Effective Document Clustering for Large Heterogeneous Law Firm Collections

Jack G. Conrad and Khalid Al-Kofahi
Research & Development Department
Thomson Legal & Regulatory
St. Paul, Minnesota 55123 USA
{Jack.G.Conrad, Khalid.Al-Kofahi}@Thomson.com

Ying Zhao and George Karypis
Computer Science & Engineering Department
University of Minnesota
Minneapolis, Minnesota 55455 USA
{yzhao, karypis}@cs.umn.edu

ABSTRACT

Computational resources for research in legal environments have historically implied remote access to large databases of legal documents such as case law, statutes, law reviews and administrative materials. Today, by contrast, there exists enormous growth in lawyers' electronic work product within these environments, specifically within law firms. Along with this growth has come the need for accelerated knowledge management—automated assistance in organizing, analyzing, retrieving and presenting this content in a useful and distributed manner.

In cases where a relevant legal taxonomy is available, together with representative labeled data, automated text classification tools can be applied. In the absence of these resources, document clustering offers an alternative approach to organizing collections, and an adjunct to search.

To explore this approach further, we have conducted sets of successively more complex clustering experiments using primary and secondary law documents as well as actual law firm data. Tests were run to determine the efficiency and effectiveness of a number of essential clustering functions. After examining the performance of traditional or hard clustering applications, we investigate soft clustering (multiple cluster assignments) as well as hierarchical clustering. We show how these latter clustering approaches are effective, in terms of both internal and external quality measures, and useful to legal researchers. Moreover, such techniques can ultimately assist in the automatic or semi-automatic generation of taxonomies for subsequent use by classification programs.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Clustering*; H.3.m [**Information Storage and Retrieval**]: Miscellaneous—*Legal Test Collections*; I.5.3 [**Pattern Recognition**]: Clustering—*Algorithms*

General Terms

Algorithms, Experimentation, Measurement

Keywords

knowledgement management, document clustering, taxonomy development, legal data

ICAIL '05 June 6-11, 2005, Bologna, Italy

1. INTRODUCTION AND MOTIVATION

Possessing unlimited access to large legal data repositories and associated search and analysis tools, such as those offered by Thomson-West or Lexis-Nexis, is no longer sufficient for legal practitioners who need to retain both their broad coverage and productivity in their practice of law. The center of professional knowledge management in today's legal field rests within the law firm. In this environment, *knowledge management (KM)* can be broadly defined as the identification and management of processes for leveraging the intellectual capital of the law firm over time and across sites [11]. Among its primary goals are improved efficiency and productivity for legal practitioners achieved through the reuse and sharing of expertise and associated work-products [20]. The ultimate goal is elevated knowledge-sharing among a highly skilled set of knowledge professionals.

Many firms currently deploy systems that offer their practitioners both full-text search capabilities and the ability to browse through a general legal classification system such as KeySearch [7], which is based on common legal practice or research areas (e.g., bankruptcy, intellectual property, torts & personal injury). Useful as such a broad-ranging taxonomy is, however, there are numerous instances when it is rendered incomplete if not irrelevant. For example, some firms offer fewer legal practice areas, but among those provided, they offer them at substantial depth. A broad-based, but shallow taxonomy would thus be inadequate for such expertise. Other examples include those firms that are now practicing areas of law that may not be well covered by an existing taxonomy (e.g., elder law, law involving abortion, gay marriage law, or law involving anti-terrorism). Still other firms find such taxonomies only partially relevant to their portfolio of practice areas. At least in the context of automatic text categorization, the strength of these firms' KM applications will depend upon the relevance of their taxonomies and the availability of exemplar documents for each category. So clearly the incompatibilities outlined above are problematic. In the absence of a relevant taxonomy and corresponding training data, an effective and intelligent document clustering application could prove to be useful [22].

The remainder of this paper is organized as follows: Section 2 reviews previous work in both the legal knowledge management space and in the field of document clustering. Section 3 introduces the concepts of hard versus soft clustering applications. In Section 4, we discuss the data and clustering resources we leverage to conduct our trials. Section 5 describes our experimental methodology and evaluation metrics. In Section 6, we present our experimental results and discuss their significance. We draw our conclusions and address future work in Sections 7 and 8.

2. RELATED RESEARCH

2.1 Knowledge Management for Legal Environments

An appreciable amount of work has addressed the general topic of knowledge management for legal environments, including reports produced within the AI & Law community [33, 26, 37, 3]. The works by Oskamp, et al. and Visser, et al., cited above, share with our work a study of AI tools and organizational structures to facilitate KM within law offices.

In a White Paper that stresses the KM value proposition for law firms, Tziahanas focuses on what he claims are the two fundamental needs under consideration: *information asset management* and *organizational requirements* [35]. He states that a coordinated response to the two disparate issues holds the key to eventual knowledge management success. The current challenge is thus to find a means to organize and share legal knowledge within the firm with minimal impact on existing processes. The author declares that this goal requires an organizational commitment to share and reuse information. The prospective result will be that the implementation cost of KM applications will be significantly decreased and disruptive effects can be avoided. He emphasizes the role that taxonomies, thesauri, metadata and associated formats like RDF will play in future KM solutions.

Edwards and Mahling describe a typical large law firm environment in terms of its tasks, structure, people, and technology. They then propose a taxonomy for classifying types of knowledge a large firm possesses and make analytical observations about the characteristics of a firm's knowledge which are relevant to knowledge management [10]. Based on those characteristics, they identify a set of high-level system and user specifications which can serve as a basis for developing usable tools for KM in a large law firm.

In presenting several models of Knowledge Management, Terrett argues that knowledge is one of the only meaningful assets in the legal workplace [34]. He defines the *strategy* behind knowledge management as:

> What every law firm seeks ... is a methodology that allows the systematic capture, development and use of legal ... knowledge, together with the development of an internal knowledge market which provides incentives and rewards for knowledge creators to share their knowledge and the promotion of a corporate culture which demonstrates the benefits of corporate knowledge sharing.

The author articulates three distinct models of knowledge management within firms, from the simple organizational learning loop within the firm (via a basic knowledge base), to an intellectual capital model that differentiates between human and structural capital, finally to a model based on explicit vs. tacit knowledge and their role within a firm (first formulated in [24]). Regarding this last model, he describes four modes of interplay between explicit and tacit knowledge that are created: socialization (e.g., via apprenticeships), externalization (i.e., expressing tacit knowledge via explicit concepts), combination (i.e., systemizing concepts into a knowledge system), and internalization (i.e., explicit to tacit, e.g., such as from reading a text). His main observation is that in order for law firms of the future to continue to be successful, they need to evolve from their past one-dimensional document-based information management approach to knowledge management in a broader sense, in which legal professionals are encouraged to contribute and exchange knowledge using appraisal and compensated practices that instill fundamental changes to their former billing-preoccupied model.

One central observation one can make about the majority of academic work performed in the knowledge management space for law firms is that with few minor exceptions [30], much has been done to answer the question of *what*, but little has been reported on thus far in terms of *how*. M. Ethan Katsh, who has written considerably on how digital technology will impact the legal practice, quotes litigator Fred Bartlit as stating that "Most of what lawyers do is store, categorize, organize ... and analyze data." Despite its major influence on other fields, digital technology has "hardly made a dent in how lawyers do their jobs" [16]. Subjective as this statement may be, we can assert that the essential components of any firm-based knowledge management system would likely include full-text search (Boolean, Fielded, Natural Language), classification (when a relevant taxonomy exists), clustering (when it doesn't) and possibly vetting (to purge duplicate or low content-bearing documents). The focus of this paper is on document clustering in those legal environments where complete and reliable taxonomies and labeled data do *not* exist, acknowledging that other applications exist within large law firms today that provide search and/or classification services [2], not to mention current awareness alerts. As such, this is the first work of its kind that explores and evaluates for effectiveness the organization of law firm documents via the strategies described below.

