

# Legal Document Clustering with Built-in Topic Segmentation

Qiang Lu, William Keenan  
Thomson Reuters  
Research & Development  
50 Broad Street  
Rochester, New York 14694 USA  
{qiang.lu, william.keenan}@thomsonreuters.com

Jack G. Conrad, Khalid Al-Kofahi  
Thomson Reuters  
Research & Development  
610 Opperman Drive  
St. Paul, Minnesota 55123 USA  
{jack.g.conrad, khalid.al-kofahi}@thomsonreuters.com

## ABSTRACT

Clustering is a useful tool for helping users navigate, summarize, and organize large quantities of textual documents available on the Internet, in news sources, and in digital libraries. A variety of clustering methods have also been applied to the legal domain, with various degrees of success. Some unique characteristics of legal content as well as the nature of the legal domain present a number of challenges. For example, legal documents are often multi-topical, contain carefully crafted, professional, domain-specific language, and possess a broad and unevenly distributed coverage of legal issues. Moreover, unlike widely accessible documents on the Internet, where search and categorization services are generally free, the legal profession is still largely a fee-for-service field that makes the quality (e.g., in terms of both recall and precision) a key differentiator of provided services.

This paper introduces a classification-based recursive soft clustering algorithm with built-in topic segmentation. The algorithm leverages existing legal document metadata such as topical classifications, document citations, and click stream data from user behavior databases, into a comprehensive clustering framework. Techniques associated with the algorithm have been applied successfully to very large databases of legal documents, which include judicial opinions, statutes, regulations, administrative materials and analytical documents. Extensive evaluations were conducted to determine the efficiency and effectiveness of the proposed algorithm. Subsequent evaluations conducted by legal domain experts have demonstrated that the quality of the resulting clusters based upon this algorithm is similar to those created by domain experts.

## Categories and Subject Descriptors

H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval—*Clustering*; H.3.5 [Information Storage and Retrieval]: Online Information Services—*Commercial Services*

## General Terms

Algorithms, Experimentation, Evaluation

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

CIKM'11, October 24–28, 2011, Glasgow, Scotland, UK  
Copyright 2011 ACM 978-1-4503-0717-8/11/10 ...\$10.00.

## Keywords

clustering, unsupervised learning, topic segmentation

## 1. INTRODUCTION

Search is but one tool of a comprehensive information retrieval system. Other tools include recommendation, navigation, clustering, query suggestion, and personalization, among others. Such tools are often featured more prominently in a vertical domain than they are in the open Web. This is primarily because the underlying use cases are more clearly defined; more information about the users is also available; and in the case of fee-based models, information providers often deploy editorial resources in addition to algorithmic means to organize and index content to further support specific information needs.

Topical taxonomies and encyclopedias are two examples of editorial tools that are designed to help legal researchers discover critical information via navigation. West's Key Number System is one noteworthy example of such a taxonomy. It segments the American system of law into common research and practice areas (e.g., Civil Rights, Negligence, and Pretrial Procedure). Other legal information providers, such as Lexis provide similar tools, but the Key Number System,<sup>1</sup> which contains about 100,000 nodes and existed since the 1870s, is considered by many to be the gold standard. The Key Number System still has its limitations as stated in [9]. We have found that document clustering can be very beneficial to supplement the traditional, often manual analysis and interpretation of the legal text corpora.

Most clustering algorithms view documents as single atomic units in that they do not segment them further into topics. This is also true of some soft clustering algorithms, where soft clustering refers to the fact that a document may be assigned to multiple clusters. While this may work well for short single-topic news stories, we claim it to be ineffective for complex legal documents. For example, a judicial opinion may deal with a driver's complaint about the liability of an automobile insurer, where the driver caused personal injury, was denied insurance coverage, and where compensation was subsequently awarded by summary judgment.

Given the complexity and multi-topical nature of the majority of legal documents, we developed a recursive soft clustering framework with a built-in topic segmentation algorithm. We have successfully applied it to millions of legal documents and generated high quality legal topical clusters.

<sup>1</sup><http://west.thomson.com/westlaw/advantage/keynumbers/>

Evaluations of these machine-generated clusters by trained legal professionals were overwhelmingly positive, as the quality was determined to be close to that of human-generated clusters.

Given a document, e.g., in a search result, a user is presented with more documents, in a manner similar to the “more-like-this” feature found on the open Web. However, we group the recommended documents by issues (or clusters), thus it can be viewed as issue-based more-like-this.

Logically, our work can be described as follows: (1) identify the universe of legal issues, (2) for every legal issue, identify the set of most important documents for that issue, and (3) associate every document in our collections with one or more of these issues. The distinction between documents that are most important to a legal issue (i.e., members of a cluster), and those that are merely related to one (i.e., are associated with a cluster) is important because it allows us to distinguish between documents that are an authority on a subject and those that are merely relevant to it. Note that under such an organization, a document can be a *member* of some clusters and *associated* with others at the same time.

Collectively, these two clustering relationships produce a powerful, coherent, utility-based approach to supplementing search results with additional relevant documents, ones that share the same common issue if not the same common query terms. Such a conceptual tool permits users to explore a topic at greater depth rather than simply surveying the topic through trial and error querying.

Figure 1 illustrates the overall workflow of the clustering processes and applications of the underlying algorithm. At the highest level, the primary tasks (described above) generate the two key types of document-to-cluster relationships, (1) membership, and (2) associations (shown in the two blue boxes delineated by dotted lines). As the sub-components indicate, both the clustering membership and associations processes rely on the topic segmentation results that precede them. Other significant components of the workflow include (a) merging, since the algorithm can be invoked recursively, and (b) labeling, which is an important piece of any outward facing rendering of the clusters. Each of these components will be discussed in the remainder of the paper to the extent necessary to explicate them.

