may standardize an ordering of concepts. For example, where the terms are related by a cause-effect relationship, pre-coordination may standardize that the 'effect' term must precede the 'cause' term. This standardization allows 'Wars-Economics' to have an entirely different meaning from 'Economics-Wars', and to be found in distant parts of the card catalog, without confusion. This effect of pre-coordination can be viewed as enhancing precision. However, in post-coordinate systems, where the (unordered) list of query terms 'war, economics' matches documents on both subjects, this positive

effect on precision is totally lost.

For every relationship among terms -- syntactic or semantic -- a cataloging formalism can standardize an ordering of terms, thus enhancing precision (along with recall). Various precoordinate schemes are variously ambitious in the number and kind of relationships they recognize and for which they standardize a citation order among the terms. The most ambitious of these schemes is PRECIS (Dykstra 1987). Since ordering of terms alone is insufficient to uniquely encode all possible compound subjects, PRECIS introduces a richer cataloging formalism in which an index entry is highly structured and expressive.

3.1.2 Intelligence of Order

Pre-coordination enhances recall and precision not only by *arbitrarily standardizing* an order of terms to capture syntactic and semantic relationships, but by selecting an *intelligent* order which buries some keyword terms within the context of others. Take, for example, the document on the case for free trade in a world of multinationals. This document would be much more intelligently grouped with other books on free trade than with other books on multinational corporations. This reasonable grouping is achieved by ordering the Free Trade term before the Multinational term i.e. by adopting the general rule that for all documents, the 'position-argued-for' term precedes the 'argument' term. This intelligent decision enhances recall by making it easier for the patron to correctly guess the citation order.

The intelligence of the choice also enhances precision. A thorough discussion of this is rather complex, so we will only briefly mention one precision-

enhancing mechanism, and focus on another, more relevant to this paper. An intelligent ordering enhances 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 mechanism through which intelligent ordering enhances precision is by avoiding improper *partial* matches. This is accomplished through the power of *context.* A partial match in the case of pre-coordination means the user has guessed the first terms of a subject heading in the correct order, but he has not guessed all the terms; he has stopped without completing the subject heading, and 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.

Take, for example, the book which argues for free trade in the era of multinationals. Suppose this book and others like it were cataloged under the subject heading 'Multinationals--Free Trade[3]'. Then a user interested in this argument for free trade, who partially matched by looking under 'Multinationals' would find himself browsing through numerous documents and subdivisions related to multinational corporations and totally unrelated to

[3] Anyone familiar with pre-coordinated subject headings will know that this is a very unlikely subject heading, as it consists of specific keywords rather than more categorical terms. The example is used because it is a natural list of keywords in a post-coordinated scheme. This difference between the nature of the keywords themselves, when comparing alternative types of coordination, is a crucial point which is discussed in section ten below.

his area of interest. In this sense, he would have partially matched the term 'Multinationals' out of context. It is better for precision if the term Multinationals is only found as a subdivision of the broader 'Free Trade' term. In that case, the Multinational term would be matched only in the context of free trade, and an inappropriate partial match would not be possible.

Pre-coordinate schemes use two mechanisms to establish the context for a term. Citation order is one mechanism, term phrasing is another. The LCSH heading 'Art, Asian' uses a term phrase to ensure that the term Asian matches only in the context of Art. The cataloger or the cataloging scheme must determine whether to form a term phrase or to coordinate two separate terms. Either mechanism can prevent spurious partial matches.

3.2 Limitations of Pre-Coordination

Traditional pre-coordination has a number of problems. First, while some relationships among terms can be represented with a standardized ordering or with more elaborate formalisms such as PRECIS, there is a limit to the complexity we can introduce into a cataloging formalism, and there will always be occasions when two different topics would be identically represented. In addition, ordinary users cannot be expected to properly encode their queries according to such elaborate rules. Precoordination also imposes the requirement on users to guess the appropriate citation order of the appropriate controlled-vocabulary terms and term phrases. Even in a relatively simple scheme such as LCSH, this is difficult for most users (Bates 1977, May; Markey 1984 pages 65-66). The historical result is that subject searches were less common than known-item searches in the era of

precoordinated card catalogs (see (Larson 1991), contrary to (Markey 1984)). One study, for example, found that only 28.2% of the users even knew the subject heading needed to be from LCSH (Steinberg and Metz 1984). Most users are clearly not expert in the particulars of the LCSH rules. Users cannot be expected to learn sophisticated cataloging rules.

3.3 Post-Coordination as a Mixed Blessing

Post-coordination alleviated the user from the requirement to learn any formalisms, including rules governing citation order. However, the benefits of standard, intelligent orderings were thereby lost.

The most significant loss which accompanied post-coordination, is the loss of *precision* which resulted from *intelligent* orderings. That gain in precision was helpful in the card catalog case by preventing inappropriate partial matches as described, but is much more important in the post-coordinate case. This is because partial matches are a much bigger concern in the post-coordinate case. In the pre-coordinate case, 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 post-coordinate systems, any document is retrieved if it matches any term in the uncoordinated list of keywords. Indeed, the main cause of information overload in the case of post-coordinate search is the large number of inappropriate partial matches

To take the example above, suppose a user is interested in the free trade debate

and how the power and prevalence of multinationals informs that debate. He would like to specify the two keywords Free Trade and Multinationals. To this user, a book on multinationals which is not about free trade is of little interest, while a book about the free trade debate which is not about multinationals is fairly relevant. But the user's query vector (Free Trade, Multinationals) will retrieve documents of both sorts with an equal partial match[4]. The user really intends to be interested in multinationals only in the context of free trade, but in free trade in any case. This is not possible in post-coordination, where every term has 'lead status' and can match a query term.