2.2 Document Clustering

Fast and high-quality document clustering algorithms play an important role in helping users to navigate, summarize, and organize an enormous amount of text available on the Internet, and in digital libraries, news sources, and company-wide intranets. Over the years a variety of different algorithms has been developed. These algorithms can be categorized along different dimensions based either on the underlying methodology of the algorithm, leading to *agglomerative* [31, 17, 12, 13] or on *partitional* approaches [19, 6, 40], or on the nature of the membership function, leading to *hard (crisp)* or *soft (fuzzy)* [23, 18, 4, 38, 5, 21, 29, 25] solutions.

2.2.1 Hard Clustering

Hard clustering solutions can be obtained by both partitional and agglomerative clustering algorithms.

Partitional clustering algorithms compute a k-way clustering of a set of objects either directly or via a sequence of repeated bisections. A direct k-way clustering groups the data set into k subsets that optimize a desired clustering criterion function. A k-way partitioning via repeated bisections is obtained by recursively partitioning one of the existing clusters into two clusters (i.e., bisections). Partitional clustering can be viewed as an optimization procedure that tries to create high-quality clusters according to a particular criterion function. Criterion functions used in partitional clustering reflect the underlying definition of the "goodness" of clusters. Research on partitional clustering algorithms has been focusing on developing both clustering criterion functions [14, 9, 40] and optimization methods [9].

Agglomerative algorithms find the clusters by initially assigning each object to its own cluster and then repeatedly merging pairs of clusters until a certain stopping criterion is met. The three basic criteria to determine which pair of clusters to be merged next are single-link [31], complete-link

[17] and group average [14]. In addition to these three basic approaches, a number of more sophisticated schemes have been developed, like CURE [12] and ROCK [13], that were shown to produce superior results.

In recent years, various researchers have recognized that partitional clustering algorithms are well-suited for clustering large document data sets due to their relatively low computational requirements [8, 1, 32]. In many document clustering applications, hierarchical clustering solutions are desirable, solutions which have been traditionally obtained using agglomerative algorithms. However, partitional algorithms can also be used to obtain hierarchical clustering solutions via a sequence of repeated bisections. A recent study [39] also showed that partitional algorithms always lead to better hierarchical solutions than agglomerative algorithms on various document data sets, making them ideal for clustering large document collections due not only to their relatively low computational requirements, but also to higher clustering quality.

For clarity and consistency throughout the remainder of this work, we refer to *hard clustering* as the assignment of a candidate document to a single cluster resulting from the partitioning of a collection. By contrast, we refer to *soft clustering* as the ability to assign a document to *multiple* resulting clusters.

2.2.2 Soft Clustering

Soft clustering that allows an object to appear in multiple clusters has been studied extensively and still remains a challenging problem. The reason why soft clustering is important in the legal domain is because legal 'taxonomies' are not composed of mutually exclusive categories, and legal documents tend to be multi-topical. In recent years, soft clustering algorithms have been studied in document clustering contexts and shown to be effective in finding overlapping clusters [23, 18]. The fuzzy C-means algorithm [4] is one of the most widely used soft clustering algorithms. It is a soft version of the K-means algorithm that uses a soft membership function. Other newly developed soft clustering algorithms differ from fuzzy C-means by employing different dissimilarity functions [23], or by including both a soft membership function and a weight function (measuring the contribution of each object in a fuzzy cluster) in the criterion functions (K-harmonic means [38]). However, fuzzy C-means and algorithms of this type have a well-known problem of their dependency on initialization [9] and are not suitable to cluster data sets containing tens of thousands of legal documents.

Another widely used approach to produce soft clustering solutions is the well-known Expectation-Maximization (EM) algorithm [5, 21, 29, 25], which generates classification probabilities for each document. The EM algorithm is a general statistical method of maximum likelihood estimation and has a strong statistical basis. However, the classical EM algorithm may have difficulty converging or may converge to an undesired solution. Reasons for this behavior may include a data set that is high-dimensional or an initial solution that is not carefully generated [25]. Even though many approaches have been proposed to improve EM, (e.g., FREM [25], On-line EM [29], and Scalable EM [5]), the EM-based approaches still cannot guarantee convergence to high-quality clustering solutions efficiently on large volumes of legal document data sets.

3. HARD AND SOFT CLUSTERING ALGORITHMS

3.1 Problem Definition

Within a law firm environment, the organization of associated legal document collections in an efficient and effective manner implies the ability to perform a number of essential document "clustering" functions that result in the following:

- topically-useful groupings;
- hierarchical groupings (algorithms can operate iteratively on more than one level);
- multiply assigned groupings (documents can belong to more than a single group);
- grouping process operates in a reasonable time and can scale to large collections.

Our research has encompassed each of these separate problem areas, and we have tested our clustering technology on corpora containing a quarter-million documents. Yet given the number of parameters involved in each of these areas, and the space limitations of this report, in this work, we report on experiments primarily involving the first three functions described above, with particular emphasis on multi-topical documents that merit membership consideration in more than a single resultant cluster (i.e., soft clustering).

3.2 Challenges

There exist several distinct challenges involved in clustering legal documents in the law firm environment. In addition to challenges of scale (clustering potentially millions of electronic documents in a DMS) and efficiency (indexing and organizing large collections in hours rather than days), there exists the complex nature of many legal documents (that involve more than a single legal topic and can regularly pertain to numerous topics). Take, for example, a legal matter involving a client's misuse of pension funds, subsequent securities investments, filing for bankruptcy and associated perjury. Lawyers and IT authorities alike acknowledge that in the legal domain, it is the rule, rather than the exception, that these documents are multi-topical [22]. And unlike in World Wide Web contexts where search and categorization services are typically free, the legal profession is a fee-for-service field where the demand for reliable IT (e.g., in terms of both precision and recall) is expected. For this reason, we place a significant emphasis in this work on both the qualitative as well as quantitative and the computational as well as human-reviewed performance dimensions of our experiments.

3.3 Preliminaries on Document Modeling

We represent the documents in our legal collections using the vector-space model [28]. In this model, each document d is considered to be a vector in the space of the distinct terms present in the collection. We employ the *tf-idf* term-weighting scheme that represents each document d as the vector.

$$d_{tfidf} = (tf_1 \log(\frac{n}{df_1}), tf_2 \log(\frac{n}{df_2}), ..., tf_m \log(\frac{n}{df_m})) \quad (1)$$

In this scheme, tf_i corresponds to the frequency of the ith term in the document and $idf_i = \log(n/df_i)$ corresponds to its inverse document frequency in the collection (df_i is the number of documents that contain the ith term). To account

for documents of different lengths, we scale the length of each document vector so that it is of unit length.

We measure the similarity between a pair of documents d_i and d_j by taking the cosine of the angle formed between the *tf-idf* representation of their vectors. Specifically, this is defined as

$$\cos(d_i, d_j) = \frac{{d_i}^t d_j}{\|d_i\|\|d_j\|}, \quad (2)$$

which can be simplified to $\cos(d_i, d_j) = {d_i}^t d_j$, since the document vectors are of unit length. This similarity measure becomes one if the document vectors point in the same direction (i.e., they contain identical set of terms in the same relative proportion), and zero if there is nothing in common between them (i.e., the vectors are orthogonal to each other).

3.4 Hard Clustering Algorithms

Our hard clustering algorithms are partitional in nature. A key feature in these algorithms is that they treat the clustering problem as an optimization process which seeks to maximize or minimize a particular *clustering criterion function* defined either globally or locally over the entire clustering solution space. The various hard clustering algorithms are available from `CLUTO` [15] (Section 4.2 provides additional information on this clustering toolkit). Recent studies have shown that these algorithms are efficient and effective in producing both hierarchical and non-hierarchical clustering solutions [39, 40].