The paper is organized as follows. Section 2 briefly discusses related work. Section 3 gives a short description of the metadata available in the legal domain. Section 4 introduces definitions and notation used throughout this paper. Section 5 presents the overall clustering framework, which includes the topic segmentation algorithm, and the recursive clustering algorithm. Some potential applications using the generated clusters are discussed in Section 6. We describe our evaluation of cluster quality and system performance in Section 7. Finally, Section 8 concludes the work while discussing future research directions.

## 2. RELATED WORK

The ability to identify and partition a document into segments is important for many Natural Language Processing (NLP) tasks, including information retrieval, summarization, and text understanding. One of the most important applications of topic segmentation is the Topic Detection and Tracking (TDT) task, as described in [2]. Much research has been done on topic segmentation. Many unsupervised, domain independent approaches [7, 13, 27] exploit

lexical cohesion information. The fact that related or similar words and phrases tend to be repeated in topically coherent segments and segment boundaries often correspond to a change in the vocabulary [24]. Other approaches rely on complementary semantic knowledge extracted from dictionaries and thesauruses, or from collocations collected in large corpora, which use additional domain knowledge such as the use of hyponyms or synonyms [3, 8, 17, 18].

Clustering is an active area of research and a variety of algorithms have been developed in recent years. Clustering algorithms can be categorized into agglomerative schemes or partitioning schemes (depending on how the final clusters are generated), or hard/soft clustering (depending on the nature of the membership function, i.e., whether single vs. multiple assignments are permitted). A comprehensive survey on clustering algorithms is presented in [4]. The legal community has also benefited from this technology [22, 28].

Soft clustering is often used when algorithm designers want to capture the multi-topical nature of documents. The fuzzy C-means algorithm is one of the most widely used of these algorithms; it allows one document to be assigned to more than one cluster by using a fuzzy membership function. It has been applied to text document collections with some success [15]. However, the well-known limitation of fuzzy C-means and algorithms of this type—its dependency on its initialization—limits their application to very large document collections. Other soft clustering solutions also exist, such as probabilistic clustering frameworks built upon Expectation-Maximization (EM) algorithms [6, 23], Fuzzy Adaptive Resonance Theory (Fuzzy-ART) neural network [16] and the Suffix Tree Clustering (STC) algorithm [29], to cite a few examples.

Tagarelli and Karypis [26] incorporate document topic segmentation into their clustering framework by first breaking documents into paragraph-based segments and then grouping these segments into clusters using the spherical K-means algorithm. These segment-clusters, based on all the documents in the collections, are then further grouped into high level clusters (segment-sets) using a “fuzzy” version of the spherical K-means algorithm. In their framework, an induction process is introduced to map the segment-sets clustering solution to document-level clusters in order to provide the user with a more useful organization of the input texts. In the case addressed by the authors, the text/topic segmentation algorithm assumes that documents are multi-topical. It further assumes that document paragraphs represent coherent topics and topics shift on or around paragraph boundaries.<sup>2</sup> Issues of scale are the prime differences between this work and our own. The segments with which we start are four magnitudes greater than those used in this study.

## 3. LEGAL DOMAIN CHARACTERISTICS

Documents in the legal domain have some unique characteristics. These characteristics include being intrinsically multi-topical, relying on well-crafted, domain-specific language, and possessing a broad and unevenly distributed coverage of legal issues.

<sup>2</sup>The extent to which a framework will actually lead to effective solutions depends in part on the validity of its assumptions. In the legal domain, there is no guarantee that paragraphs will possess a single topic only. This work is performed on document sets of 2.5K-6.5K using 15-25 underlying classes and where the number of clusters is tied to the number of known classes, a rather unrealistic scenario.

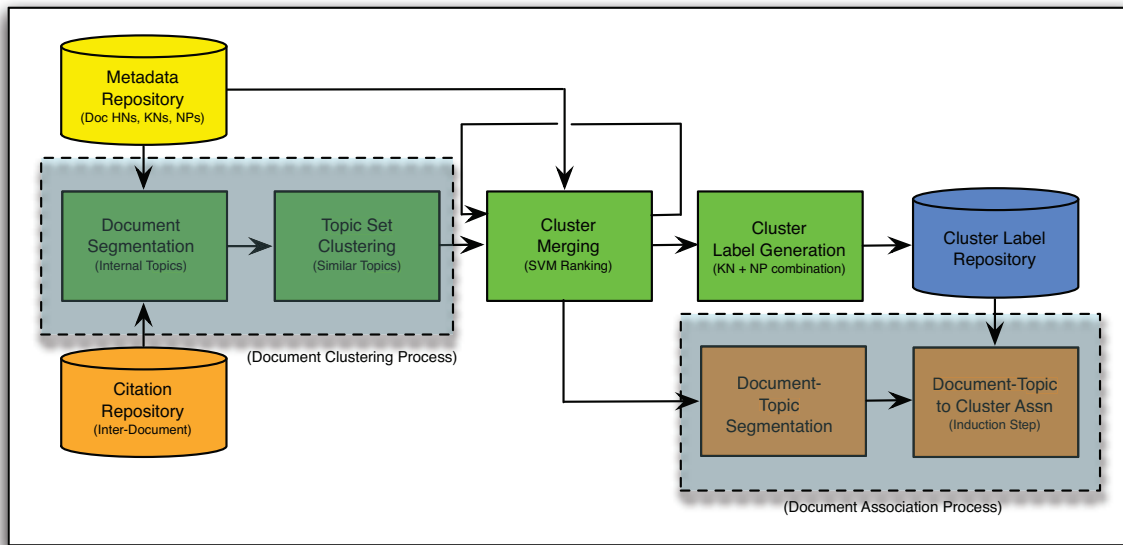


Figure 1: Workflow of topic clustering process and subsequent document-to-cluster associations application.