We have seen that pre-coordination enhances precision by specifying both *some* order and an *intelligent* order on terms. Post coordination does not recognize either of these benefits, even though the problem of inappropriate partial matches is actually worse in the post-coordinate case.

4 Prior Research on Improving Pre- and Post-coordination

Various efforts have been made to improve on either pre- or post-coordination. The most widespread improvement to post-coordinate search is the use of boolean operators. However, boolean operators are often not used or are ineffectively used by information seekers (Cooper 1988; Fox and Koll 1988). In a traditional boolean environment it also is impossible to rank documents according to the extent of a match. For such a ranking, the document keywords or the query terms must be assigned weights. The definition of boolean operators must then be extended to the non-binary case as in (Bookstein 1980, July; Salton et al. 1983, November; Waller and Kraft 1979), or a ranking

[4] Term weighting is of little benefit here, as discussed in section four.

function which can account for weights must be added to a traditional boolean retrieval (Radecki 1988).

Weighting of document and/or query terms may enhance precision. But term weights cannot afford the benefits of pre coordination. If a document term has a high weight, then it has a heigh weight w.r.t. any query; if a query term has a hight weight, then it has this high weight w.r.t. every document. If weights are given to individual terms, and if the scoring function is additive across terms, then term weights cannot recognize good or bad *combinations* of terms -- i.e. context. We return in section nine below to the question of desired mathematical properties for scoring functions.

It is possible for an extended boolean approach to realize some of the benefits of pre coordination. For example, suppose AND were defined as SUM, and OR as MAX, and all term weights are equal to one. Then the expression (A AND B) OR B can be used to indicate that A is of interest only in the presence of B (and the score with respect to this query of a document labeled with only A is zero, while if it is labeled with both A and B the score will be 2), but B is of interest in any case (and the score of a document labeled only with B would be 1). This is roughly analogous to hiding the term A within the context of the term B. It is unclear, whether sophisticated use of extended boolean operators can both express arbitrarily complex patterns of desired keywords and their contexts, as well as intelligently rank search results. We return to this point in section nine, after demonstrating the power of partial coordination.

Most information retrieval research in recent years has focused on full-text indexed documents. Full text search creates its own problems. Empirical

research results are mixed when comparing recall and precision results of full-text versus assigned-keywords (Svenonius 1986, September). In fact, the problem of false drops seems to be greater in the case of full text, as discussed above. Achieving good recall is also difficult. Negotiating the nuances of natural language places its own burdens on the user (Lancaster 1986).

Full text does allow for new techniques which enhance recall and/or precision, such as the proximity operator. This operator allows one to specify that two terms must occur within a specified number of words (or sentences, paragraphs). The problems here are similar to the limitations of boolean operators. First, the burden is traditionally placed on the user, who is given no guidance -- practical or theoretical -- to inform his choice of operator (should I require that the two terms be adjacent, within five words or within three paragraphs?). A more promising use of the proximity operator relieves the user of the need to specify the exact required proximity. Instead, the user merely specifies A near B, and the system ranks documents according to the *extent* of (the presence and) nearness of A and B. But the other problems remain: Specifying that A is of interest in the near context of B, but B is of interest in any case, raises the same difficulties posed just above for the extended-boolean operators.

In any case, whatever promises full-text retrieval may hold for the future, no libraries currently provide it, and none are likely to provide it in the foreseeable future, at least not for their paper document collection. This paper, on the other hand, aims to propose a practical enhancement to current OPAC's.

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, June; Robertson and Jones 1976, May-June) 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, in this approach, 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. Of course, this approach is ideal, and can certainly incorporate the notion of context. However, this most general approach has been considered impractical because of the exponential number of parameters involved.

A limited version of this model is the tree term dependence model (Rijsbergen 1977, June; Salton et al. 1982), in which 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, September; Salton et al. 1982).

The major limitations of this approach are discussed below. First, it requires user feedback to estimate the probabilistic parameters. This requires that a dumb retrieval method be used in the first pass, resulting in a possibly unmanageable result set which the user must not only scan through for documents of interest, but must now deliberately and reliably review to assess

each document for relevance feedback[5]. This is exactly the situation we were trying to avoid. This aspect of probabilistic retrieval is enough for us to consider alternative approaches.

Another aspect of the tree term dependence model which renders it unable to capture the notion of context, is that the selection of which term dependencies to consider -- recall that each term is considered to depend on only one other term -- depends upon the co-occurence data over the whole document base, and is not particular to each document. The dependency between two terms is included in the probability estimation if the two terms have a high expected mutual information measure or "EMIM". (see (Harper and Rijsbergen 1978, September)) over the whole document base. So, for example, if some document is about Nazi Comic-books, the system would most likely not be able to prevent spurious partial matches of this document while allowing it to match a relevant query; this is because the system would not choose to consider the dependencies between these two terms, since they would not likely have a EMIM over the whole document base[6].

The most difficult problem with the probabilistic models is the reliability of the parameter estimations, including the independent and joint probabilities

---

[5] It is possible to use a probabilistic approach to document retrieval without such feedback by using reasonable ballpark estimates as suggested in [Salton 1989 p. 288]. However, while this sort of estimate allows the use of a probabilistic approach to retrieval, it loses that aspect of probabilistic retrieval which is of interest to us here, i.e. the use of joint or conditional probability estimates to capture the notion of context.