`CLUTO` provides a total of seven different criterion functions that have been shown to produce high-quality clusters in low- and high-dimensional data sets, which optimize various aspects of intra-cluster similarity, inter-cluster dissimilarity, and their combinations. The mathematical definitions of the two clustering criterion (or objective) functions we used in our experiments are shown in equations (3) and (4), and they are derived and analyzed in [39, 40].

$$\mathcal{I}_1 = \text{maximize} \sum_{i=1}^{k} \frac{1}{n_i} \left(\sum_{v,u \in S_i} \text{sim}(v,u) \right) \quad (3)$$

$$\mathcal{I}_2 = \text{maximize} \sum_{i=1}^{k} \sqrt{\sum_{v,u \in S_i} \text{sim}(v,u)} \quad (4)$$

The notation used in equations (3) and (4) is as follows: k is the total number of clusters, S_i is the set of objects assigned to the ith cluster, n_i is the number of objects in the ith cluster, v and u represent two objects, and $\text{sim}(v,u)$ is the similarity between two objects.

An important aspect of partition-based, criterion-driven clustering algorithms is the method used to optimize their criterion function. `CLUTO` uses a randomized incremental optimization algorithm that is greedy in nature, has low computational requirements, and produces high-quality clustering solutions [40]. Our greedy optimizer computes the clustering solution by first obtaining an initial k-way clustering and then applying an iterative refinement algorithm to further improve it. During initial clustering, k documents are randomly selected to form the *seeds* of the clusters and each document is assigned to the cluster corresponding to its most similar seed. The iterative refinement strategy that we use is based on the *incremental* refinement scheme described in [9]. During each iteration, the documents are visited in a random order and each document is moved to the cluster that leads to the highest improvement in the value of the criterion function. If no such cluster exists, then the document does not move. The refinement phase ends, as soon as an iteration is performed in which no documents were moved between clusters.

We use a two-step clustering method, referred to as the "`rbr`" method, to utilize the optimization procedure described above [15, 40]. In the first step, a k-way partitioning via repeated bisections is obtained by recursively applying the optimization procedure to compute 2-way clustering (i.e., bisections). Initially, the objects are partitioned into two clusters, then one of these clusters is selected and is further bisected, and so on. This process continues $k-1$ times, leading to k clusters. Each of these bisections is performed so that the resulting two-way clustering solution optimizes a particular criterion function. In the second step, we apply the incremental refinement procedure again on the k clusters obtained from repeated bisections, where each document can be moved to all $k-1$ clusters. At the end of this step, we obtain a k-way clustering solution that optimizes a particular criterion function. Note that we perform $k-1$ repeated bisections before a k-way refinement to ensure a more balanced clustering solution.

3.5 Soft Clustering Algorithms

We harness two efficient soft clustering algorithms: *segment clustering* and *case clustering*. In addition to an efficiency advantage over other traditional soft clustering algorithms, our soft clustering algorithms can also control the maximum number of clusters each document can belong to. In the remainder of this section, we describe the two soft clustering algorithms in detail.

3.5.1 Segment Clustering

The first clustering approach relies on the segmentation of legal documents, which results in segments presumed to correspond to the topical areas within those legal documents. Following such segmentation, fast hard clustering algorithms are applied to those segments. A document is assigned to a cluster if one of its segments is assigned to that cluster by the hard clustering algorithm. Since the segments from the same document may be placed into different clusters, we indirectly assign the original document to multiple clusters.

Note that to a large extent segment clustering relies on an effective segmentation system. Developing such a system is a part of our on-going research. In this study, the focus is on evaluating the resulting clusters.

3.5.2 Case Clustering

The second soft clustering approach contains two stages: an initial hard assignment stage and a multiple assignment stage. The basic idea is to first obtain an initial hard clustering solution, and then consider which of the additional resultant clusters a given document might participate in. The pseudo-code for the case clustering algorithm is shown in Figure 1.

There are two steps to select the clusters that a given document is assigned to:

1. The clusters with an internal Z-score greater than a certain threshold are selected [line (C) in Figure 1];
2. The candidate clusters from step (1) are sorted by intra-cluster similarities, and the tightest (high intra-cluster similarity values) clusters are selected [line (D)].

```
Given a document set, D, consisting of n documents,
D = {d1, d2, d3, ..., dn}, and an initial cluster set, S,
consisting of k clusters, S = {S1, S2, S3, ..., Sk},
produced by a hard clustering algorithm,

  for (i = 1 to n), consider di for a multiple cluster asmt. (A)
    {
    if (di warrants multiple assignments)                  (B)
      {
      for (j = 1 to k)
        {
        Calculate izscore(i, j);

        if (izscore(i, j) > threshold)                     (C)
          {
          Add j to candidate list of clusters;
          }
        }
      Sort candidate list based on intra-cluster similarity; (D)
      Assign di to MaxNum clusters from the list;
      }
    }
```

Figure 1. Pseudo-code for Case Clustering Algorithm

Given document i and cluster j, the internal Z-score, $izscore(i, j)$, is calculated as follows: If document i is part of the cluster, the internal Z-score is defined as the average similarity between document i and all other objects in cluster j normalized by the mean and standard deviation of the same average similarities of all the documents in cluster j. If document i is not part of the cluster, the internal Z-score is defined as the average similarity between document i and all the documents in cluster j normalized by the mean and standard deviation of the same average similarities of all documents that are not in cluster j.

One of the important parameters of the case clustering algorithm is the internal Z-score threshold, which determines the candidate clusters that are valid for future membership consideration. If the threshold is too high, there will be a small number of cases that have candidate clusters satisfying this threshold. As a result, we only make multiple assignments for a few cases. On the other hand, if the threshold is too low, the candidate clusters may not have sufficient quality, and thus we may make wrong assignments. If we evaluate soft clustering solutions using the F-score, increasing the threshold value will improve precision but degrade recall.

4. RESOURCES

4.1 Data Sets

The approach we took to constructing and researching the three principal data sets reported on in this paper is as follows. We began with readily available in-house legal collections of case law, law reports (ALR) and law review (JLR) documents in order to convince ourselves that clustering in general and our clustering toolkit in particular performs reasonably well. Once we met this objective, we proceeded to secure two other types of corpora, those that permitted us to compare the reliability of case clustering with segment clustering (50-Topics Headnotes), and those that facilitated the assessment of new soft-clustering techniques on actual law firm data (LRC). A more detailed description of these collections follows.

4.1.1 Traditional Legal Data Collection

Our first data set consists of 951 (of originally 1,000) "traditional" legal documents made available from Thomson–West. The goal was to assemble a relatively diverse set of legal documents, both in terms of topics (roughly a dozen) and document-types (at least three), which would arguably simulate the type of variance one might expect to encounter within a law firm's document repository, in both form and content. The composition of the collection is shown in Table 1. Approximate percentages for the participating content of this pseudo-law firm collection are included in the second column. The 11 topics represented in the traditional legal collection are the starred entries shown in Figure 2. In order to make this clustering problem more of a challenge, the distribution of topical classes throughout this collection was skewed. Ten of the classes were similar in distribution, whereas Bankruptcy topics represent nearly one-half of the entire document set.