### 3.1 Data and Metadata Resources

Legal documents in the U.S. are complex in nature because they are the product of a highly analytical and adversarial process that involves determining relevant law, interpreting such law and applying it to the dispute to which it pertains. In doing so, courts must be careful not to diverge from established precedents or risk being overturned on appeal or criticized by future courts. This is because each case law document not only attempts to resolve a particular legal dispute, but also serves to help resolve similar disputes in the future. Legal publishers not only collect and publish the judicial opinions from the courts, but also summarize and classify them into topical taxonomies – such as the Key Number System (described in 3.1.3).

We mention some features of Thomson Reuters’ products below to illustrate a point about the kinds of resources legal publishers harness in order to offer researchers multiple entry points and search indexes into the content. Other legal publishers have their own analogous means of accessing their content.

#### 3.1.1 Judicial Opinions making Case Law Corpus

A judicial opinion (or a case law document) contains a court’s analysis of the issues relevant to a legal dispute, citations to relevant law and historical cases to support such analysis and the court’s decision. In other words, a judicial opinion expounds the law as applied to the case, and details the reasons upon which the judgment is based. By contrast, caselaw or legal cases refer to the *collection* of reported cases that forms a body of jurisprudence (i.e., the law on a particular subject formed by the decided cases), and is distinct from statutes and other sources of law. Generally considered among the most important legal documents, judicial opinions represent the bedrock of our clustering environment. A judicial opinion typically consists of several conventional components. These include the parties involved in the case, the jurisdiction, the court and judge(s) hearing the case,

the background and facts of the case, the case history (if this is an appellate court document), the holdings made by the court, and the mandate or binding legal decision. In addition, Thomson Reuters’ Westlaw System adds several annotations to these documents to summarize the points of law within and make them more accurate using a consistent language for the purposes of legal research. These include a synopsis of the case, a series of summaries of the points of law addressed in the case (3.1.2), classification of these points to a legal taxonomy (3.1.3), and an historical analysis of the case to determine whether its holdings and mandate remain intact or whether they have been overruled in part or in whole (3.1.4). Westlaw currently has just over 7 million annotated caselaw documents in the system and an additional 5 million unannotated caselaw documents available to legal researchers.<sup>3</sup>


#### 3.1.2 Caselaw Annotated Points of Law

Thomson Reuters’ Westlaw System creates “headnotes” for case law documents, which are short summaries of the points of law made in the cases. A typical case law document produces approximately 7 headnotes, but cases with over one hundred headnotes are not rare. On average, about 500,000 new headnotes are created each year, and the total repository now contains over 22 million headnotes corresponding to over 7 million annotated cases. West has been writing headnotes for over 120 years and the 7 headnotes per case average is more reflective of the last 10-15 years.

#### 3.1.3 Headnote Classification, Key Number System

Headnotes are further classified to a legal taxonomy known as the West Key Number System, an hierarchical classification of the headnotes across more than 100,000 distinct legal categories. Each category is given a unique alpha-numeric code, known as a Key Number, as its identifier along with

<sup>3</sup>Unannotated cases on Westlaw tend to be cases from lower level courts such as county courts or very short cases such as certain summary judgments or other cases with few if any citations.

- ↔ [313A Products Liability](#)
  - ↔ [313AII Elements and Concepts](#)
    - ↔ [313Ak132 Warnings or Instructions](#)
      - ↔ [313Ak134 k. Obvious Danger. \[Most Cited Cases\]\(#\)](#)  
(Formerly 313Ak54)
- ↔ [313A Products Liability](#)  [KeyCite Citing References for this Headnote](#)
  - ↔ [313AIII Particular Products](#)
    - ↔ [313Ak256 k. Ladders and Scaffolds. \[Most Cited Cases\]\(#\)](#)  
(Formerly 313Ak54)

Danger or potential danger of electrocution posed by use of aluminum ladder near high-voltage overhead power lines was danger which was generally known and recognized such that manufacturer of ladder had no duty to warn users to stay clear of power line, and thus manufacturer of aluminum ladder was not liable in products liability action based on theory of strict liability brought by personal representative of roofer who was killed by electrocution when conveyor ladder came into contact with power lines. (Per Goolsby, J., with one Judge concurring and Chief Judge dissenting.)

Figure 2: A example of a headnote with its assigned key number.

a descriptive name.<sup>4</sup> An example of a headnote with its assigned key number is shown in Figure 2.

### 3.1.4 Citation System

Equivalent to the links among Web pages, legal documents contain rich citation information just as documents from other domains do, such as scientific publications and patents. A case law document tends to cite previous related cases to argue for or against its legal claims; therefore, it is not unusual to have landmark cases decided by the U.S. Supreme Court with hundreds of thousands of cites to them.

KeyCite, a citation system created and maintained by West, is an example of such a citation network that keeps track of these rich relations between legal documents. Two types of citations are maintained in the system, *citing* (out-links to other legal documents) and *cited* (in-links to the instant document). KeyCite’s key functionality includes: indicating whether a document represents good (valid) law (i.e., has a decision been overruled or weakened in a subsequent opinion?), the depth of treatment a citation may receive in the citing document, and whether the citing document directly quotes the cited document.<sup>5</sup>

These rich metadata provide useful information not only to legal researchers but also to data mining algorithms for better understanding of complicated legal issues.

## 4. DEFINITIONS AND NOTATION

Let  $D = \{d_1, \dots, d_N\}$  denote the set of documents. Each document  $d_i \in D$  is seen as being comprised of a set of topics which contains text and other metadata information. A set of topics,  $T$ , is called a topic-set. We denote with  $T_j$  a topic-set from a document  $d_i$ , and further with  $T_{ij}$  being the  $j_{th}$  topic in the document  $d_i$ . Also we use  $TS = \cup_{i=1 \dots N} \cup_{d_i \in D, j=1 \dots k} T_{ij}$  for  $k$  topics in  $d_i$  to represent the set of topic-sets from all the documents in the collection.