[6] Van Rijsbergen proposes limiting the 'dimensionality' of the problem by considering only dependencies among *query terms* which appear in the document and *adjacent terms* in the maximum spanning tree (see (Harper, 1978)). Considering only query terms would alleviate the problem being discussed. But experimental results indicated that including the adjacent terms in the MST is necessary for realizing the benefits of the probabilistic model generally. With inclusion of those adjacent terms, the problem being discussed in the text returns.

for the general model, and the independent and dependent probabilities for the limited tree term dependence model. It is very telling that in the seminal paper introducing experimental results from the various probabilistic models, it turned out that the feedback which was necessary to estimate the probabilistic parameters, was then sufficient to enhance precision/recall results even without the probabilistic models (Harper and Rijsbergen 1978, September). Thus, in the end, while the method of relevance feedback is very powerful if it is practical, the estimation of conditional probabilities which could capture the notion of context proved unreliable. For all these reasons, we consider the problem anew in this paper, and are not satisfied with the probabilistic approach.

Our 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 , January/March). One of these approaches, 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." 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).[7] In this way, the precision of pre-coordination is not lost, while

---

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

not requiring that the patron guess the one correct citation order.

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

The proposal put forward 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, and partial matches are supported. The technique we put forward concentrates on the benefits of *intelligent* ordering, not the benefits of *arbitrary* ordering. The reason for this is that the benefits of having some *arbitrary* ordering rules are only reaped by a user who knows or guesses those rules. As we will discuss below, the benefits of an *intelligent* ordering can be realized in a computerized environment without any additional effort on the part of the user; the user enters his keywords as before, but the intelligent ordering of documents' keywords by professional catalogers prevents inappropriate partial matches.

5 Partial Coordination

5.1 Introduction

The central aim of partial coordination is to inhibit inappropriate partial

matches of keyword searches through the use of context. This is accomplished by defining scores for documents with respect to (hereinafter " w.r.t"). queries to depend not on the separate presence of individual matching terms, but on the particular *combination* of terms in 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. In general, given a query, we would have to individually specify the score for each document with a different combination of matching terms.

Two alternative approaches are possible. In one approach, the *user* specifies his *query* terms, then further specifies the score a document should have as a function of which query terms it matches. In the extreme case, the user specifies a score for each of the possible 2^n partial matches (e.g. if terms A and B match, score is 3, if terms B and D match, score is 0, etc.). In the opposite approach, the *cataloger* specifies a *document's* keywords, and further specifies the score this document will have in terms of potential queries. In the extreme case of this approach, the cataloger specifies the score of this document w.r.t. each of the 2^n queries which partially match the document's keywords.

In this paper we adopt the second approach, though not in its extreme form, in order to relieve the user of all efforts beyond listing his keywords. For each document, the cataloger specifies keyword dependencies, defined below, and this results in a function for each document's score w.r.t. all possible queries.

5.2 Vector Space Model Environment

We assume users are in a post-coordinate information retrieval environment based on the simplest form of Salton's vector-space model (VSM) (i.e. assume

independence of term occurrences in document subject headings) (Salton 1989) The user enters a list of keywords. This list need not contain any boolean operators at all; each document is given a score and ranked according to the weighted number of query terms appearing in the document's keyword list. Formally, the VSM assumes a given set T of possible terms for both queries and documents. This set creates the vector space. Each document and query lies in this space. Suppose the set T of possible terms is of size t. Then a document is represented as a vector of the form $D_i = (a_{i1},a_{i2},...a_{it})$ and a query as $Q_j = (q_{j1},q_{j2},...q_{jt})$, where the coefficients a_{ik} represent the weight for term k in document i and the coefficients q_{jk} represent the weight for term k in query j. Then, assuming term independence, the score of a query Q_j w.r.t. a document D_i is the dot product of the vectors, i.e. $a_{i1}* q_{j1} +... a_{it}* q_{jt}$. For simplicity, the examples in this paper will assume that all weights are unit and will assume the non-normalized simple dot product similarity measure, so the score of a document w.r.t. a query is just the number of matching terms. The proposal put forward in this paper is trivially extended to the more complex cases of weighted terms. It is not clear at this point how the proposal put forward in this paper would be combined with the extended VSM to account for non-independence of terms.

5.3 Introductory Example

Suppose we have a document describing how Nazi Germany distorted 19th Century French philosophy for its own propaganda, and we are considering the following as keywords: 19th Century, France, Philosophy, Nazism, Propaganda. Now, a user submitting the (overly) simple query 'Philosophy' (or the query 'France') will retrieve this book with a positive score, though a

score of zero would seem more appropriate.

Partially coordinating the document's keyword terms, however, can solve this problem by *hiding* the keyword "Philosophy" behind the keyword "Nazism". The cataloger determines that the document keyword "Philosophy" should be coordinated after the keyword "Nazism". Then, a user entering just the query term "Philosophy" will not retrieve this document at all, as the document term 'Philosophy' will not create a partial match with the query. According to the cataloger's specifications, the document keyword "Philosophy" will match the query term "Philosophy" *only if the query also contains the term Nazism.*

Our proposal is not to re-invent pre-coordination, but to marry it to the advantages of computers. This is illustrated with a brief overview and an example

6 Mechanics of Partial Coordination

6.1 Overview: Dependencies Replace Orderings

A key motivation for an intelligent ordering of terms is the cataloger's belief that a term C should only match a user's query in the context of another term A. In the card catalog environment, this is only accomplished by physically ordering the term C after the term A. Moreover, a card catalog requires a complete linear ordering of all the terms and users must duplicate the cataloger's thinking and arrive at the same ordering of terms.

Partial ordering allows the cataloger to specify a context for each term, but relieves the user of any need to order his query terms. This is done by utilizing

the computer's ability to permute, i.e: rather than requiring that a term C can only match *after* another term A, the cataloger requires that a term C can only match if A appears *somewhere* in the user's query. In this way, a term C will only match if the user is also interested in A, yet the user does not need to order his query terms. Instead of C *after* A, the cataloger specifies that C 'depends on' A.