Document-type	**Target Percentage**	**Actual Doc Count**
Case Law	70%	684
w/ headnotes	60%	586
w/o headnotes	10%	98
American Law Reports	15%	132
Journals & Law Reviews	15%	135
Total	100%	951

Table 1: Doc-type Distribution in Initial Legal Corpus

4.1.2 LRC Collection

The LRC collection represents a corpus of facsimile law firm documents representing briefs, memoranda, and letters to clients. In some instances the letters are represented as short digital transmittals. The list of LRC topics is shown in Figure 2. Their distribution by document-type is presented in Table 2, and by topical class in Figure 3.

1. Administrative Law
2. Alternative Dispute Resolution
3. Antitrust & Trade Regulation
4. Art, Entertainment & Sports Law
5. Bankruptcy (*)
6. Business Organizations
7. Civil Procedure (*)
8. Civil Remedies
9. Civil Rights (*)
10. Commercial Law & Contracts (*)
11. Communications (*)
12. Conflict of Laws
13. Constitutional Law (*)
14. Construction Law
15. Criminal Justice
16. Education
17. Elections & Politics
18. Employment Law (*)
19. Energy & Utilities
20. Environmental Law
21. Family Law
22. Finance & Banking
23. Government
24. Health (*)
25. Immigration Law
26. Insurance
27. Intellectual Property
28. International Law
29. Legal Services (*)
30. Maritime Law
31. Products Liability (*)
32. Professional Malpractice (*)
33. Property
34. Science, Computers & Technology
35. Securities Law
36. Taxation
37. Torts & Personal Injury
38. Transportation
39. Wills, Trusts & Estate Planning

Figure 2. List of Topics for the LRC Collection

Document-type	**Percentage**	**Doc Count**
Briefs	28%	1,265
Memoranda	39%	1,762
Letters	33%	1,490
Total	100%	4,517

Table 2: Doc-type Distribution in LRC Corpus

We also performed a series of experiments involving hierarchical clustering on a portion of this set, which is illustrated in Figure 4.

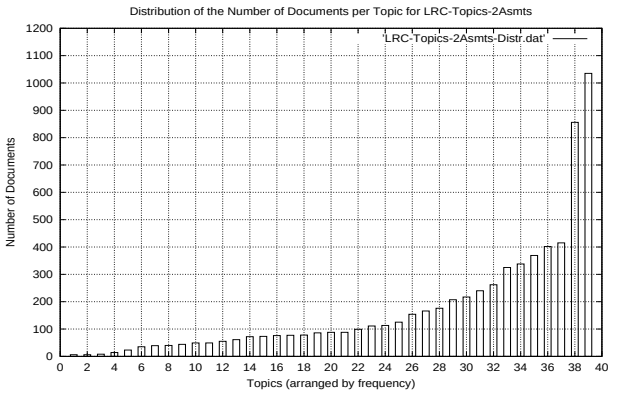

Figure 3. Distribution of the Number of Documents per Topic for the LRC Collection (sorted by frequency)

The total number of topical classes associated with the complete LRC collection is shown in Table 3, along with those for the other test collections. It is worth observing that the underlying topical classes referred to in Table 3 ("Asso. Topics") all derive from the same global KeySearch taxonomy [7], which at the top-level possesses 50 core "topics."

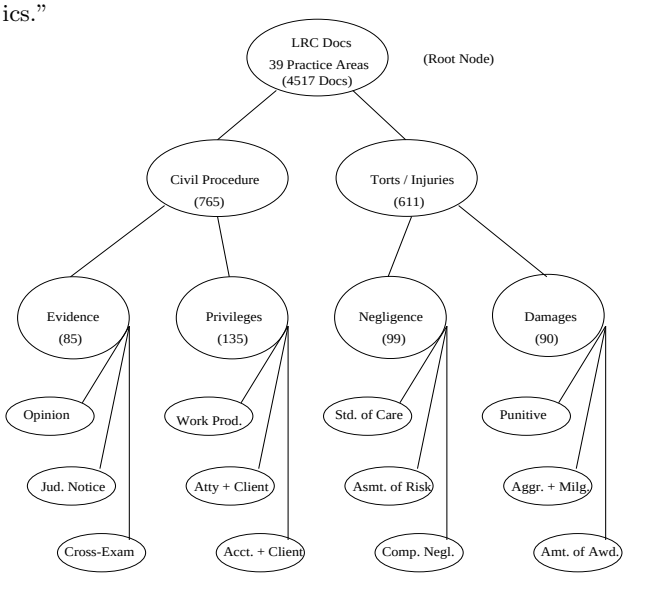

Figure 4. Partial Hierarchical Representation of the LRC Collection including Sub-collection Document Counts

4.1.3 *Topic-based Legal Data Collection*

This data set was derived from a larger U.S. Case Law collection containing over 4 million cases. Each case contains annotated "headnotes" (i.e., summary points of law) classified by human experts to the West Key No. System.[1]

The 50Topics data set consists of 5000 cases whose headnotes are from 50 selected topics. In particular, we focused on two subsets from the 50Topics data set in order to test the effectiveness of our soft clustering algorithms when the maximum number of topics per case is two or three. To this end, we created two data subsets, 50Topics-2HNs and 50Topics-3HNs, where the cases in the data set contain two and three headnotes respectively. The distribution of the number of cases per topic for the first of these two data sets is shown in Figure 5. The distribution for the second data set closely resembles that shown in Figure 5.

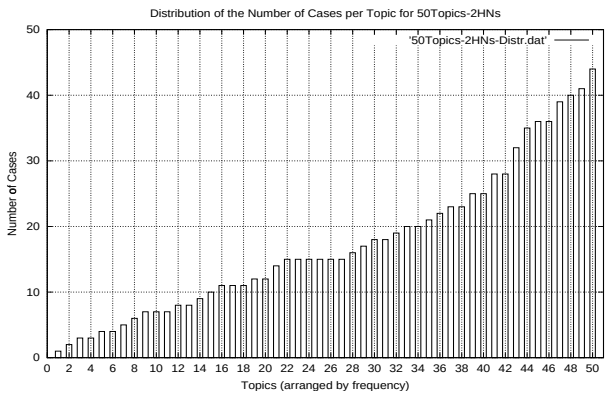

Figure 5. Distribution of the Number of Cases per Topic for 50Topics-2HNs

4.1.4 *Comparative Sizes of Legal Collections*

The total number of documents, tokens and unique tokens (excluding numbers), as well as associated KeySearch topical classes for the data sets is shown in Table 3.

4.2 Clustering Toolkit

This section presents a brief overview of `CLUTO` (release 2.1), a software package for clustering low- and high-dimensional data sets and for analyzing the characteristics of the various clusters. It was designed by the University of Minnesota's algorithms group and is available at `www.cs.umn.edu/~karypis/cluto`. `CLUTO` has been developed as a general purpose clustering toolkit.

`CLUTO`'s distribution consists of both stand-alone programs (`vcluster` and `scluster`) for clustering and analyzing these clusters, as well as a library through which an application program can access directly the various clustering and analysis algorithms implemented in `CLUTO`.[2]

5. EXPERIMENTAL METHODOLOGY AND METRICS

5.1 Experimental Methodology

In order to determine the effectiveness and reliability of document clustering on law firm collections, we designed three sets of experiments with two distinct phases for each. The three sets of experiments focus on legal applications which (1) evaluate hard clustering, (2) evaluate and compare soft clustering approaches (i.e., segment and case clustering algorithms), and (3) conduct a parameter study on our internal Z-score threshold while harnessing the F-score metric and its β parameter. The *first phase* of these experiments focuses on pseudo-law firm data and is designed to indicate whether or not clustering technology can perform dependably when applied to corpora in the legal domain. It also indicates whether added experiments are worth pursuing in an operational setting. These tests are conducted on our Traditional Legal (Case+JRL+ARL) and 50Topics collections. In the *second phase*, we verify whether earlier

[1] Each Key Number classification consists of one topic identifier and one key number identifier, which correspond to the first level and the leaf note level in the Key Number hierarchy, respectively.