For clustering purposes, we use  $C = \{c_1, \dots, c_M\}$  to denote the distinct cluster set that exists in the document collection  $D$ , of which  $c_k$  is one cluster consisting of similar topics from different documents.

In general, the vector-space model is used to represent the documents to be clustered. A vector not only contains

items from textual space, such as terms, but also contains items from other metadata, such as legal classification assignments, citator information based on inter-document citing and cited relationships, and click stream data from user behavior databases. For textual data, unless otherwise specified, text term relevance is weighted by using the standard *tf.idf*, which computes the weight of any term  $w$  as  $tf.idf(w) = tf(w) \times \log(N/N(w))$ , where  $tf(w)$  is the number of occurrences of  $w$  in a document (term frequency),  $N$  is the total number of documents in the collection  $D$ , and  $N(w)$  is the portion of text documents in  $N$  that contains term  $w$ . Weights for the metadata representations are defined in the next section. The length of each vector is normalized so that it is of a unit length.

The cosine similarity is applied to compute the similarity between two vectors  $x_1$  and  $x_2$  in the vector-space model, which is defined to be  $cos(x_1, x_2) = (x_1 \cdot x_2) / (||x_1|| \times ||x_2||)$ , and  $||x||$  to be the length of a vector.

## 5. LEGAL DOCUMENT CLUSTERING VIA TOPIC SEGMENTATION

At a high level, our clustering approach consists of two steps. First, each document in the document set  $D$  is processed to identify its topic-set and the collection of topic-sets from all documents in  $D$  is aggregated to generate the topic-sets,  $TS$ . In the second step, similar topics in the topic-sets  $TS$  are grouped together to form final clusters using a soft clustering algorithm. See Figure 3, where the topic sets,  $TS = \{ts_1, ts_2, ts_3, \dots, ts_N\}$ , populate clusters  $C = \{c_1, \dots, c_M\}$ . The rest of this section describes how these two steps are performed.

### 5.1 Topic Segmentation in Legal Documents (Document Segmentation by Topics)

The topic segmentation algorithm leverages available metadata such as headnotes, key numbers, and citations. We found this approach to provide better results (in terms of coverage and quality of topics) over traditional topic segmentation algorithms that rely upon lexical cohesion and utilize only document text. For the purposes of emphasizing the segmentation task, we will subsequently focus only on legal documents possessing headnotes, i.e., case law documents. As stated in 3.1.2, headnotes are short summaries of points of law in case law documents, there-

<sup>4</sup> <http://west.thomson.com/westlaw/advantage/keynumbers/>  
<http://west.thomson.com/documentation/westlaw/wlawdoc/wlres/keynmb06.pdf>

<sup>5</sup> <http://west.thomson.com/documentation/westlaw/wlawdoc/web/kwclqr4.pdf>

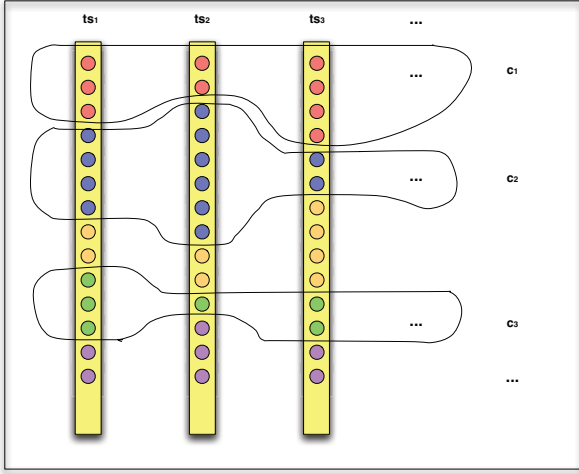


Figure 3: Topic sets populating final soft clusters.

fore have a near-complete coverage of the main legal issues in them. By grouping headnotes based on their “similarities” within a case law document, it is possible to identify the main legal topics within a document.

We use the vector-space model to represent headnotes in a case law document. A headnote is represented in terms of four types of features: text, key numbers, KeyCite citations, and noun phrases.<sup>6</sup> Thus, a headnote feature vector,  $h$ , is composed of four separate feature components, one per feature type. The similarity between a pair of headnotes,  $sim(h_i, h_j)$ , is defined as the weighted sum of the similarities between the corresponding component vectors. The weights are determined using heuristics.<sup>7</sup>

The similarity functions for the component vectors are defined as follows. For text-related features (i.e., headnote text and noun phrases) we use cosine similarity with a *tf.idf* weighting scheme. An analogous similarity function is also used for the key number features, where each key number is treated as if it was a word. For the KeyCite component vector we define similarity in terms of co-citations:

$$cite\_sim(h_i, h_j) = \frac{cite(h_i \cap h_j)}{cite(h_i \cup h_j)} \quad (5.1)$$

in which  $cite(h_i \cap h_j)$  represents the number of documents that cite both headnotes  $h_i$  and  $h_j$ , and  $cite(h_i \cup h_j)$  the number of documents that cite either headnote  $h_i$  or  $h_j$ .

An agglomerative clustering algorithm groups similar headnotes to generate the topic-set for a document. The algorithm merges two headnotes together while maximizing the following equations,

$$F = maximize \frac{\tau}{\epsilon} \quad (5.2)$$

where,

$$\tau = maximize \sum_{r=1}^k \sum_{h_i \in T_r} sim(h_i, \bar{T}_r) \quad (5.3)$$

<sup>6</sup> An in-house study found that noun phrases, e.g., from headnotes, closely approximated the associated key legal concepts, e.g., “product liability.”