6.2 Dependencies

In partial coordination, each document is represented in two parts: A list of index terms, and dependencies among those terms. The list of index terms is similar to pre- or post-coordinate terms. In post-coordination, the cataloging effort is completed with the list of terms. In pre-coordination, the cataloger (or the cataloging scheme which supplies the allowable subject headings) would struggle with *forming* term *phrases* and with *ordering* those term phrases, in order to prevent spurious partial matches. With partial coordination, these two efforts are replaced by the specification of term dependencies, the second part of a document index.

In the second part of a partial-coordination document index, each term is specified by the cataloger to depend on zero, one, or more other terms. If a term A is specified by the cataloger to depend on another 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.* The cataloger can thus specify a context for each term in the document index. In this way, a 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 other contextual terms, as specified by the cataloger.

6.3 Example

The aim of this example is to demonstrate how document terms can be prevented from matching queries out of context, without requiring the user to order his query terms. The example document describes how Nazi Germany distorted 19th Century French philosophy for its own propaganda purposes.

6.3.1 The document index

In this document, the list of index terms consists of four keywords: Nazism, 19th Century, France, Philosophy (Propaganda is omitted only because it adds nothing given our instructive purposes). The term dependencies are specified with a graph notation: An arrow from term A to term B indicates that term A depends on term B, and will add to the match only if the query also contains the term B. For example, figure one shows that the term France depends on the term Philosophy. Thus, even if a query contains the term France, this document will not match on the term France unless the query also contains the term Philosophy. The cataloger has thus indexed the document with a list of terms, and has further specified dependencies among the terms.

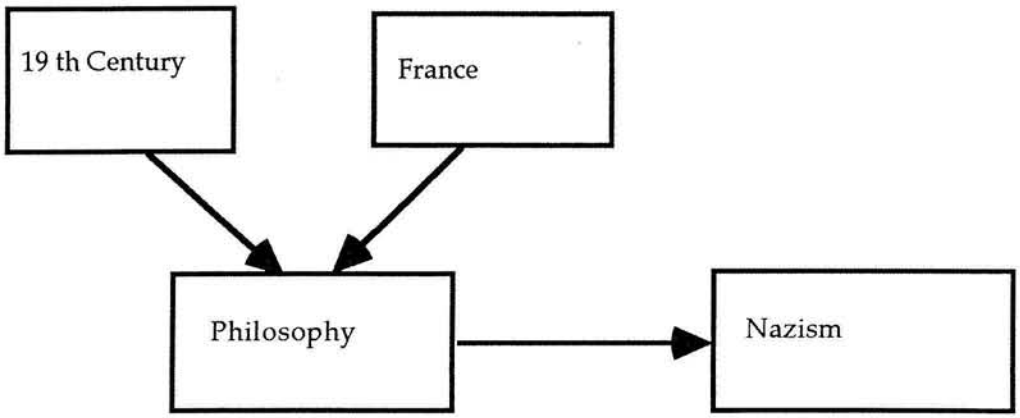


Figure 1

Query Terms	Using Partial Coordination	Traditional
19th Century	0	1
France	0	1
Philosophy	0	1
Nazism	1	1
19th Century France	0	2
19th Century Philosophy	1	2
19th Century Nazism	1	2
France Philosophy	1	2
France Nazism	1	2
Philosophy Nazism	2	2
19th Century France Philosophy	2	3
19th Century France Nazism	1	3
19th Century Philosophy Nazism	3	3
France Philosophy Nazism	3	3
19th Century France Philosophy Nazism	4	4

Table 1

6.3.2 Scoring of Document for Various Queries

Table one shows the score of this document w.r.t. some possible queries. We chose all the queries which only use terms in the document's list of keywords. Then the scoring of each query is described by the following rules:

> For each query term q, if q appears in the document's index terms, and if all that term's dependencies appear somewhere in the query, then q matches, and we add one point to the score; otherwise, q is no match and we go on to the next query term.

Consider, the (rather uninformed) query '19th Century Nazism'. The term 19th Century appears in the document's index terms, but its dependency -- i.e. Philosophy -- does not appear in the query. The term 19th Century is therefore prevented from matching out of context and adds nothing to the score. We go to the next query term, Nazism. This term also appears in the document's index, and all its dependencies (there are none) appear in the user's query, so this term is a match and one point is added to the score.

Consider another example from the table. For the user interested in 19th Century French Philosophy, this document will have points added to its score for the fact that the sort of philosophy of interest is the 19th Century variety (one point) and the French variety (one point), but will not have any points added for the ultimate Philosophy term, since, in the cataloger's opinion, this document is only really about philosophy in the context of Nazism. Note that the dependency relationships are not transitive: France and 19th Century each depend on Philosophy, and Philosophy depends on Nazism, but this does not mean than 19th Century and France depend on Nazism. The cataloger would

have to explicitly specify this dependency if he wished. This non-transitive definition of scoring allows points to be added for 19th Century and France, even though no point is given for the term on which they depend -- i.e. Philosophy. There are two related reasons for this non-transitivity. First, is the notion that this book should certainly score higher for someone interested 19th Century French philosophy than for someone interested only in philosophy in general. Thus, we want the scores of *this document w.r.t. various queries* to make sense. We also would like this book to score higher than a book on philosophy w.r.t. a query on 19th Century French Philosophy. Thus, we want the scores of *various documents w.r.t. a given query* to make sense.

After presenting the scoring algorithm more formally in terms of a query and a dependency graph, we will return in section eight to discuss more fully the strengths of partial coordination.

7 Formalism

7.1 Document Definition

The simple notation of a directed dependency graph (cycles permitted) can express the cataloger's preferred partial coordination of one term w.r.t. any others. For a document with u index terms, a dependency graph is a graph with u nodes. Each node is labeled with one term from the index terms. An arrow from one term to another indicates that the term at the arrow's tail is dependent upon the term at the arrow's head. We say the first term is dependent (for its contribution to the document's score w.r.t. a query) on the term at the arrow's head. It will then only add to the score of a document with

respect to a query in the presence of the term(s) on which it depends. Any term may depend on more than one other term, and this is simply indicated by multiple arrows originating from the same node (i.e. term). Cycles are allowed.