[2] To date, `CLUTO` has been successfully used to cluster data sets arising in many diverse application areas including information retrieval, commerce and science (e.g., biological applications).

Data Set	Doc Cnt	Total Terms	Unique Terms	Associated Topics	Average No. of Topics per Doc
Case+ALR+JLR	951	769,019	46,980	11	—
50-Topics-2HNs	533	19,391	3,431	50	1.57
50-Topics-3HNs	756	32,614	4,362	50	2.44
LRC (Law Firm)	4,517	1,445,669	49,975	39	1.25

Table 3: Statistics for Four Test Collections Used

results from hard and soft topical clustering extend to actual law firm document collections such as LRC.

5.1.1 Hard Clustering Experiments

The first set of experiments focuses on applying hard clustering techniques to our Traditional Legal and Law Firm corpora. In phase 1, we apply hard clustering algorithms to our Case-ALR-JLR collection to generate 11 or more clusters for 11 classes (which include several distinct sub-classes of bankruptcy). In phase 2, we apply these hard clustering algorithms to the LRC collection, with ± 39 clusters for as many classes at the top (root) node (representing 4,517 documents), and with fewer clusters for comparable classes at lower nodes in the KeySearch hierarchy (with sizes of document sets indicated in parentheses in Figure 4).

5.1.2 Soft Clustering Experiments

The second set of experiments compares our two soft clustering algorithms. In the phase 1, we analyze the performance of case clustering relative to segment clustering when applied to the 50Topics data set. In phase 2, attempt to extend these results via applications to the LRC data set.

Phase 1: *Segment Clustering*—Since we have segment information for the 50Topics data set, we evaluate the performance of segment clustering on the two associated corpora. Of the 5,000 cases covering 50 specific topics, 533 of these cases contain exactly two headnotes, while 756 of these cases contain exactly three headnotes (Table 3). We use headnotes to represent the segments of the associated case law documents. In the first part of this test, we use those cases with two headnotes, 50Topics-2HNs; in the second part, we use those cases with three headnotes, 50Topics-3HNs. *Case Clustering*—we also test case clustering against the 50Topics corpora. In this experiment, we keep the sets of two and three headnotes from the given cases "bundled" together and apply the case clustering algorithm. As described in Section 3.5.2, this consists of first obtaining an initial hard clustering solution, and then considering which additional resultant clusters a given document might participate in. For 50Topics-2HNs, we limit the number of clusters to which a document may be assigned to two; for 50Topics-3HNs, the limit is three. After the experiments have been conducted, we are able to compare the performance of segment clustering and case clustering, relative to the original topic numbers of the associated headnotes. Note that for both data sets, we use our two soft clustering algorithms to generate 50 clusters for 50 topics (classes).

Phase 2: We proceed to examine the performance of soft clustering on the LRC collection. Since the segmentation of this collection is not available, in this study we focus on evaluating the performance of case clustering. We use the case clustering algorithm to generate 39 clusters for 39 classes and limit the maximum number of clusters to which a document may be assigned to two.

5.1.3 Parameter Inspection

In our third set of experiments, we conduct a parameter study on the soft clustering internal Z-score threshold (described in Section 3.5.2). In phase 1, we examine the behavior of the threshold in conjunction with the 50Topics data sets, while in phase 2, we compare this behavior with that for the LRC data set. This investigation is undertaken by tracking the performance, in terms of the F-score metric (described in Section 5.2.3), against the internal Z-score. Recall that the level of this threshold regulates the number of cluster candidates for multiple assignments after the initial hard clustering portion of the algorithm has been executed.

5.2 Metrics

To evaluate the performance of the various clustering algorithms, we employ (1) metrics that utilize the information provided to the clustering algorithms (i.e., **internal metrics**); (2) metrics that utilize *a priori* knowledge of the classification information of the data set (i.e., **external metrics**); and (3) assessments by legal researchers. We describe these next.

5.2.1 Internal Metrics for Hard Clustering

The basic idea behind internal metrics stems from the definition of clusters. A meaningful clustering solution should group objects into various clusters, so that objects within each cluster are more similar to each other than objects from different clusters. In particular, **intra-cluster similarity**, *ISim*, is defined as the average similarity between objects within each cluster, and **inter-cluster similarity**, *ESim*, is defined as the average similarity between objects within each cluster *and* the remainder of the objects in the data set. After obtaining *ISim* and *ESim* values for all clusters in a clustering solution, we calculate the average *ISim* ($ISim_{avg}$) and *ESim* ($ESim_{avg}$) to evaluate the quality of the entire clustering solution. We also report the ratio between the $ISim_{avg}$ and $ESim_{avg}$ measures. The higher the ratio, the better the clustering solution is.

5.2.2 External Metrics for Hard Clustering

External metrics rely on the true class memberships in a document set. We use two such metrics, **entropy**, which is a function of the distribution of classes within the resulting clusters, and **purity**, which is a function of the relative size of the largest class in the resulting clusters.

Given a particular cluster, S_r, of size n_r, the entropy of this cluster is defined as

$$E(S_r) = -\frac{1}{\log q} \sum_{i=1}^{q} \frac{n_r^i}{n_r} \log \frac{n_r^i}{n_r}, \tag{5}$$

where q is the number of classes in the data set, and n_r^i is the number of documents of the ith class that were assigned to the rth cluster. The entropy of the entire clustering solution

is then defined as the sum of the individual cluster entropies weighted according to the cluster size. That is,

$$\textbf{Entropy} = \sum_{r=1}^{k} \frac{n_r}{n} E(S_r). \quad (6)$$

A perfect clustering solution will result in clusters that contain documents from only a single class, in which case the entropy will be zero. In general, the smaller the entropy values, the better the clustering solution is. In a similar fashion, the purity of a cluster is defined as

$$P(S_r) = \frac{1}{n_r} \max_i(n_r^i), \quad (7)$$

which is the number of documents of the largest class in a cluster divided by the cluster size. The overall purity of the clustering solution is obtained by taking a weighted sum of the individual cluster purities and is given by

$$\textbf{Purity} = \sum_{r=1}^{k} \frac{n_r}{n} P(S_r). \quad (8)$$

In general, the larger the values of purity, the better the clustering solution is.

5.2.3 F-score Metric for Soft Clustering

Given the true class label of each document, we use average F-score to evaluate various soft clustering solutions [36].

Given cluster i and class j, the F-score is calculated as follows:

$$\textbf{F-score} = \frac{(\beta^2 + 1.0)(P * R)}{(\beta^2 * P) + R} \quad (9)$$

where P is the *precision* of cluster i (defined by the number of cases that are both in cluster i and class j, divided by the number of cases in cluster i), and R is the *recall* of cluster i (defined by the number of cases that are both in cluster i and class j, divided by the number of the cases in class j).

We measure the quality of our resulting soft clustering by first associating clusters with classes and then measuring the similarity between the resulting cluster-class pairs using equation (9). We define **average F-score** as the average of these individual F-score values.

Varying the β coefficient provides a means of biasing F-score towards precision or recall (e.g., $\beta = 0.5$ biases it towards precision; $\beta = 1.0$ weights precision and recall equally; $\beta = 2.0$ biases it towards recall). In our study, we test all three cases to observe the relative performance of the various soft clustering algorithms, while emphasizing the quality of the resulting clusters (P), complete coverage (R) , and both (P+R).

