Exploiting Coauthorship to Infer Topicality in a Digital Library of Computer Science Technical Reports

James C. French and Charles L. Viles *
Department of Computer Science
University of Virginia
Thornton Hall
Charlottesville, Virginia 22903
{french, viles}@virginia.edu

Technical Report No. CS-96-20

December, 1996

Abstract

We propose a method of mapping the topical content of distributed digital libraries and demonstrate the technique using data from the Networked Computer Science Technical Report Library (NCSTRL) digital library project. This method seeks to exploit information derived from document coauthorship to produce improved automatic subject classifications of the documents. In a distributed digital library, these subject classifications are useful in characterizing both intra-site and inter-site content. They are also helpful in providing secondary retrieval services. We present the method and describe an experiment and results showing that improved clusterings can be achieved relative to traditional document clustering.

1 Introduction

Digital library projects[6] place heavy demands on existing infrastructure and hence, the performance of these systems is an object of considerable study. With the increased availability of digital libraries and distributed information retrieval systems, we can expect increased demand for additional services providing better information access. This paper looks at *topical clustering* as a mechanism both for examining issues affecting retrieval performance and for providing additional access to digital library resources. These two points are discussed separately below.

Viles and French[8, 20] showed that the distribution of content in a distributed information retrieval system could manifestly affect the retrieval effectiveness of the system. This concept, termed *content skew*[8], is one motivation for the current study. The main result from this work was that in content-skewed systems, sites need to share information in order to maximize effectiveness. Thus the presence or absence of skew has a direct influence on the communication requirements

*This work supported in part by NASA Goddard Space Flight Center under NASA Graduate Student Research Program Fellowship NGT-51018 and by NSF grant CDA-9529253.

of a distributed digital library. The methodology proposed in this paper is aimed at providing a mechanism for quantifying content skew in deployed digital libraries.

Another area in which knowledge of the topical distribution of a system can be used is improving retrieval efficiency. Performance benefits can be gained in widely distributed IR systems if many sites can be pruned from a search list. The remaining sites can be most efficiently searched by directing the query to the most promising sites first. Knowledge of the topical distribution is essential for both these purposes.

The second motivation for investigating topical distribution in distributed IR systems is to provide adjunct retrieval capabilities to the IR system. We have discussed a number of possible capabilities in earlier work[7]. Here we are specifically interested in the ability to broaden a query along the line of *topical relatedness*. For example, when users are examining the results of their queries, they might like to identify particular documents and ask for others on the same topic. This is easily done if the holdings have been cataloged but many free text systems will not have this luxury. They will have to deduce topical relatedness by other means. Knowledge of the topical content distribution will help provide a useful browse mechanism.

We have chosen NCSTRL[5], the Networked Computer Science Technical Report Library[1], as an appropriate digital library for our work in topical clustering. The NCSTRL project provides a perfect testbed for an initial study of this type. It is an internationally distributed collection of computer science technical report archives that is treated as a cohesive whole for searching and retrieval purposes. Each individual participating site has complete autonomy in the policies governing the publication of its technical reports and uses one of two software suites, Lite or Standard, to connect itself into the larger digital library. Bibliographic records are maintained at the issuing site for Standard sites and collected into a central server for Lite sites. User queries are processed at all sites by means of a distributed search.

At first blush, one is tempted to say that NCSTRL is a topically homogeneous digital library. After all it is a collection of computer science technical reports. At some level of granularity this is true, but as a practical matter there are many identifiable topics within computer science just as in any other discipline. The goal of this research is to identify these topics with reasonable accuracy and entirely automate the process.

Our strategy is to analyze each individual site to discover its topical content and then try to reconcile these topics across all the sites in the digital library. We are using clustering techniques to collect "topically similar" technical reports into topics. Ideally these topics would be well defined such as: theory of computation, database management, or artificial intelligence. Realistically we might not be able to apply such labels unambiguously. Our strategy is to cluster all the technical reports at a particular site into some reasonable number of strongly topically related clusters without worrying about the particular label to apply to these clusters. After all sites have been clustered in this manner, we will attempt an inter-site cluster phase with the goal of gathering similar site-topics together into a uniform set of topics. We can then relabel the topics at each site with these new labels. Having created such a topical map we are then in a position to attempt to analyze the content-skew of the collection or provide query expansion to topically related documents.