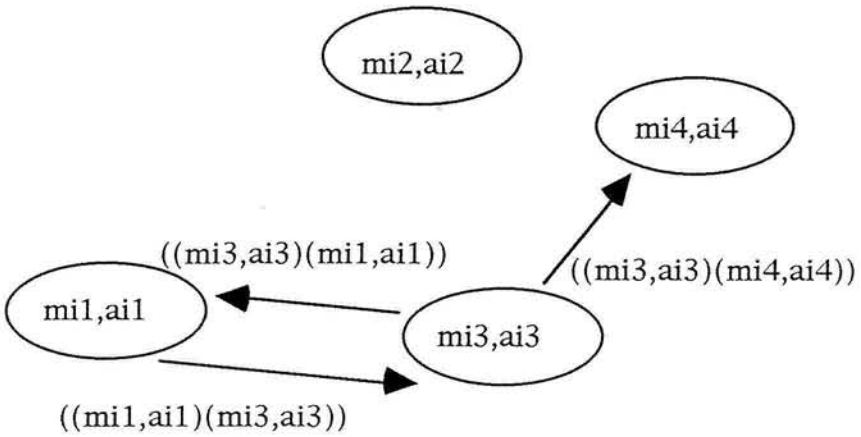
A document is formally represented as a dependency graph. A dependency graph consists of a set of nodes and a set of arcs among them.

The set of u nodes of dependency graph G_i for document i is denoted $N_i = \{ (m_{i1},a_{i1})..(m_{iu},a_{iu}) \}$, where m_{ij} represents the literal term j in document i and a_{ij} represents the weight of term j in document i. An example node is (computers, 0.5). A simple transformation V maps the jth term into its proper position in the vector space, $V(m_{ij}) \rightarrow \{1..t\}$. We will continue our formalism without this transformation, and return to VSM below.

An arc of dependency graph G_i is an ordered pair of nodes $A_{i,jk} = ((m_{ij},a_{ij}),(m_{ik},a_{ik}))$ where $1 \leq j,k \leq u$. The (possibly empty) set of arcs for dependency graph G_i is denoted A_i.

A dependency graph G_i for document i is a pair (N_i,A_i).

An example for the case of $u = 4$



If a term has no dependencies, it is represented as a node with no emanating arrows. If a term has no dependencies, and no other term depends on it, then that keyword will be represented in a disconnected node. Thus, in general, the representation of a document is an unconnected directed graph, with one node for each keyword term.

A query of w terms is $Q_i = (m_{i1},a_{i1}),(m_{iw},a_{iw})$, where m_{ij} represents the jth literal term in query i and a_{ij} represents the weight of term j in query i. Once again, the simple transformation $V(m_{ij})$ --> $\{1..t\}$ can map a query term into the VSM.

7.2 Scoring Algorithm

A node with n originating arrows indicates that the term is coordinated after all n terms. That is, according to the formal scoring algorithm below, it contributes to the document's score on a query only in the presence of ALL the terms at the heads of the arrows emanating form it. The arcs, then, are AND arcs. Any combination of AND and OR arcs would be feasible, but we restrict

ourselves to AND arcs for simplicity.

The score of a query with respect to a document is defined in terms of the document's dependency graph. The scoring algorithm of a query w.r.t. a document is as follows:

(0) For each term (m_{ir},a_{ir}) in the query:

(1) If the graph contains no node (m_{ij},a_{ij}) such that $m_{ij} = m_{ir}$, then add 0 to the score and go to next term

(2) If the graph contains a node (m_{ij},a_{ij}) such that $m_{ij} = m_{ir}$, make that the new current node

(3) For each arc $((m_{ij},a_{ij}),(m_{ik},a_{ik}))$ in A_i emanating from the current node

(4) If the query contains the term (m_{ik},a_{ik}), then continue to the next arc

(5) Otherwise add 0 to the score and go to next term (m_{ir},a_{ir}) (i.e. back to step (0)

(end ForEach arc)

(*) Add one to the score of the document w.r.t. the query

(end ForEach term)

Note that extending the algorithm to account for weighted document and/or query terms involves modifying the last step 'Add one..' to 'Add $a_{ij} * a_{ir}$...'

Although each document and query can be viewed in terms of the VSM, the score of a document is now defined by the above algorithm rather than by a simple inner product. Therefore, nothing is gained by using the *V* transformation to map documents into vector space, since even if the

document and query terms are transformed to 'line up', the score still cannot be defined as the simple inner product. We have therefore omitted the V transformation in our definitions.

8 Benefits of Partial Coordination

8.1 Example Discussed

In our example the scores for this document w.r.t. these possible queries are a clear improvement over the traditional scores. For example in the given table we can detect improvements by comparing whether the relative scores w.r.t. each query make more sense under the partially coordinated or traditional system. The benefits would also be visible by comparing (under each scheme) the score of the example document to the scores of other obviously-more-relevant or obviously-less-relevant documents. That analysis is omitted to avoid confusion.

In the example document, each of the terms 19th Century, France, and Philosophy have dependencies. For example, the relative increase of this book's score due to the presence of the term 'Philosophy' in the query, depends on the context -- i.e. it depends on whether the user is known to be interested in Nazism. If he is, then the philosophy discussed in this book resembles what the user may have had in mind with the query term Philosophy, and a point is added to the document's score for matching the term Philosophy. If he is not, then the philosophy discussed in the book should not even add the one point to the document's score, since this is not likely the sort of philosophical discussion the user had in mind.