5.2.4 Human Assessment

We also ask paralegal researchers to assess the quality of the resulting clusters in two respects: coherence and usefulness. For both assessments, they use a five point Likert scale ranging from 1 (low coherence with the current cluster's central topic or low usefulness to a legal researcher) to 5 (high coherence with the current cluster's central topic or high usefulness to a legal researcher).[3] A cluster is useful insofar as it facilitates knowledge sharing, which means that it groups documents together that are about the same topic.[4]

[3]The second metric was included since it is possible to have a cluster of documents that all share certain characteristics (e.g., a set of digitally-scanned fax messages), but which as a group are not particularly useful to a lawyer.

6. EXPERIMENTAL RESULTS AND DISCUSSION

All of our data sets are indexed by the `doc2mat` utility provided in the CLUTO package. We supply our own stopword list containing roughly 300 words; we also stem terms using the Porter stemmer [27]. During preliminary experiments on our initial legal collections, we examined various combinations of optimization methods and criterion functions provided by CLUTO.

6.1 Hard Clustering Experiments with Traditional Legal Data and the LRC Collection

6.1.1 Results from Traditional Legal Data

In our series of experiments with the traditional legal data collection (described in Section 4.1.1), we determined that, on average, the top-performing hard clustering algorithm for legal text is repeated-bisection with a global-optimization step (`rbr`). We also determined empirically that the two top-performing hard clustering criterion functions are $\mathcal{I}_1$ and $\mathcal{I}_2$ [equations (3) and (4)]. With few exceptions, the results reported in the remainder of this paper are produced by the `rbr` partitional algorithm and the $\mathcal{I}_2$ criterion function.

We should also note that the number of underlying topical classes associated with this collection is 11 (from Section 4.1.1). When specifying the input number of clusters for the toolkit, we would thus typically select the number of clusters to be significantly larger than n=11, in order to determine whether `CLUTO` could not only partition the top-level documents into their respective classes, but also at least partially partition Bankruptcy into reasonable sub-classes, which indeed it did for n=20.

In one representative and close to optimal trial with the data set, the ratio of $ISim_{avg}$ to $ESim_{avg}$ is 7.95 while the Purity value is 0.74 and the Entropy value is 0.27. In general, when $ISim_{avg} \approx 10 \times ESim_{avg}$, then the resulting clusters are of noteworthy quality. In the case of the traditional legal data collection, the assessor rated our best clustering results both topically coherent and useful to legal practitioners. The majority of the clusters examined received scores of 4 and 5 (out of 5) from our human assessor.

6.1.2 Results from the LRC Collection

Results from hard clustering trials performed on the LRC collection are presented in Table 4, and are discussed below.

6.1.2.1 Clustering at the Top-level.

Hard clustering results from trials performed at the root-level of the LRC collection (Figure 4) are presented in the top row of Table 4. Although we did not achieve an optimal $ISim_{avg}$ to $ESim_{avg}$ ratio, our mean human assessment scores nonetheless suggest that the resultant clusters are of high quality. These averages are based on a random sampling of at least 25% of the clusters, with standard deviation shown in parentheses.

It is also worth observing that in cases where the number of clusters created exceeds the number of topical classes

[4]The instructions to an assessor on what is "useful" to a lawyer thus tended to emphasize practical topical categories (e.g., Criminal Justice, Personal Injury, etc.) over, for instance, more rudimentary means of grouping such as by document type (e.g., briefs, administrative letters, etc.).

Data Set	Internal Measures			External Measures		Human Assessment	
	Intra-cluster Similarity, $ISim_{avg}$	Inter-cluster Similarity, $ESim_{avg}$	$\frac{ISim_{avg}}{ESim_{avg}}$ Ratio	Purity	Entropy	Coherence [1=Low ... 5=High]	Usefulness [1=Low ... 5=High]
LRC (complete)	0.128	0.025	5.22	0.441	0.530	4.34 (0.95)	4.17 (1.06)
Civil Procedure	0.166	0.033	5.09	0.490	0.476	4.56 (0.74)	4.51 (0.84)
Torts / Injuries	0.203	0.034	6.02	0.507	0.373	3.30 (1.83)	2.69 (1.89)
C.P.–Evidence	0.346	0.032	*10.81*	0.612	0.369	4.07 (1.46)	3.93 (1.58)
C.P.–Privileges	0.193	0.037	5.17	*0.859*	*0.177*	—	—
T.-I.–Negligence	0.211	0.034	6.30	0.545	0.440	—	—
T.-I.–Damages	0.200	0.034	5.95	0.667	0.387	4.39 (1.04)	4.06 (1.47)

Table 4: Performance Metrics for Hard/Hierarchical Clustering of Law Firm Corpus (cf: Fig. 4)

Data Set	β	Random Assignment	Case Clustering			Segment Clustering
			Init. Hard Asmt.	Overall	% Improvement	
50Topics-2HNs	0.5	0.129	0.482 (3.7)	0.488 (3.8)	+1.2%	0.472 (3.7)
	1.0	0.136	0.416 (3.1)	0.460 (3.4)	+10.6%	0.437 (3.2)
	2.0	0.161	0.384 (2.4)	0.460 (2.9)	+19.8%	0.424 (2.6)
50Topics-3HNs	0.5	0.114	0.432 (3.8)	0.433 (3.8)	+0.2%	0.415 (3.7)
	1.0	0.118	0.342 (2.9)	0.386 (3.3)	+13.5%	0.384 (3.3)
	2.0	0.142	0.299 (2.1)	0.391 (2.8)	+30.8%	0.382 (2.7)
LRC	0.5	0.057	0.200 (3.5)	0.276 (4.8)	+38.0%	-
	1.0	0.060	0.209 (3.5)	0.251 (4.2)	+20.1%	-
	2.0	0.070	0.246 (3.5)	0.264 (3.8)	+7.3%	-

Table 5: Average F-score from Two Soft Clustering Algorithms

associated with the data set, the final one or two clusters can and often do serve as lower-scoring "garbage-collector" bins which tend to consist of short or low content-bearing documents.

6.1.2.2 *Hierarchical Clustering.*

Results from hard clustering trials performed iteratively on the second and third levels of the LRC collection (Figure 4) are presented in the lower rows of Table 4. Worth mentioning is that in a number of instances such as those at level 3 under Civil Procedure—Evidence and Privileges—certain values remain affirmative, in particular, the ratio of $ISim_{avg}/ESim_{avg}$ for Evidence and the *Purity* and *Entropy* values for Privileges (in italics). Perhaps more important, however, is that the human assessments of each of the resultant clusters have generally remained in the high-coherence and high-usefulness end of the Likert scale.

6.2 Soft Clustering Experiments with Topic-based and LRC Collections

In Table 5, we present the average F-scores achieved by the segment clustering approach and the case clustering approach for the LRC and two topic-based collections. Since segment information is not available for the LRC collection, we only report the results of case clustering for this data set. For the case clustering approach, we also report the average F-scores obtained by performing the initial hard clustering only, in order to show how effective the multiple assignment stage is (quantified in the "% Improvement" column). We also create random assignment baselines by randomly assigning two class labels to each case for the 50Topics-2HNs and LRC collections, and three class labels for 50Topics-3HNs. The various F-scores achieved by random assignment are shown in the column labeled "Random Assignment." The values in parentheses represent the improvement ratio over the corresponding random baseline. For each data set, we calculate F-scores with $\beta = 0.5$, 1.0, and 2.0, and report them in three separate rows.