The remainder of the paper is organized as follows. Section 2 surveys related work. Our new clustering approach is described in section 3 and some experimental results are given in section 4. We conclude with a discussion of some future directions.

[1]NCSTRL can be accessed via the URL *http://www.ncstrl.org/.*

2 Related Work

Citation analysis has been studied from the early days of information retrieval (c.f., [1, 2, 16]). Two measures often considered are *bibliographic coupling*[13] and *co-citation*[18]. The bibliographic coupling of two documents is the number of references they have in common; their co-citation strength[9] is the number of documents that jointly reference them. Thus, bibliographic coupling is static while co-citation strength may vary. The calculation of these measures is facilitated by access to a citation index[9].

Citation indexing is a well-known technique that has been used successfully for query expansion. In this technique a retrieved document is considered to be relevant to documents it cites. As Green[10, p. 649] has noted, the main assumption upon which citation indexing is predicated is that "if one document relevant to a user need has already been identified, it may be used as a stepping stone to other relevant documents." Clearly this assumption also underlies the topical relationship that we are trying to infer.

Of particular relevance to this paper is the study of social science journal citations due to Arms and Arms[1]. Although that study concluded that "hierarchical methods of clustering or classification are unsuitable for the social sciences," it held a different view of the natural sciences suggesting that "hierarchical methods could well prove adequate in a well-structured scientific field and the results could be expected to be relatively independent of the precise method used."[1, p. 11] Although our approach does not use citations, it has the same goal of subject classification as the Arms study.

Others have also used citations in a variety of ways for subject classification. Carpenter and Narin[3] and Small and Koenig[19] clustered scientific journals into subdisciplines. The former study was based on cross citing among the journals while the latter used a novel variant of bibliographic coupling. Kwok[14] reported on a method of classifying documents based on titles and cited titles. This method provided compact document representations with good properties.

As valuable as citation analysis appears to be for document classification, we cannot use it on the NCSTRL collection because we only have access to bibliographic information about the documents and, therefore, have no access to their citations. This led us to a different strategy for content analysis and subsequent topical clustering. Our approach uses the indicative words present in titles together with a novel use of coauthorship to achieve good levels of topical clustering.

3 Topical Clustering for Content Analysis

Some form of document clustering seems to be the appropriate approach to this problem. Of the three hierarchical agglomerative clustering methods most often used to cluster document collections — single link, complete link, and group average — we believe that group average offers the most promise for the present application and have chosen it as the preferred method for our work. Croft[4] and Voorhees[21] have shown these algorithms to be practical for reasonably large collections. Our approach will be seen to extend the range of document collections over which these algorithms can be applied.

After considerable research in the 70's, document clustering has still not been shown to increase retrieval effectiveness to any significant degree. However, the efficacy of document clustering as an aid to browsing, the use that we are exploring, is still an open issue. The reader is referred to

Sparck Jones[12] for an excellent retrospective of clustering research in automatic classification and to standard IR[15] and clustering[11] textbooks for specific details of clustering techniques.

Our first approach toward identifying the topical content of the NCSTRL collection was a straightforward application of a generic clustering algorithm to all the document titles in the collection. We simply collected all he titles of all the technical reports in the NCSTRL pilot[2] and applied a group average clustering algorithm using several different similarity measures. This approach proved ineffective and did not adequately differentiate the topical content. The most likely source of this failure is short titles.

This led us to consider alternative approaches to the problem and, more particularly, to try to exploit characteristics of the document collection. The key observation to the work here is:

> *Coauthorship hypothesis*: Several reports coauthored by the same, or a substantially similar set of authors, are very often topically related.

Reports exhibiting this characteristic grow out of joint work within a department or a research group within a department. Preprocessing site bibliographic information based on this observation forms the basis of our proposal. This strategy is designed to aggregate several titles when possible into a larger chunk of text that can be more effectively clustered by the subsequent conventional document clustering phase.

An observation by Salton[17] in an early study on the use of bibliographic information in document retrieval lends support to this hypothesis as well. Salton noted that[17, p. 445]:

> In particular, it may be conjectured that information associated with the *author* of a given document, for example data contained in related publications of the same author, may furnish usable content indicators.

Our preliminary investigations lend substantial credence to the coauthorship hypothesis and it forms the basis of the topical clustering method described below.