<sup>7</sup> We determined empirically that the key number feature was the most discriminating. Given that it is assigned by humans, is independent of any particular terms or phrases, and is highly granular (having O(100K) leaf nodes), the finding is not surprising.

and

$$\epsilon = minimize \sum_{r=1}^k n_r sim(\bar{T}_r, \bar{T}) \quad (5.4)$$

$$\bar{T}_r = \frac{\sum h}{n_r} \quad (5.5)$$

$$\bar{T} = \frac{\sum \bar{T}_r}{k} \quad (5.6)$$

where  $\tau$  is the intra-cluster similarity and  $\epsilon$  is the inter-cluster similarity,  $k$  denotes the total number of topics in a document,  $T$  denotes the topic-set for a document,  $T_r$  denotes an individual topic, and  $n_r$  is the number of headnotes in the topic  $T_r$ . Also,  $\bar{T}_r$  and  $\bar{T}$  represents the center of a single topic and all topics, respectively.

Notice that the algorithm does not require the number of topics as an input parameter; rather, it depends on an intra-topic similarity threshold to control the granularity of the topics in a document. The threshold is determined empirically by analyzing the histogram of intra-cluster similarities. We use a set of documents with known topic-segmentations to guide our threshold selection process.

## 5.2 Topic Set Clustering

The above document segmentation process resulted in a large number of topics, over 10 million in total. Clearly, many of these topics are duplicative and thus called for further merging or clustering. This section describes the clustering process.

First, we needed to reduce computational complexity of the underlying problem. We do this in two ways: we reduced the dimensionality of the topics themselves, and we reduced the computational complexity of the clustering algorithm.

We performed aggressive feature selection to reduce the dimensionality of the topics from hundreds of words, noun phrases, and key numbers to a much smaller set. We use a ranker support vector machine (SVM) [25] to rank noun phrases in a cluster, and select the top  $n$  most descriptive and discriminative ones ( $n$  was set empirically to 5).

We then used a technique similar to that described in [21] to reduce the computational complexity of the algorithm. In [21], McCallum, et al. describe a “canopy” based clustering technique for large and high-dimensional data sets. The main idea is to use a cheap, approximate distance measure to efficiently divide the data into overlapping subsets, or “canopies” and then apply (traditional) clustering algorithms by measuring exact distances only between points that occur in a common canopy. This reduces the overall computational complexity of clustering dramatically.

Our methods are similar in that we first index the set of 10 million+ topics and for each topic we retrieve the top  $n$  most similar topics using several simple features. To merge topics, we use measurements in more extensive feature sets, but only against the set of most similar topics, thus producing dramatic improvements in processing speed without negatively affecting quality.

A topic can be a member of as many clusters as appropriate, a general property of soft clustering algorithms. A seed topic is a topic that can start its own cluster. However, a topic that is already a member of another cluster cannot be

used as a seed for a new cluster. This means that the resulting clusters are order dependent. To increase the likelihood that popular topics are represented in their own clusters we order topics according to their popularity (in terms of number of citations to the headnotes in a cluster), and start with most popular topics first.

We used a document classification engine, CaRE [1], to retrieve similar topics (clusters). CaRE allows the utilization of an ensemble of classifiers and comes equipped with a number of meta-classifiers to be used to combine the results of individual classifiers. For each topic (e.g., cluster), we have three classifiers – one per feature type – headnote texts, key numbers, and citation patterns. Other indexing engines (e.g., Lucene<sup>8</sup>) could have been used.

Given a seed topic, we use CaRE to retrieve a pool of candidate topics. We then use a ranker SVM to determine which of the topics in the candidate pool can be merged with the seed topic. To do this, we represent each seed-candidate pair in terms of a feature vector. The features include CaRE scores, as well as the four similarity functions described in section 5.1. In addition, the feature vector includes a *click* similarity feature. The basic idea is that documents that are frequently viewed (clicked on) in the same session, by different users, tend to share common topics. This feature is computed as follows:

$$click\_sim(d_i, d_j) = \frac{click(d_i \cap d_j)}{click(d_i \cup d_j)} \quad (5.7)$$

in which  $click(d_i \cap d_j)$  represents the number of times that both documents  $d_i$  and  $d_j$  have been clicked in the same session, and  $click(d_i \cup d_j)$  is the number of times that either document  $d_i$  or  $d_j$  has been clicked in all sessions.

The above similarity features are then used to train a ranker SVM to rank clusters in the retrieved pool and top ranked clusters are then merged into the seed cluster.

The algorithm is designed as a recursive process such that multiple rounds can be performed if needed. The initial input of the algorithm is the set of 10 million+ topics, and the output of each round is the input to subsequent rounds. The process stops when the inter-cluster similarities between any two clusters are lower than a predefined threshold if a subsequent merge took place. In all, we performed 3 rounds. After the first round, we ended up with 1.4 million clusters, and these were further reduced to approximately 360,000 clusters after the third and final round.

### 5.3 Additional Notes about the Clusters

We relied upon two Linux servers, each with 32 GB RAM, and two quad core 2.66 MHz CPUs. One of them was used for the central clustering and merging tasks and the other for CaRE similarity services. These servers reduced processing time from days to tens of hours. Table 1 contains a sample of clusters as well as their metadata. Notice that the majority of the 360,000 clusters are jurisdiction neutral. The clusters represent the universe of legal issues, some of which are state specific, others are federal and others are a mix of both. However, not all legal issues are represented in all jurisdictions. In fact, the resulting clusters only contain descriptions of legal issues. The process of “populating” these clusters with legal documents is evaluated in section 7.

No clustering algorithm is complete without a discussion about labeling. As one can see in Table 1, cluster labels are

<sup>8</sup><http://lucene.apache.org>

hierarchical (separated by ‘/’) and coherent. The algorithm for generating these labels is quite complex and will be the subject of another paper.