This ordering of terms -- Nazism before Philosophy -- is roughly comparable to the card-catalog decision not to file this document under Philosophy but only under (say) Nazism--Philosophy. A card-catalog user looking under Philosophy will not find this book, while if he looks under Nazism first, then looking under the subdivision Philosophy will help him. On the other hand, there are many differences between partial and pre-coordination, as will be discussed in the following section.

Ten of the fifteen possible queries resulted in lower scores under partial coordination. The query '19th Century France', for example, would give a score of zero under partial coordination, but a score of two under post-coordination. All of these score-lowerings seem beneficial. Note that given a set of index terms, adding dependencies can only lower scores in appropriate matches for greater precision.

One possible objection to this supposed improvement is that the choice of terms '19th Century, France, etc.' lends itself to out-of-context partial matches, and that a good cataloger would never include such terms is a post-coordinate environment for fear of such poor matches. We agree completely with this objection and take it up in section 10.2. It may very well be that catalogers would be afraid to include such terms in the document's index. But the document is, after all, about an abuse of 19th Century French Philosophy. Certainly the terms 19th Century, France, and Philosophy are relevant for recall and ought to be included in the document index, were it not for fear of poor precision. Thus, partial coordination should also be viewed as enhancing recall by allowing catalogers the freedom to include relevant terms without fear of poor precision! In sum, for a *given* set of index terms, partial

coordination can serve only to lower inappropriate scores for increased precision; it cannot ever raise a score above the traditional post-coordinate score. But the choice of terms is not, in fact, fixed. Rather, partial coordination may serve to allow inclusion of additional recall-enhancing terms. We return to this point below in section 10.2.

We believe that an examination of the scores in this example demonstrates the potential of partial coordination to significantly shift out the recall/precision curve for post-coordinate keyword indexing. This is accomplished by combining the strength of post-coordination -- i.e. relieving the user from any need to know about the rules of citation order -- with the ability of pre-coordination to enhance precision by preventing inappropriate partial matches. The appropriately lowered scores of irrelevant or partially relevant documents w.r.t. the various queries will serve to improve the rank ordering of documents.

8.2 Partial as Generalization of Pre-Coordination

Partial coordination differs from full pre-coordination in three ways. First, even if we view the dependencies (e.g. B *depends on* D) as traditional orderings (i.e. B *after* D), the cataloger does not specify one linear citation order of all the document terms. Rather, he specifies a set of individual restrictions on that order. Many complete orderings may satisfy the individual restrictions -- imposed by the cataloger -- of a document's terms. For example, the cataloger may specify 'A after B, D after E, and no restrictions on B, C, or E'. Many complete orderings would meet these restrictions.

Second, partial matches are possible as in the post-coordinate or VSM environment; any partial match may be recognized, and the greater the match, the higher that document's score w.r.t. 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.

Third, orderings are replaced by our existential dependencies[8]. That is, the actual ordering of query terms never matters. What matters is whether a contextual term appears *somewhere* in the query or is totally absent from it.

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

It is important to note that the term dependencies are specified for each document. Two documents may be indexed with the identical keywords but with different dependencies among those keywords, as when the same keywords can be differently related syntactically and semantically to have different meanings. This leads us to note one circumstance in which our 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

---

[8] Replacing orderings with dependencies may be viewed as simply allowing any of the many partial orderings which fulfill the dependencies.

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.

9 Desirable Mathematical Properties of Extended VSM

Recent work (Lee 1994) has analyzed the properties of various proposed extended boolean operators which allow for non-binary operand and outputs. For example, in some extensions of boolean logic, AND is defined as product. Lee describes the "negative compensation problem" of this definition, as follows: Take a document with weighted keywords

D = { (Information .7) (Retrieval .7) } and two queries

q1 = Information AND Retrieval

q2 = Information

The queries are evaluated w.r.t. the document as .49 and .7 respectively. This quality of the product definition for AND is counter-intuitive, as most people would say the document is a better match for q1 than for q2.

We propose that this sort of analysis is properly applied to non-boolean environments such as VSM. In the VSM environment which we assume in this paper, the user adds terms to a query with an implicit operator. It is not a boolean OR or AND, but it is an operator which indicates the user's interest is not simply in the directions of terms 1..(t-1), but also (simultaneously) in the direction of term t. We will therefore refer to the implicit VSM query operator as the ALSO operator. A VSM query $Q_j = (q_{j1}..q_{jt})$ can thus be viewed as q_{j1} ALSO q_{j2} ...ALSO q_{jt}. An analysis after (Lee 1994) would then focus on the properties

of the ALSO operator. Such as analysis must differ slightly from (Lee 1994), as there is no way to evaluate the properties of such an expression except w.r.t. another point in the vector space -- viz. a document. An analysis of the ALSO operator is therefore an analysis of the properties of ALSO=ALSO($q_{ji}..q_{jt}$,D_i).

What properties ought the ALSO operator have? It can be shown that the ALSO operator satisfies all the properties delineated as favorable by (Lee 1994), if the definition of those properties is extended to be a function the query term weights *and a given document.* But there is another property which we propose as desirable. If the ALSO operator is not homomorphic w.r.t. addition of disjoint queries, then it will be more flexible. Two queries $Q_j = (q_{j1}..q_{jt})$ and $Q_2 = (q_{21}..q_{2t})$ are disjoint if for no *i*, are both q_{1i} and q_{2i} non-zero. Not homomorphic w.r.t. addition means that the following does not hold: $ALSO(Q_1+Q_2) = ALSO(Q_1) + ALSO(Q_2)$. In particular, we can implement the notion of context if the operator is not homomorphic w.r.t. addition. In those cases where $ALSO(Q_1+Q_2) > ALSO(Q_1) + ALSO(Q_2)$, the inequality can be used to implement the notion of context. A document index term which matches a query term may add more to the document's score, depending on the other contents of the query.