The topic analysis proposed here has two clustering phases. In phase one, all the reports at a given site are clustered. This *intra-site* clustering phase is based on document authorship. In phase two, the clusters from phase one are treated as documents comprised of all titles and keywords of the constituent documents. This *inter-site* phase trys to associate topics across sites. Both phases are based solely on bibliographic records obtained from the NCSTRL index servers. This is described in more detail below.

Before describing our approach in detail it will be useful to summarize some of the advantages that we see for the inclusion of authors in the clustering process.

- Titles are frequently short and often misleading.
- There is additional information in the collection that is embodied in the coauthorship of the reports. This approach seeks to take advantage of this additional information.
- The collection is large and growing. It is hoped that NCSTRL, deployed in November 1995, will have 200 participating sites by November 1997. This will place a premium on algorithmic efficiency if secondary services are dependent on topical content analysis.
- As will be seen, much of the processing can be distributed. A small amount of preprocessing at each site will provide enormous efficiency gains in the topical analysis.

[2] At the time of this work there were approximately 8,500 reports in the collection.

3.1 Using Coauthorship for Clustering

The general strategy for mapping the topical content of the NCSTRL collection proceeds in three steps.

1. *Intra-site clustering:* Preprocess the documents at each site by coauthorship criteria. This results in a provisional placement of documents into topic groups but makes no effort to decide the final topic set. The result at a site is a partition of its document collection into subsets the members of which are considered to be topically related. It should be noted that this document partition is not a partition by topic; two or more subsets in the document partition may in fact be from the same topic.

2. *Inter-site clustering:* All the provisional topical clusters generated at each site are collected together and clustered to form final topics. This step employs conventional document clustering techniques and results in a hierarchical organization of the topic space. The hierarchy may provide a useful browsing structure, but for the work reported here we chose to break the hierarchy into a partition of the topic space.

3. *Make final topic assignments:* Relabel and coalesce the provisional topics at each site into the topic classes identified in step 2.

These steps are discussed in greater detail next.

3.1.1 Intra-site Clustering

The intra-site clustering proceeds in two steps. First we generate *seed clusters* and then we merge them where appropriate to reduce the number of clusters in the final set representing each site.

Forming the Seed Clusters

The steps employed to form the seed clusters are as follows.

1. Put author names in canonical form. Our choice is to use the last name and first initial as a unique author identifier. This will clearly not uniquely identify all authors in the collection but we believe the clustering error rate due to this will be small and isolated to specific sites. The difficulty in locating a second initial consistently from the data prevented us from further qualifying the names.

   Thus, John Doe would be identified as *DoeJ*. If in this process and attempt is made to duplicate a label, an integer is added. So Jane Doe would become *DoeJ1*. Any reasonable strategy could be used as long as the authors can be distinguished. In the running example in this paper we use initials only for brevity.

2. Sort the author identifiers for each technical report to form a label for the report. Examples of these final labels can be found in Table 1.

3. Merge all reports with the same set of authors (i.e., the same report label). These form the seed clusters. Each seed cluster contains:

   (a) a list of all report identifiers for the constituent reports;

(b) a list of all the words in the titles of all the constituent reports; and

(c) a list of the authors.