A number of observations can be made from Table 5. First, as shown, the two soft clustering algorithms work effectively in all instances with relative improvements over random assignment ranging from 2.6 to 4.8. The case clustering approach performs comparatively or better than the segment clustering approach in all instances. Second, the multiple assignment stage clearly improves the clustering solution produced by initial hard clustering for all three data sets. For the 50Topics-2HNs and 50Topics-3HNs data sets, the greatest relative improvements of the multiple assignment stage are achieved with $\beta = 2.0$, and the least with $\beta = 0.5$. Note that $\beta = 2.0$ biases F-score towards recall and $\beta = 0.5$ biases F-score towards precision. Hence, the results suggest that the multiple assignment stage "boosts" the clustering solution especially from a recall point of view. On the other hand, in terms of precision, the solid performance of the initial hard clustering on these two data sets makes it difficult for the multiple assignment stage to achieve significant improvement. Third, for LRC collection, the relative improvements of the multiple assignment stage are significant with all three β values. Recall that the LRC collection differs from the two 50Topics data sets in two ways: the average number of classes per document is significantly lower (Table 3); and it contains a more diverse document set (e.g., briefs, memos, letters, electronic transmittals, and others). The former directly explains why the relative improvement of the multiple assignment stage for LRC is not as significant as those for the other two data sets when evaluated with $beta = 2.0$ (biased toward recall). The latter makes it difficult for hard clustering algorithms

to generate high-quality solutions for LRC and leaves more room for soft clustering algorithms to improve upon them.

Table 6 shows that the quality of the clusters for the LRC collection can be traced to the type of document being clustered. Although these LRC clusters may consist of any combination of document-types, we present human assessments of them by document-type because this attribute tends to correspond to general length or content-bearing qualities of the relevant text, with briefs possessing the most substantial textual discourse, memoranda nearly as much, and letters, in general, the least.[5] Even though the resulting clusters formed primarily by letters tend to be of low coherence and usefulness, our method is able to construct high-quality clusters for greater content-bearing documents such as briefs and memoranda. The significance of this finding for systems designed to yield high-precision performance is that KM managers may wish to rely upon briefs and memoranda over letters in order to establish highly coherent initial clusters.

LRC Document-type	Human Assessment	
	Coherence [1=Low ... 5=High]	Usefulness [1=Low ... 5=High]
Briefs	3.76 (0.99)	3.48 (1.14)
Memoranda	3.63 (1.01)	3.16 (1.57)
Letters	2.09 (1.36)	1.64 (1.17)

Table 6: Human Assessment of Resultant LRC Clusters

6.3 Parameter Study

As described above, one of the essential parameters of the soft ("case") clustering algorithm is the internal Z-score threshold, which controls the number of candidate clusters that are eligible for a document's subsequent assignment. By evaluating our soft clustering solutions with the F-score, we permit the internal Z-score threshold to improve precision at the expense of lowering recall when the threshold is raised, and the converse, when it is lowered.

We plot the average F-score of the soft clustering solutions obtained with various internal Z-score threshold values for 50Topics-2HNs and 50Topics-3HNs in Figures 6 and 7, respectively.[6] As shown in both figures, the best performance was achieved with the threshold value between 3 and 4. By contrast, when this same study was performed on the LRC collection (Figure 8), we see the relative F-score performance is reduced, likely attributable to the underlying disparate document types. We also notice that the performance curves are flatter, possibly due in part to the generally lower F-score behavior of these documents as a whole, yet when considering the three β curves simultaneously the top performance remains close to a Z-score threshold of 3.5.

7. CONCLUSIONS

In law firm environments where representative taxonomies or labeled training documents are not available, technologies complementary to search and classification are required to foster the organization, delivery and reusability of internal legal work products. In the research that we have conducted, we have seen that document clustering, especially clustering that performs well in hierarchical and multiple assignment contexts, holds promise to answer such requirements.

[5]Standard deviation for these assessments is presented in parentheses. The internal Z-score threshold used in this experiment is $z = 3.5$.

[6]The "1Asmt" plots represent hard clustering baselines, while the "2-3Asmt" plots represent the contributions of soft clustering.

Moreover, essential technologies such as clustering can serve to establish a foundation that supports the longer term knowledge mangement goals of leveraging analytical resources across a firm. If the hard and soft clustering approaches that we have examined continue to deliver effective solutions to rapidly expanding information environments, such techniques may be suitable for a host of related large-scale applications in the legal domain, including cross-firm or enterprise KM and Electronic Data Discovery (EDD).

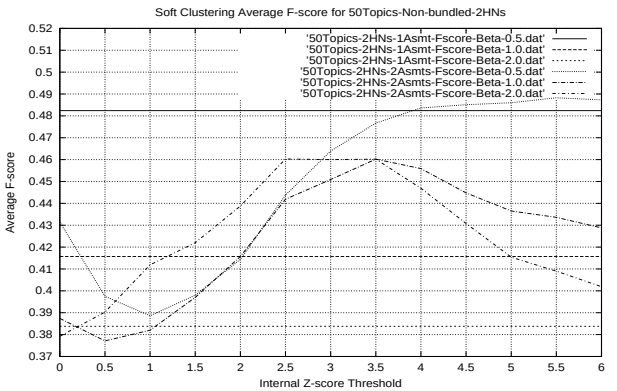

Figure 6. Average F-score for soft clustering solutions obtained for 50Topics-2HNs

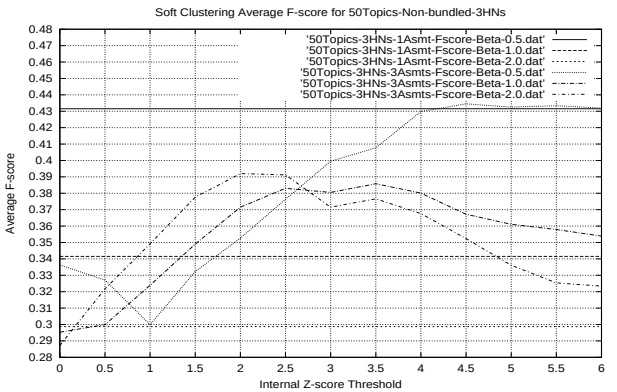

Figure 7. Average F-score for soft clustering solutions obtained for 50Topics-3HNs

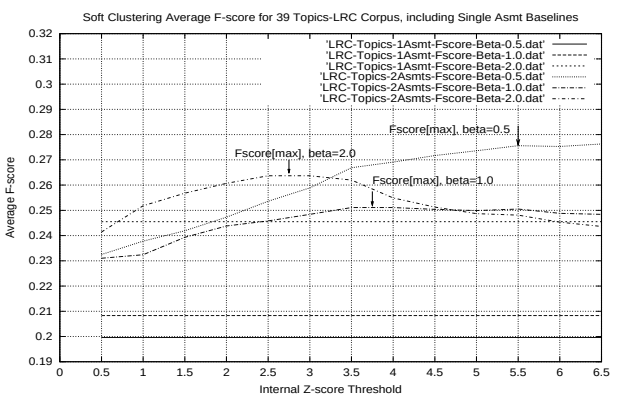

Figure 8. Average F-score for soft clustering solutions obtained for 39Topics-LRC Corpus

8. FUTURE WORK

One of our goals is to further enhance the soft ("case") clustering algorithm presented in this work. By expanding

the study of the number of clusters a document may be assigned to, we may be able to develop additional heuristics to determine a document's "cluster quotient." In parallel with this effort, we have begun to conduct experiments that rely on hybrid feature sets, such as document profiles constructed from high content-bearing word-*pairs*. We are also exploring the application of hierarchical clustering to taxomony development in order to enable the subsequent deployment of automatic categorization tools.

9. ACKNOWLEDGMENTS

We thank Po Yang-Stephens and Arun Vaccher for their help in creating our test collections derived from West data. We are also grateful for the assistance of Dan Dyke for his detailed assessments of the quality and utility of our resultant cluster sets.

10. REFERENCES

[1] C. C. Aggarwal, S. C. Gates, and P. S. Yu. On the merits of building categorization systems by supervised clustering. In *Proceedings of the Fifth Int'l Conference on Knowledge Discovery and Data Mining (KDD'99)(San Diego, CA)*, pages 352–356. ACM Press, Aug. 1999.