## 6. ASSOCIATING DOCUMENTS WITH CLUSTERS

By design, clusters are meant to contain the most important case law documents on a legal topic. Hence, not every case law document is a member of a cluster. Yet although clusters are case-centric, they are also populated with other types of legal documents such as statutes, regulations, administrative decisions. The utility of clusters as a means to organize legal content around issues or topics is as much a function of the quality of the clusters themselves as it is a function of their coverage. Without this universal coverage, one could not envision an issue-based “more like this,” where legal researchers can discover topics related to a document they are examining or dig deeper into a topic of interest. In other words, it is critical that most if not all legal documents (regardless of their type) be linked to these clusters. To address this challenge, we designed a process by which documents are associated with (or linked to) existing clusters. In other words, clustering defines the space of legal topics as well as the most important case law documents which are cluster members under that topic, while association indexes all types of legal content relative to the discovered topics.

As such, *association* is not part of the algorithm for cluster *membership*, but we mention it here because it significantly improves the utility of the resulting clusters. Very briefly, the document association algorithm consists of two steps: (1) segment documents into topics and (2) associate topics with clusters. The document segmentation step is tailored to the various content types but is analogous to the process described in Section 4. The association step is based on the similarity between topics and clusters. In summary, we have established a two-tier relationship between documents and clusters. The strongest relationship, for the most *authoritative* documents under a given topic, is one of *membership*. By contrast, the second relationship for documents that remain *relevant* to a given topic, is that of *association*.

## 7. EVALUATION AND PERFORMANCE

To assess the performance of our clustering approach, we clustered the set of headnoted case law documents on Westlaw. This consists of about 7 million U.S. case law documents, and collectively contains more than 22 million headnotes classified to about 100,000 key numbers.

### 7.1 Evaluation Design

We designed three different experiments as follows:

#### 7.1.1 Evaluation I: Cluster Quality – Coherence and Utility

Coherence was defined as the extent to which the documents in a given cluster address the same specific legal issue. Utility was defined as the usefulness of the documents in the given cluster to a legal researcher. The rationale behind why utility was assessed in addition to coherence was because it would be possible to have a cluster with a high coherence score, but not be very useful to a legal researcher, for example, if the documents contained within were clustered based upon a common dimension such as “all litigation involving a company.” So a cluster is considered to be useful to a legal

Cluster ID	Label	Noun Phrases	Key Numbers
12799254	Eminent Domain/ Compensation/ Nature of Property and Neighborhood and Estimated Replacement Value	“fair cash” “right way” “landowner right access” “diversion traffic” “access controlled highway”	148k221 148k141(1) 148k107
12436744	Opinion Evidence/ Challenged Portion of Expert Witness Opinion	“admissible expert testimony” “qualify expert witness” “product liable action” “direct examination manufacturer expert” “subject matter testimony”	157k536 157k546 157k539
12957372	Automobiles/ Arrest, Stop, or Inquiry; Bail or Deposit/ Stop and Arrest by Police Officers	“police officer” “valid traffic stop” “defendant drive license” “reason suspicion criminal activity” “law enforcement officer”	48Ak349(4) 35k63.5(6) 48Ak349(2.1)
13080456	Mines and Minerals/ Titles, Conveyances, and Contracts/ Prior Conveyance of Rights of the Oil Gas and Mineral Royalty Estate	“half mineral interest” “severance mineral estate” “gas mineral” “surface land” “possession surface”	260k55(4) 260k55(2) 260k55(5)
12879050	Child Custody/ Removal from Jurisdiction/ Best Interests and Welfare of Children and Rights of Other Parent	“admissible expert testimony” “modify child custody” “noncustodial parent child” “child father” “child quality life”	76Dk261 76Dk76 76Dk921(1)

Table 1: A sample set of clusters and their metadata

researcher not simply because it groups documents containing a similar “topic” together, but it also presents a useful legal concept with respect to the topic. In Evaluation I the top ranked 25 case law documents were evaluated for 128 clusters. These documents were ranked by relevance to the cluster definition. For both metrics, the reviewers used a five-point Likert scale ranging from 1 (low coherence with the current cluster’s central topic or low utility to a legal researcher) to 5 (high coherence with the current cluster’s central topic or high utility to a legal researcher).

### 7.1.2 Evaluation II: Cluster-to-Document Association Quality

In Evaluation II, the quality was evaluated indirectly through a document association process briefly described in Section 6. The quality of the associated clusters to documents of different types was graded by expert researchers using a five-point scale from A (high quality of associated clusters) to F (low quality of associated clusters).

### 7.1.3 Evaluation III: Legal Issue Detection Quality, Legal Issue Clustering Quality

In Evaluation III, a group of legal professionals were involved in creating 10 research reports from a cross-section of U.S. jurisdictions and covering different topical areas. Each of the reports included 7 or fewer of *the most authoritative documents*, including both primary sources, such as case law and statutes, and secondary sources such as analytical materials. Legal topics were identified manually for each of the documents by these experts. Further, they found that each of the 10 reports in this study has a common ‘thread’ (i.e., a common legal issue) running through it. However, the common thread did not always appear in each document in each of the reports.

Our algorithm was applied to the same set of reports to detect clusters in the documents. The objective of the assessment in this evaluation was two-fold: (1) is the clustering algorithm able to discover all the legal clusters in these documents, and (2) is the clustering algorithm able to find the common legal issue in each of the reports.

We used precision P and recall R to measure the performance for the first objective, in which:

- **P** is defined by the number of correctly identified clusters of a document (compare to the manually identified clusters) divided by the total number of clusters, and
- **R** is defined by the number of correctly identified clusters of a document divided by the total number of manually identified clusters of a document (as ground truth).

For the second objective, we used precision P for evaluation, which is defined by the number of common legal issues identified among documents in all reports divided by the number of common legal issues manually identified among documents in all reports by experts.

## 7.2 Performance

### 7.2.1 Evaluation I

Table 2 shows the quality assessment by legal experts of 128 randomly selected clusters. Each cluster was given three scores, for coherence, utility, and an overall score (an additional score to assess the overall cluster quality, using a simpler 3-point scale, A-C-F, representing high-medium-low).