Label	Author Tag	Title
a	FJ	A Global Time Reference for Hypercube Multicomputers
b	FJ;VC	A Software Toolkit for Prototyping Distributed Applications (Preliminary Report)
c	FJ;GA;RP;WA;WW	A Synopsis of the Legion Project
d	FJ;JJ	An Archive Service with Persistent Naming for Objects
e	FJ;GA;PJ	An Introduction to the ADAMS Interface Language: Part I
f	FJ;VC	Availability and Latency of World Wide Web Information Servers
g	BP;FJ;KD;OR;PJ;SS	Basic Data Concepts in ADAMS
h	FJ;JA;PJ	Scientific Database Management (Panel Reports and Supporting Material)
i	FJ;GA;KJ	Breaking the I/O Bottleneck at the National Radio Astronomy Observatory (NRAO)
j	FJ;VC	Dissemination of Collection Wide Information in a Distributed Information Retrieval System
k	FJ	Electronic Distribution of Technical Reports and Working Papers: A Simple Cooperative Approach
l	DM;FJ;PT	Performance Measurement of a Parallel Input/Output System for the Intel iPSC/2 Hypercube
m	FJ	Heuristic Clustering of Signature Files: A New Approach (Preliminary Results)
n	FJ;GA;KJ	High Performance Access to Radio Astronomy Data: A Case Study
o	BJ;FJ	Scalable Database Support for Correlation and Fusion Algorithms
p	BP;FJ;GA;JS;LY;...	Implementation of the ADAMS Database System
q	BJ;FJ;KP;MW;WJ	Indexing Multispectral Images for Content-Based Retrieval
r	FJ;GA;RP;WA;WW	Legion: The Next Logical Step Toward a Nationwide Virtual Computer
s	FJ;PJ	Multiple Inheritance and the Closure of Set Operators in Class Hierarchies
t	FJ;PT	Performance Measurement of Two Parallel File Systems
u	FJ;GA;KJ	Extensible File Systems (ELFS): An Object-Oriented Approach to High Performance File I/O
v	FJ;JA;PJ	Scientific Database Management (Final Report)
w	FJ;PJ;WJ	Scoping Persistent Name Spaces in ADAMS
x	BJ;FJ;KP;MW	System for indexing multi-spectral satellite images for efficient content-based retrieval
y	BP;FJ;GA;JS;KA;...	The ADAMS Database Language

Table 1: List of reports with labels and author tags.

Merging into Final Clusters

The general strategy for merging the clusters is as follows.

1. Divide seed clusters into two groups, those that are topically prominent and those that are not. As an example of reports in these two categories, an author represented in the collection by a single report with no coauthors[3] would not be considered topically prominent; a group of three reports with the same four coauthors would be considered topically prominent.

[3]This situation is characteristic of a Master's thesis or Ph.D. dissertation issued as a technical report.

2. Apply combining rules to place non-topically prominent reports within topically prominent groups and to combine topically prominent groups. Note that this step is based on coauthorship criteria and therefore, the extreme case of single report with single author would not be coalesced by this step; these reports would be passed through to the inter-site cluster step described later.

Let $C = (A, n) = (\{a_1, a_2, ..., a_p\}, n)$ denote a cluster, C, of n papers attributed to the set A of p authors listed. If C is a seed cluster, all n papers have the same p authors. If C has been formed by one or more applications of a combining rule, then some of the articles may have been authored by more or fewer than the p listed authors.

Given two clusters (A, m) and (B, n) we have three cases to consider:

1. they should remain two independent topical entities;
2. they are topically similar and (A, m) should be absorbed into (B, n); or
3. they are topically similar and (B, n) should be absorbed into (A, m).

Before we can apply any combining rule we must first determine if the two clusters are topically similar. A general discussion of the algorithm follows next.

For this work we used a weighted Jaccard coefficient[15] to determine topical similarity. Let $C = (A, n)$ denote a cluster. For all $a \in A$ we define the *weight* of a, $w(a)$, to be the number of papers on which a appears as an author. The quantity $w(a)$ is the weight of the author a in the collection. These weights form the basis of the similarity calculation. Given two clusters C_1 and C_2, we define the similarity of C_1 and C_2 by the weighted Jaccard coefficient, $J_w(C_1, C_2)$, given by

$$J_w(C_1, C_2) = \frac{\sum_j w(a_{1j})w(a_{2j})}{\sum_j w(a_{1j})^2 + \sum_j w(a_{2j})^2 - \sum_j w(a_{1j})w(a_{2j})} \tag{1}$$

where $a_{1j} \in A_1$ and $a_{2j} \in A_2$. This is a normalized measure taking on values in $[0, 1]$ where 0 is completely dissimilar and 1 is identical.

We use a weighted Jaccard coefficient to help resolve issues such as the following. Suppose that we have two clusters $C_1 = (\{a_1\}, m)$ and $C_2 = (\{a_1, a_2\}, n)$. Let $w(a_1) = 10$. If $w(a_2) = 1$, we have $J_w(C_1, C_2) = .99$ whereas if $w(a_2) = 10$ the similarity is $J_w(C_1, C_2) = .5$. This reflects the fact that in the former case ($w(a_2) = 1$) it is very likely that the author a_2 collaborated with author a_1 on a single paper in a_1's topic area. In the latter case where both a_1 and a_2 have written a larger number of reports, the lowered similarity reflects our increased uncertainty as to whether C_1 and C_2 should be combined, that is, whether a_1 or a_2 better characterizes C_2 or whether C_2 should just be regarded as a topic by itself.