[2] K. Al-Kofahi, A. Tyrrell, A. Vachher, T. Travers, and P. Jackson. Combining multiple classifiers for text categorization. In *Proceedings of the 10th Int'l Conference on Information and Knowledge Management (CIKM'01) (New Orleans, LA)*, pages 97–104. ACM Press, Nov. 2001.

[3] T. J. Bench-Capon and P. R. Visser. Ontologies in legal information systems. In *Proceedings of the Sixth Int'l Conference of Artificial Intelligence and Law (ICAIL'97) (Melbourne, Australia)*, pages 132–141. ACM Press, June 1997.

[4] J. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum Press, New York, 1981.

[5] P. S. Bradley, C. Reina, and U. M. Fayyad. Clustering very large databases using EM mixture models. In *Proceedings of the Int'l Conference on Pattern Recognition (ICPR '00)*, volume 2, pages 2076–2080, 2000.

[6] P. Cheeseman and J. Stutz. Baysian classification (AutoClass): Theory and results. In U. Fayyad, G. P.-Shapiro, P. Smith, and R. Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, pages 153–180. AAAI/MIT Press, 1996.

[7] C. Curling. KeySearch, West's Key Number System, & Lexis' Search Advisor. *Law Library Resource Exchange*, May 2001. http://www.llrx.com/features/keysearch.htm.

[8] D. Cutting, J. Pedersen, D. Karger, and J. Tukey. Scatter/Gather: A cluster-based approach to browsing large document collections. In *Proceedings of the 15th Int'l Conference on Research and Development in Information Retrieval (SIGIR'93) (Copenhagen, Denmark)*, pages 318–329, Copenhagen, June 1992.

[9] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*, chapter 10: Unsupervised Learning and Clustering, pages 3–87. Wiley-Interscience, 2nd edition, 2000.

[10] D. L. Edwards and D. E. Mahling. Toward knowledge management systems in the legal domain. In *Proceedings of the Int'l ACM SIGGROUP Conference on Supporting Group Work: The Integration Challenge (Phoenix, AZ)*, pages 158–166. ACM Press, Nov. 1997.

[11] P. Gottschalk. Use of IT for Knowledge Management in Law Firms. *The Journal of Law and Information Technology (JLIT)*, 3, 1999.

[12] S. Guha, R. Rastogi, and K. Shim. CURE: An efficient clustering algorithm for large databases. In *Proceedings of the Int'l Conference on Management of Data (SIGMOD'98) (Seattle, WA)*. ACM Press, June 1998.

[13] S. Guha, R. Rastogi, and K. Shim. ROCK: a robust clustering algorithm for categorical attributes. In *Proceedings of the 15th Int'l Conference on Data Engineering*, pages 512–521, March 1999.

[14] A. Jain and R. C. Dubes. *Algorithms for Clustering Data*. Prentice Hall, 1988.

[15] G. Karypis. *CLUTO: A Software Package for Clustering High-Dimensional Data Sets*. University of Minnesota, Dept. of Computer Science, Minneapolis, MN, Nov. 2003. Release 2.1.1 (www-users.cs.umn.edu/ karypis/cluto).

[16] M. E. Katsh. *Law in a Digital World*, page 172. Oxford University Press, Oxford, 1995.

[17] B. King. Step-wise clustering procedures. *Journal of the American Statistical Association*, 69:86–101, 1967.

[18] D. H. Kraft, J. Chen, and A. Mikulcic. Combining fuzzy clustering and fuzzy inference in information retrieval. In *Proceedings of the IEEE Int'l Conference on Fuzzy Systems (FUZZ-IEEE'00)*, pages 375–380, May 2000.

[19] J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the 5th Berkeley Symposium of Mathematical Statistical Probability*, pages 281–297, 1967.

[20] K. Martin. 'Show me the money' – measuring the return on knowledge management. *Law Library Resource Exchange*, Oct. 2002. http://www.llrx.com/features/kmroi.htm.

[21] G. J. McLachlan and T. Krishnan. *The EM Algorithm and Extensions*. Wiley, New York, 1997.

[22] C. Meltzer. *Personal Communication*. Chief Information Officer, Dorsey & Whitney, LLP, Minneapolis, MN, Feb. 2004.

[23] M. E. S. Mendes and L. Sacks. Evaluating fuzzy clustering for relevance-based information access. In *Proceedings of the IEEE Int'l Conference on Fuzzy Systems (FUZZ-IEEE'03)*, pages 648–653, May 2003.

[24] I. Nonaka and H. Takeuchi. *The Knowledge-Creating Company*. Oxford University Press, 1995.

[25] C. Ordonez and E. Omiecinski. FREM: fast and robust EM clustering for large data sets. In *Proceedings of the 11th Int'l. Conference on Information and Knowledge Management (CIKM'02) (McLean, VA)*, pages 590–599. ACM Press, Nov. 2002.

[26] A. Oskamp, M. W. Tragter, and A. R. Lodder. Mutual benefits for AI & Law and knowledge management. In *Proc. of the Seventh Int'l Conf. of Artificial Intelligence and Law (ICAIL '99) (Oslo, Norway)*, pages 126–127. ACM Press, June 1999.

[27] M. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.

[28] G. Salton. *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley Longman Publishing Co., Boston, MA, 1989.

[29] M. Sato and S. Ishii. On-line EM algorithm for the normalized gaussian network. *Neural Computation*, 12:407–432, 2000.

[30] M. Schireson. Does technology matter for knowledge management? In *KMWorld: Content, Document, and Knowledge Management*, page S12. Information Today, Nov/Dec 2004. Special Supplement on Best Practicies on Enterprise Knowledge Management.

[31] P. H. Sneath and R. R. Sokal. *Numerical Taxonomy*. Freeman, London, UK, 1973.

[32] M. Steinbach, G. Karypis, and V. Kumar. A comparison of document clustering techniques. In *Notes from KDD Workshop on Text Mining, held at the Sixth Int'l Conference on Knowledge Discovery and Data Mining (KDD'00)(Boston, MA)*, Aug. 2000.

[33] R. E. Susskind. The Spirit of AI and Law: Reflections on emerging technology in legal practice. In *The 9th Int'l Conference of Artificial Intelligence and Law (ICAIL'03) (Edinburgh, Scotland)*, June 2003. Keynote Address.

[34] A. Terrett. *Knowledge Management and the Law Firm*, pages 67–76. Emerald Group, Bradford, England, Sept. 1998.

[35] G. T. Tziahanas. *White Paper: Legal Knowledge Management: A Holistic Model*. Legal Research Center, Minneapolis, MN, April 2003.

[36] C. J. van Rijsbergen. *Automatic Information Structuring and Retrieval*. Ph.D. Diss., University of Cambridge, July 1972.

[37] P. R. Visser, R. W. van Krulingen, and T. J. Bench-Capon. A method for the development of legal knowledge systems. In *Proceedings of the Sixth Int'l Conference of Artificial Intelligence and Law (ICAIL'97) (Melbourne, Australia)*, pages 151–160. ACM Press, June 1997.

[38] B. Zhang, M. Hsu, and U. Dayal. K-harmonic means a data clustering algorithm. Technical Report HPL-1999-124, 1999.

[39] Y. Zhao and G. Karypis. Evaluation of hierarchical clustering algorithms for document datasets. In *Proceedings of the 11th Int'l. Conference on Information and Knowledge Management (CIKM'02) (McLean, VA)*, pages 515–524. ACM Press, Nov. 2002.

[40] Y. Zhao and G. Karypis. Empirical and theoretical comparisons of selected criterion functions for document clustering. *Machine Learning*, 55(3):311–331, 2004.