This is exactly what has been proposed in this paper. The simple VSM (assuming term independence) evaluation of a query is homomorphic w.r.t. addition. But by extending the definition of a document to include not only (possibly weighted) terms but also a dependency graph, and by extending the definition of the evaluation of a query w.r.t. a document to the algorithm in section 7.2, we have altered the implicit VSM ALSO operator so it is no longer homomorphic, and is used to implement the notion of context as discussed.

It is not obvious how this property is properly discussed for the extended boolean operators. For example, if, as in the traditional fuzzy set extension to boolean operators, OR is defined as MAX and AND as MIN, we would not be interested in homomorphism w.r.t. addition which makes no sense at all. What property, then, would we want those extended boolean operators to have in order to implement the notion of context ? This is not at all clear.

This question can also be viewed from another perspective. As mentioned in section four, it is conceivable that some combination of extended boolean operators can be used to implement any arbitrary dependency graph. Let S be the scoring algorithm of section 7.2, and D_i be a document and Q_j a query of our partial coordination system as defined in section 7.1. Then what is required is

1) a mapping *f* from a pair $<D_i Q_j>$ to a pair $<D_i', Q_j'>$ (where D_i' is a simple vector and Q_j' is an extended boolean query) $f: <D_i Q_j> \rightarrow <D_i', Q_j'>$ and

2) a definition of score S': $<D_i', Q_j'> \rightarrow$ Real

such that $S(<D_i Q_j>) = S'(f(<D_i', Q_j'>))$ for all $<D_i Q_j>$

In section four we showed the beginning of such a mapping and scoring function. Take a dependency graph (i.e document) of the form: A--->B, and a query of the form (A, B), with A and B as defined in section 7.1 and the scoring function of section 7.2. Then the query would be mapped to an extended boolean query of the form (A AND B) OR B, the document would be mapped to a simple vector (A B), and a scoring function would be introduced which defines AND as SUM and OR as MAX. But it is not at all clear how to complete such a mapping. And in any case, this approach would place an enormous burden on

the user, which our approach avoids.

We remain, then, with the following questions: Can a property be identified which extended boolean operators should possess to enable easy implementation of the notion of context ? Alternatively, can a mapping be found which would allow sophisticated users to specify the required context as part of the query? Until these questions are answered, our proposal for extended VSM is alone in facilitating an implementation of the notion of context.

10 Choice of Terms for Pre- and Partial Coordination

The mechanics and improvements of partial coordination have a secondary effect. Catalogers are able to choose desirable keywords with partial coordination, which would be impossible to include in pre-coordination, or unwise to include in post-coordination. The following two sub-sections raise these two points, respectively.

10.1 Freedom to Choose Arbitrarily Related Keywords

Partial coordination is more flexible than pre-coordination not only because it allows more flexible ordering of terms, but because the cataloger is not limited in his choice of keyword terms[9]. Partial orderings are applicable between any

[9] It is surprising that little or no attention has been paid in the academic literature to the differences between keyword terms which appear in pre-coordinated subject headings and those which are used for post-coordinate searching, and the implications for proper user

two terms, and any term can be used as a keyword. In pre-coordination, only a small number of terms is available. The limitation we have in mind is not due to the use of controlled vocabularies, which is incidental, but is an inherent result of the philosophy of pre-coordination.

The reason a document cannot be labeled with any arbitrary relevant keyword in pre-coordination, is that pre-coordinate schemes attempt to represent the semantic relationships between terms, and are usually rather limited in these representations. So, for example, our example document could not have been labeled Nazism--Philosophy, because this would *mean* the philosophy of Nazism, and the book is not about that (it is, rather, about the abuse by Nazism of some *other* philosophy). It is the nature of pre-coordinate schemes to focus on one complete order with a clear meaning. This focus limits the use of coordination to those terms which happen to relate semantically or syntactically in a way that can be represented in the pre-coordinate scheme. There is no way to include a term (such as French Philosophy in our example above) whose relationship to the other terms is not representable in the scheme. Of course, if such a term cannot be included, it also cannot be well ordered.

Partial coordination purposely abandons the attempt to represent particular relationships, and does not even recognize the difference between a syntactic and semantic relationship among terms. If, the cataloger believes that one term should be hidden behind another in a given document, this is specified. At the same time, the user is also not required to ponder the relationships among terms or any cataloging rules.

---

behavior and retrieval performance.

Pre-coordination was mostly interested in establishing good, complete citation orders. The whole notion of a partial match is rather obscure in the case of card catalogs, and so that emphasis was appropriate. In the computerized environment, however, where inappropriate partial matches are a big problem, we propose shifting the emphasis from ideal complete orderings, to the ability to hide individual terms behind others. Abandoning the attempt to represent relationships among terms simultaneously frees the patron from wondering about those relationships, and allows the cataloger to hide any term behind any other, regardless of whether those terms relate in one, big, recognized, fully-ordered manner.

10.2 Freedom to Choose Non-Categorical Keywords

There is a second manner in which partial coordination allows greater flexibility in the cataloger's choice of terms than either pre- or post-coordination. Because the patron must correctly guess the terms and their citation order, and in order to keep the scheme workable, pre-coordinate subject headings are rather short, i.e. they are not deeply nested; this shortening limits the number of places the patron can go wrong as he attempts to guess the terms and their order. Related to this, the individual terms tend to be categorical rather than particular, since the shallow subject headings must cover the whole topic. In the case of the book on Nazism's abuse of French philosophy, even if we found in some scheme a subject heading Nazism--Propaganda--Misrepresentations, we would certainly never see any particular mention of French Philosophy -- a particular *instance* of Misrepresentation -- in the pre-coordinated subject heading. Beyond that, we

are very unlikely to see such a deep subject heading altogether. The scheme would become unwieldy, and there would be just too many places where the patron could go wrong. We would therefore expect something no deeper than (say) Nazism--Propaganda. As for the possibility of a short yet *instantiated* subject heading such as Nazism--Philosophy,French, this is also out of the question, since, as described above in section 10.1, the relationship between the two terms is then obscure, contrary to the aims of pre-coordination.