To determine the best pair of clusters to combine, we calculate the pairwise similarity for all clusters and take the pair with the greatest weighted Jaccard coefficient as the best pair to attempt to combine. Although we use a weighted Jaccard coefficient to determine whether two clusters are sufficiently similar to combine, we use an unweighted Jaccard coefficient[15] to determine whether they should actually be combined. Given two clusters C_1 and C_2, the unweighted Jaccard coefficient, $J_u(C_1, C_2)$, is given by

$$J_u(C_1, C_2) = \frac{|A_1 \cap A_2|}{|A_1 \cup A_2|}. \tag{2}$$

J_u is also a normalized measure with the same range and interpretation as J_w. The unweighted Jaccard coefficient is simply the fraction of authors the two clusters have in common; it is a measure of the joint coauthorship of the two clusters.

Since we are in effect building a classifier, the purpose of introducing this step is to prevent overfitting the classifier to the data by refusing to combine highly similar clusters under some circumstances. The criterion we used is $J_u > .5$, that is, we only combined similar clusters if more than half their authors were in common. The intuition here is that of a core group of authors collaborating on a single project versus a number of common authors participating in two different projects that may be topically unrelated. If they are in fact topically related then the subsequent text clustering should bring them together so this is a conservative step.

If two highly similar clusters are not combined, we choose to no longer consider them in the algorithm. We refer to this process as *freezing* the clusters. The rationale is that due to the order in which we consider clusters, there is no better pairing of C_1 and C_2 possible so if we elect not to combine C_1 with C_2, any other pairing would be inferior.[4] Hence, we choose to retain C_1 and C_2 as provisional topics at this point.

To summarize, the steps involved in the author group (AG) clustering phase at each site are given below.

1. Calculate weights, $w(a)$, for all authors, a, in the site collection.
2. Create seed clusters as described earlier.
3. Calculate the pairwise similarity scores for all seed clusters.
4. Combine clusters as shown by the algorithm of Figure 1.

```
while clusters remain unfrozen
    pick the most similar clusters max J_u(C_i, C_j)
    if J_u(C_i, C_j) > .5
      then merge(C_i, C_j)
      else freeze both C_i and C_j
```

Figure 1: Algorithm for combining clusters.

A word about merging clusters is in order here. When two clusters are merged, we must: (1) select a label for the new cluster; and (2) determine whether the cluster similarities must be recalculated. Generally, when we merge $C_1 = (A_1, m)$ with $C_2 = (A_2, n)$ we get the new cluster $C_3 = (A_1 \cap A_2, m + n)$ where the new label is the joint authors. Whenever, $A_1 = A_1 \cap A_2$ (or $A_2 = A_1 \cap A_2$) it is not necessary to recalculate the similarities since C_3 will have the same similarity with other clusters as did C_1 (or C_2). Whenever we freeze clusters it is also not necessary to recalculate similarities. We only recalculate when the new cluster label is one that has not been used before. The algorithm terminates when all clusters have been frozen.

[4] Technically there may be another cluster C such that $J_w(C, C_1) = J_w(C_1, C_2)$ or $J_w(C, C_2) = J_w(C_1, C_2)$. This complicates the decision to freeze somewhat but only to the degree that each of the other possibilities must be checked before freezing either or both of C_1 and C_2.

To see how this all works consider clustering the documents shown in Table 1. Table 1 shows 25 documents together with their author labels. The complete author group (AG) clustering of these documents is shown in Table 2. Column A of Table 2 shows the resulting 19 seed clusters. The final 11 merged clusters resulting from the AG clustering are depicted in Column B of the figure.