The result clearly demonstrates the effectiveness of the clustering framework even when applied to such a large scale data set. The results are significantly higher than the ones reported in the authors’ earlier work in [9].

The slight discrepancy between the coherence score and utility score suggests a unique advantage of this human assessment. A cluster can contain highly coherent topics from different documents, but that in itself does not guarantee they will be equally useful, thus a slightly lower score for utility is possible. For example, a “summary judgment” cluster is a less interesting and less useful topic to a legal researcher due to its commonality in the legal litigation process; therefore one would expect it to receive a lower utility grade, even though it may contain highly related documents. This type of quality characteristic can only be revealed through human expert assessments. Traditional machine-generated cluster quality measurements, such as entropy and purity, as used in [9], are not suitable for this purpose.

Grade	Expert Assessment		
	Coherence	Utility	Overall Cluster Grade
5	41	37	65
4	71	54	0
3	15	29	60
2	1	7	0
1	0	1	3
<b>Average</b>	4.11	3.93	3.97

**Table 2: Quality of Randomly Selected Clusters**

### 7.2.2 Evaluation II

In Evaluation II, more than a thousand documents from different content types were run through our association algorithm to associate clusters to each of the documents. Besides cases, documents from nine other types were tested by the system. These include U.S. statutes, U.S. government regulations, U.S. secondary law materials (a.k.a. analytical materials), and others indicated in Figure 4. Figure 4 shows the quality of the associated clusters of tested documents from all ten types.

When legal experts were evaluating the quality of the associated clusters, they reached a consensus on the clusters with grade A being “excellent”, B being “good”, C being “acceptable”, D being “marginal”, and F being “poor”. In addition, the experts defined the “precision rate” as the ratio of A+B graded clusters, and the “success rate” as the ratio of A+B+C grade clusters, to the total associated clusters, respectively. The algorithm shows a consistently high success rate and good precision rate in documents across different content types.

One reason why two of the content types have poorer performance when compared to others, e.g. Analytical Materials and Expert Witness Reports, is that the language used in these types is quite different. For example, a typical expert witness report is recorded in a “Question and Answer” format with less specific legal terminologies such as those corresponding to common layman’s terms. By contrast, Analytical Materials may address more conventional legal topics. However, Analytical content is difficult as it does not follow standards in terms of style, breadth or depth of material; so without tuning the segmentation algorithm to such stylistic variations, performance is bound to suffer. As an illustration, *American Law Reports (ALR)* is very different from *American Jurisprudence (AmJur)* in terms of document and section lengths, while both are dramatically different from the format of law reviews.

### 7.2.3 Evaluation III

Regarding the first objective, the precision and recall of the system on 10 reports across different document types is shown in the Figure 5. Overall, we achieved reasonably high precision, but the recall was quite low especially for case law documents. The main reason for this is the aggressive filtering, by adopting much higher thresholds, in the post processing of the system to achieve high precision.

For the second objective, Table 3 shows the performance of the system in each individual report, as well as overall.

In this analysis, each of the 10 reports has a common legal issue running through it. However, the common ‘thread’ did not appear in each document in each of the reports.

In summary, the experts manually created clusters by identifying a common thread through all documents in 7 of the 10 reports; our system identified a common thread through all documents in 6 of the 10 reports. In one of the reports, our system missed a common thread in one of the documents in that report, and is thus considered as a failure. Across the entire set, experts manually created clusters and identified the common thread in 52 of the 58 documents (89.7%). Our system created clusters and identified the “common thread” in 51 of the 58 documents (87.9%).

As demonstrated in three different evaluations, the assessment of our proposed clustering system for legal documents consistently achieved significantly reliable results overall.

## 8. CONCLUSIONS AND FUTURE WORK

This paper describes a large scale soft clustering algorithm that relies on topic-segmentation. The performance of the algorithms is encouraging, especially given its validation by human legal experts through different test assessments.

Clustering large document collections remains a challenging problem, especially in the legal domain where documents with multiple topics are very common. Traditional clustering algorithms have shown limited success in this area. In the research we have conducted, we have shown that our topic segmentation-based soft clustering framework not only successfully incorporates metadata information into the topic segmentation process for a given document, but also develops a practical soft clustering system which is highly scalable.

In this paper, we have attempted to demonstrate the utility of highly refined issue-based clusters created through two important kinds of document-cluster relationships. The first, *membership*, identifies and populates these document clusters by defining a comprehensive set of legal issues. The second, *association* associates these clusters with the documents retrieved by users who seek “more-like-this” functionality, whether or not these documents have any terms in common with the original query. As such, this paper makes three contributions to the field. First, it implements a highly practical means for defining and populating clusters along issue-based dimensions. Second, it demonstrates how one can expand the set of original relevant legal documents using “more like this” functionality, one that does not require the intersection between original user query terms and those in the candidate documents. And third, by focusing on high precision clusters, we show how users can expand both the breadth and depth of their legal research, rather than conducting research that only surveys the documents returned by a given query. This approach has been