Thus, we can potentially include particular, non-categorical keyword terms only in partial or post-coordination. The trouble with this sort of specific, instantiated keyword in post-coordination, however, is that including non-categorical terms can lead to spurious partial matches. In particular, the specific keywords may match *out of context*, as an instance of one theme can also be an instance of another. For example, if France and Philosophy are included as keywords, then a query about the influence of Sartre on French Philosophy may retrieve the document. Partial coordination prevents exactly this sort of mismatch, and allows catalogers to fully exploit the recall and precision-enhancing effects of particular, non-categorical keywords.

11 Feasibility for Libraries and Internet

One aspect of the practical elegance of partial coordination is that retrospective conversions are not required, since the traditional VSM search is just a special case of partial coordination, i.e. where no dependencies have been specified for documents' keywords. Thus, one retrieval engine can be used immediately for all the documents in a collection. However, one would

have to work out a scoring mechanism which could properly relate the two types of scores, or else separately score the results into two retrieved sets, one for documents which have been cataloged with partial coordination, one for documents which have not.

We believe that partial coordination is a simple enough extension to the cataloging task -- which, in any case, is usually carried out by a very small number of catalogers whose decisions are then shared with others -- that it is practically feasible as a new method of keyword cataloging. Of course, empirical testing of such an extension needs to be carried out on a large document collection to demonstrate the promised improvements.

As the Internet grows providing larger numbers of online documents, full text indexing will diminish in value due to losses of precision. Post coordinate search enabled by full text indexing of content on the Internet will result in unmanageably large retrieval sets. Under these circumstances, keyword indexes -- perhaps augmented by partial coordination -- can yet prove to be a feasible and effective form of indexing content on the Internet. While projects such as INTERCAT are under way to catalog Internet resources in a mostly traditional way, the best hope for up-to-date subject access to the whole Internet is good indexing by the individual document authors. HTML authoring environments could easily be re-configured to encourage publishers to include keywords for their documents, and could also easily provide the simple graphical tools for partial coordination. Second as the Internet community evolves new location independent standards for referencing documents (such as URI), keywords can be specified in the header or document identifier. These and other Internet-specific issues will be more

fully addressed in forthcoming work. In the context of this paper, we argue that partial coordination is a simple enough extension to a familiar enough indexing technique (i.e. keyword labels) that it is a promising alternative for practical implementation and use.

Summary

The proposal put forward in this paper combines the strengths of pre- and post-coordination. Partial coordination improves precoordination by allowing any term to be dependent on any other, without limiting the cataloger to specifying one complete ordering. Furthermore, any term may be specified to depend on any other, regardless of whether the particular relationship between terms can be identified or specified. Better than pre-coordination, the searchers's query is a simple list of keywords. And partial coordination is an improvement over post-coordination as it avoids spurious partial matches.

At the same time, our proposal can be viewed as an extension to the traditional VSM. In this extension, a document is represented by a vector of terms and a dependency graph of those terms. The score of a document w.r.t. a query is defined by an algorithm which selectively includes individual terms of the traditional inner product, depending on the context of the users' query.

The chief benefit of partial coordination is enhanced precision. This is achieved by avoiding matches of individual query terms out of context. Moreover, this benefit is achieved while requiring of the user only what is required in the traditional VSM -- i.e. a list of terms. It is the cataloger who must, for each document, specify a dependency graph of document terms in

addition to a list of document keywords.

We believe this proposal is attractive for many practical reasons. First and foremost is the aforementioned ease of use for patrons. Second is the relatively easy extension of the cataloger's work. Third is that a retrospective conversion is unnecessary. We further believe that for many of these reasons, and because it promises to enhance precision, the proposal is promising and practical for use on the Internet, where large retrieval sets are a big concern.

The immediate follow-up to this work will need to be an experiment to test the promised enhancements in precision for a given level of recall. Whatever the empirical benefits of the kind of partial coordination proposed here, many benefits of pre-coordination are lost in the post-coordinate and VSM environments. It is a worthy research objective to recapture some of those lost benefits. Finally partial coordination may improve document search and retrieval in large networks such as the Internet.

References

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

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

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

Cochrane, P. A. (1983, March). "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.

Cooper, W. S. (1988). "Getting Beyond Boole." *Information Processing and Management, 24*(3), 243-248.

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

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

Fox, E. A., and Koll, M. B. (1988). "Practical Enhanced Boolean Retrieval: Experiences with the SMART and SIRE Systems." *Information Processing and Management, 24*(3), 257-268.

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

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

Lancaster, F. W. (1986). *Vocabulary Control for Information Retrieval,* Information Resrources Press, Washington.

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.

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

Lee, J. H. (1994)."Properties of Extended Boolean Models in Information Retrieval." *Proceedings of 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, (182-190).Dublin, Ireland,

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

Markey, K. (1980). "Analytical Review of Catalog Use Studies." *ED186041*,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, January/March). "Subject-Searching Experiences and Needs of Online Catalog Users: Implications for Library Classification." *Library Resources & Technical Services, 29*(1), 34-51.

Markey, K. (1988, September). "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, December). "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.

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

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

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 and Knowledge about an Online Catalog." *College & Research Libraries* (January 1984), 66-70.

Svenonius, E. (1986, September). "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.

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.