Initial Groups (A)	After AG Clustering (B)	Final Clustering (C)
FJ;PJ;WJ = $\{w\}$	FJ;PJ = $\{h, s, v, w\}$	$\{e, g, h, o, p, s, v, w, y\}$
FJ;JA;PJ = $\{h, v\}$		
FJ;PJ = $\{s\}$		
BP;FJ;GA;JS;KA;PJ... = $\{y\}$	BP;FJ;GA;JS;PJ = $\{p, y\}$	
BP;FJ;GA;JS;LY;PJ... = $\{p\}$		
BP;FJ;KD;OR;PJ;SS = $\{g\}$	BP;FJ;KD;OR;PJ;SS = $\{g\}$	
BJ;FJ = $\{o\}$	BJ;FJ = $\{o\}$	
FJ;GA;PJ = $\{e\}$	FJ;GA;PJ = $\{e\}$	
FJ;GA;KJ = $\{i, n, u\}$	FJ;GA;KJ = $\{i, n, u\}$	$\{i, l, n, t, u\}$
FJ;PT = $\{t\}$	FJ;PT = $\{l, t\}$	
DM;FJ;PT = $\{l\}$		
FJ;VC = $\{b, f, j\}$	FJ;VC = $\{b, f, j\}$	$\{b, f, j, k, m\}$
FJ = $\{k\}$	FJ = $\{k\}$	
FJ = $\{m\}$	FJ = $\{m\}$	
FJ;JJ = $\{d\}$	FJ;JJ = $\{d\}$	$\{d\}$
BJ;FJ;KP;MW = $\{x\}$	BJ;FJ;KP;MW = $\{q, x\}$	$\{q, x\}$
BJ;FJ;KP;MW;WJ = $\{q\}$		
FJ;GA;RP;WA;WW = $\{c, r\}$	FJ;GA;RP;WA;WW;= $\{c, r\}$	$\{c, r\}$
FJ = $\{a\}$	FJ = $\{a\}$	$\{a\}$

Table 2: Illustration of the clustering methodology. We start with raw author groups (column A). Then we treat each group as a whole and cluster them further by common co-authors (column B). These clusters then form the the input to a general hierarchical agglomerative clustering method which produces the final clusters (column C).

3.1.2 Inter-site Clustering

After the documents at each site have been clustered by the coauthor criteria, the clusters are treated as "documents" for a separate clustering phase applied to the words derived from the titles. For the work here we used a group average hierarchic agglomerative clustering method. The terms in each cluster vector were weighted by term frequency and the vectors were normalized. The result of this second phase clustering on the data of Table 1 is shown in Table 2(C). Column C represents the system's view of the topical content; each of the 7 subsets is considered a topic.

4 An Experiment

In the last section we described the mechanics of our approach to content analysis via coauthor clustering. In this section we describe an experiment that we conducted to determine the efficacy of this method.

Cluster ID	Document ID	Author Label	Title
A	1	BP;FJ;KD;OR;PJ;SS	Basic Data Concepts in ADAMS
	2	FJ;PJ;WJ	Scoping Persistent Name Spaces in ADAMS
	3	BP;FJ;GA;JS;KA;...	The ADAMS Database Language
	4	BP;FJ;GA;JS;LY;...	Implementation of the ADAMS Database System
	5	FJ;GA;PJ	An Introduction to the ADAMS Interface Language: Part I
	6	FJ;JA;PJ	Scientific Database Management (Final Report)
	7	FJ;JA;PJ	Scientific Database Management (Panel Reports and Supporting Material)
B	8	FJ;GA;KJ	Breaking the I/O Bottleneck at the National Radio Astronomy Observatory (NRAO)
	9	FJ;GA;KJ	Extensible File Systems (ELFS): An Object-Oriented Approach to High Performance File I/O
	10	FJ;GA;KJ	High Performance Access to Radio Astronomy Data: A Case Study
	11	FJ;PT	Performance Measurement of Two Parallel File Systems
	12	DM;FJ;PT	Performance Measurement of a Parallel Input/Output System for the Intel iPSC/2 Hypercube
C	13	FJ	Electronic Distribution of Technical Reports and Working Papers: A Simple Cooperative Approach
	14	FJ;JJ	An Archive Service with Persistent Naming for Objects
	15	FJ;VC	Availability and Latency of World Wide Web Information Servers
	16	FJ;VC	Dissemination of Collection Wide Information in a Distributed Information Retrieval System
D	17	BJ;FJ	Scalable Database Support for Correlation and Fusion Algorithms
	18	BJ;FJ;KP;MW	System for indexing multi-spectral satellite images for efficient content-based retrieval
	19	BJ;FJ;KP;MW;WJ	Indexing Multispectral Images for Content-Based Retrieval
	20	FJ	Heuristic Clustering of Signature Files: A New Approach (Preliminary Results)
E	21	FJ;VC	A Software Toolkit for Prototyping Distributed Applications (Preliminary Report)
	22	FJ;GA;RP;WA;WW	A Synopsis of the Legion Project
	23	FJ;GA;RP;WA;WW	Legion: The Next Logical Step Toward a Nationwide Virtual Computer
F	24	FJ	A Global Time Reference for Hypercube Multicomputers
G	25	FJ;PJ	Multiple Inheritance and the Closure of Set Operators in Class Hierarchies

Table 3: Hand clustering of 25 technical reports. Author labels are included to show the strong relationship that co-authorship has with topicality.

4.1 Methodology

We chose to compare our approach to conventional clustering methods by means of a clustering experiment. We first took the 25 documents shown in Table 1 and clustered them by hand into topical groups using all the information we had available.[5] The final hand clustering was completed before any of the automated procedures were run. This clustering of the documents is shown in Table 3 and represents "ground truth" for the clustering experiment. This clustering is also shown on the first line of Table 4. Here the document numbers are shown grouped into the topical subsets to which they have been assigned.

Next we clustered the documents using a group average algorithm on the titles of the documents. We show several of the clusterings, including the best that resulted, in Table 4. Finally we ran the group average algorithm over the merged clusters shown in Table 2(B). We chose the best clustering from the run. This clustering is shown in Table 2(C) and as the last line of Table 4.

4.2 Evaluation of Clusterings

To evaluate the quality of each of these clusterings (Table 4), we calculated the "distance" each was from the hand clustering shown in Table 4. This distance calculation took the form of determining how many documents were misclassified by each of the candidate clusterings. The first step in this calculation is to match each subset with the appropriate topic given by the hand clustering. After that it is relatively straightforward to count the number of misclassifications. The misclassified documents are shown in boldface type in Table 4.

Method	Clustering	Nbr. Missed
By Hand	[1,2,3,4,5,6,7] [8,9,10,11,12] [13,14,15,16] [17,18,19,20] [21,22,23] [24] [25]	0
Title Only 1	[1,2,**14**] [3,4,5] [6,7,**17**] [8,9,10,11,12] [13,**20,21**] [15,16,**18,19**] [22,23] [24] [25]	6
Title Only 2	[3,4,5] [6,7,**17**] [**1,2**,8,9,10,11,12,**14**] [13,**20,21**] [15,16,**18,19**] [22,23] [24] [25]	8
Title Only 3	[6,7,**17**] [**1,2,3,4,5**,8,9,10,11,12,**14**] [13,**20,21**] [15,16,**18,19**] [22,23] [24] [25]	11
AG	[1,2,3,4,5,6,7,**17, 25**] [8,9,10,11,12] [13,15,16,**20,21**] [14] [18,19] [22,23] [24]	4

Table 4: Comparison of Author Group (AG) clustering and Title Only clustering to the manually generated clusters. Misplaced documents are in bold. Document numbers correspond to those in the previous table.

As can be seen from Table 4, the best title clustering resulted in 6 document misclassifications with the worst having misclassified 11 documents. The clustering based on document coauthorship considerations resulted in only 4 document misclassifications, a considerable improvement over the best title clustering. While this is only one small example, it does suggest that specific improvement gains are possible and that this approach is deserving of further study.

We recognize that a sample of 25 documents is far too small to generalize from and are in the process of conducting larger experiments. The main problem is the difficulty in acquiring hand classifications for complete sites with many hundreds of documents. We included this small example for two reasons. First, it is small enough to demonstrate the mechanics of the approach.

[5] These 25 documents were chosen because they were all coauthored by one of the authors of this paper. While we understand that this is a biased sample, we took every precaution, including having them hand clustered by others, to assign them objectively to subject areas.

Second, it does in fact represent a fairly difficult problem. The small size does not allow for much improvement. The fact that we were able to obtain improvement is important. We do recognize the need for much more thorough testing though.

5 Conclusions and Future Directions

We have described a novel technique for augmenting conventional document clustering to improve subject classification. The purpose of the proposed algorithm is to map the topical content of distributed IR systems which are expected to form the backbone of digital libraries. Knowledge of the topical content distribution can be used to assess content-skew for the purpose of communication management within the system. The specific methodology for achieving this is the subject of a forthcoming paper. The knowledge can also be used to provide adjunct retrieval services such as browsing, and improve the efficiency of query processing by pruning sites.

These early results are very encouraging and we are now in the process of fully automating the procedures so that they can be applied to the complete NCSTRL collection. The coauthorship algorithm evaluated above is the simplest that we have considered. There are several improvements that we think may be necessary before running against the entire collections. Refinements may be necessary in three areas: the cluster combining rules; the relabeling rules; and the similarity measure. There is also work to be done on the algorithm termination criterion.

There is ample evidence in the current study to suggest that major gains in topical clustering accuracy are possible when coauthorship information is employed. One difficulty with upcoming work will be quantifying the degree of that gain. We will need to acquire hand clusterings for several complete NCSTRL sites in order to mount a large-scale evaluation effort.

We are also interested in topic hierarchies for use in retrieval applications. The approach described here offers promise for automatically generating subject classifications that can be usefully employed in that capacity.

References

[1] W. Y. Arms and C. R. Arms. Cluster Analysis Used on Social Science Journal Citations. *Journal of Documentation*, 34(1):1–11, March 1978.

[2] Julie Bichteler and Edward A. Eaton III. The Combined Use of Bibliographic Coupling and Cocitation for Document Retrieval. *Journal of the American Society for Information Science*, 31(4), July 1980.

[3] Mark P. Carpenter and Francis Narin. Clustering of Scientific Journals. *Journal of the American Society for Information Science*, 24(6), Nov.-Dec. 1973.

[4] Bruce W. Croft. Clustering Large Files of Documents Using the Single-link Method. *Journal of the American Society for Information Science*, 28(6):341–344, November 1977.

[5] J. R. Davis. Creating a Networked Computer Science Technical Report Library. *D-Lib Magazine*, (September), 1995.

[6] Edward A. Fox, Robert M. Akscyn, Richard K. Furuta, and John J. Leggett (editors). Special Issue on Digital Libraries. *Communications of the ACM*, 38(4), 1995.

[7] James C. French. DIRE: An Approach to Improving Scientific Communication. *Information and Decision Technologies*, 19:527–541, 1994.

[8] James C. French and Charles L. Viles. Ensuring Retrieval Effectiveness in Distributed, Digital Libraries. *To Appear in Journal of Visual Communication and Image Representation*, 1995.

[9] E. Garfield. *Citation Indexing: Its Theory and Application in Science, Technology and Humanities.* John Wiley, New York, 1979.

[10] R. Green. Topical Relevance Relationships. I. Why Topic Matching Fail. *JASIS*, 46(9):646–653, 1995.

[11] Anil K. Jain and Richard C. Dubes. *Algorithms for Clustering Data.* Prentice Hall, Englewood Cliffs, NJ, 1988.

[12] Karen Sparck Jones. Notes and References on Early Automatic Classification Work. *SIGIR Forum*, 25(1):10–17, 1991.

[13] M. M. Kessler. Bibliographic Coupling Between Scientific Papers. *American Documentation*, 14(1):10–25, January 1963.

[14] K. L. Kwok. The Use of Title and Cited Titles as Document Representation for Automatic Classification. *Information Processing and Management*, 11(8.12):201–206, 1975.

[15] Gerard Salton. *Automatic Text Processing.* Addison Wesley, 1989.

[16] Gerard Salton. Automatic Indexing Using Bibliographic Citations. *Journal of Documentation*, 27(2):98–110, June 1971.

[17] Gerard Salton. Associative Document Retrieval Techniques using Bibliographic Information. *Journal of the ACM*, 10(4):440–457, October 1963.

[18] H. Small. Co-citation in the Scientific Literature: a New Measure of the Relationship between Two Documents. *JASIS,*, 24:265–269, 1973.

[19] Henry G. Small and Michael E. D. Koenig. Journal Clustering Using a Bibliographic Coupling Method. *Information Processing and Management*, 13(5):277–288, 1977.

[20] Charles L. Viles and James C. French. Dissemination of Collection Wide Information in a Distributed Information Retrieval System. In *Proceedings of the 18th International Conference on Research and Development in Information Retrieval (SIGIR95)*, pages 12–20, Seattle, WA, July 1995.

[21] E. M. Voorhees. Implementing Agglomerative Hierarchic Clustering Algorithms for Use in Document Retrieval. *Information Processing and Management*, 22(6):465–476, 1986.