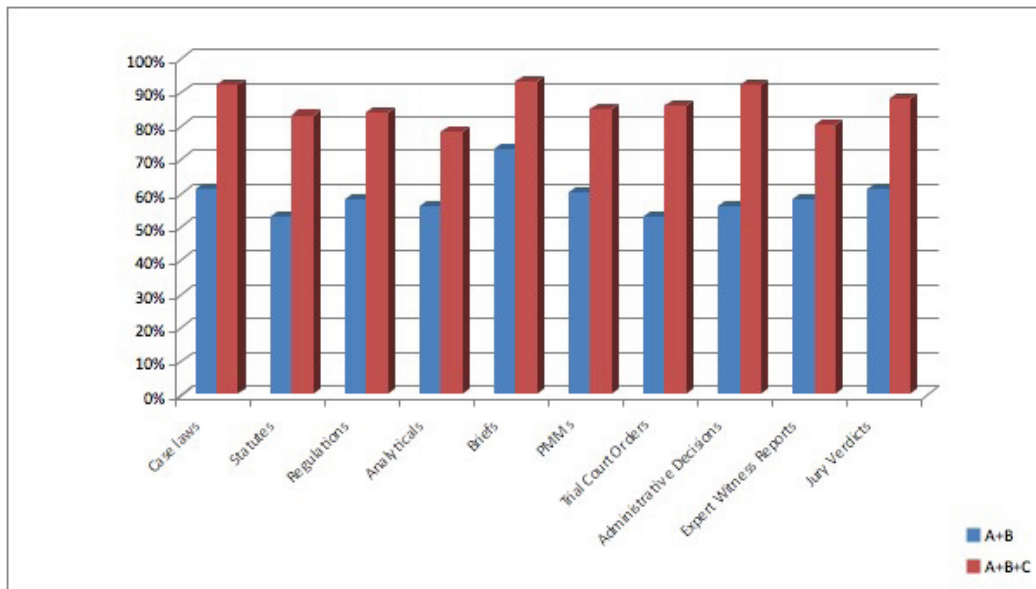


Figure 4: Quality Assessment of Associated Clusters for Different Content Types

Report IDs	Total No. of Documents	Documents with Common Theme Editorially Created Associations	Documents with Common Theme Algorithmically Created Associations
AE_2	6	6	6
AE_20	6	6	6
AE_22	4	4	4
AE_24	5	3	3
AE_31	6	6	6
AE_32	7	5	5
AE_35	6	6	5
AE_36	6	6	6
AE_37	5	3	3
AE_39	7	7	7
Total	58	52	51
Precision		89.7%	87.9%

Table 3: Precision of common legal issues of Evaluation III

scaled initially to address  $O(10M)$  clusters, which is another strength of the underlying techniques. Users, especially legal researchers, often prefer to have ability to drill down and focus on key concepts within a document set as opposed to getting a high-level overview of a document collection. Attention to fine-grained legal issues, robustness and resulting heterogeneous document clusters, and scalability are the characteristics that transform this cluster-based application into a highly resourceful research tool.

Topic segmentation for documents with little metadata information will be one focus of our future research work. Topic modeling algorithms, such as latent semantic indexing (LSI) [11] or probabilistic LSI [14], and Latent Dirichlet Allocation (LDA) [5], have been shown to be capable of modeling topics beyond lexical coherence from text documents into some conceptual aspects of basic linguistic notations, such as synonymy and polysemy. They have yet to be applied to complex documents—as far as we know, such as those in the legal domain—and produce promising outcomes.

The MapReduce framework introduced by Google [10] to

support distributed computing on very large data sets of clusters across commodity or dedicated computers has ignited much excitement given its ability to tackle very large scale document processing problems [19]. The development of the Apache Lucene Mahout [20] machine learning algorithm libraries implemented on top of the Apache Hadoop MapReduce paradigm [12] also holds promise to resolve critical problems, including those of very large scale document clustering and classification. We are looking into these subjects as another future research direction.

## 9. ACKNOWLEDGEMENTS

We thank John Duprey, Helen Hsu, Debanjan Ghosh and Dave Seaman for their help in developing software for this work, and we are also grateful for the assistance of Julie Gleason and her team of legal experts for their detailed quality assessments and invaluable feedback.

## 10. REFERENCES

- [1] K. Al-Kofahi, et al. Combining multiple classifiers for text categorization. In *Proceedings of the 10th International*

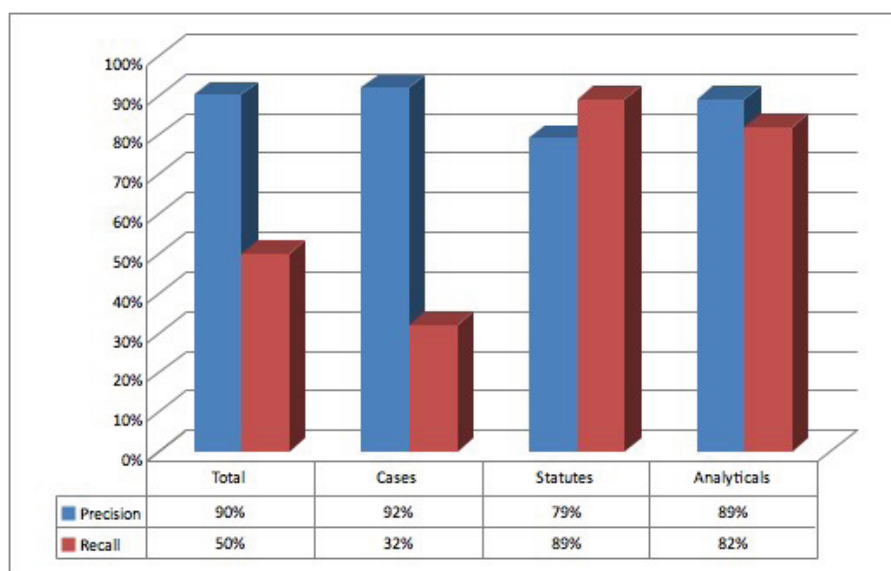


Figure 5: Precision and Recall of Evaluation III

- Conference on Information and Knowledge Management (CIKM01)*, pages 97–104, 2001.
- [2] J. Allen, et al. Topic detection and tracking pilot study – final report. In *Proceedings of the DARPA Broadcast News Transcription and understanding Workshop*, 1998.
- [3] D. Beeferman, A. Berger, and J. Lafferty. A model of lexical attraction and repulsion. In *Proceedings of the ACL*, pages 373–380, 1997.
- [4] P. Berkhin. A survey of clustering data mining techniques. *Grouping Multidimensional Data*, pages 25–71, 2006.
- [5] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2002.
- [6] P. Bradley, C. Reina, and U. Fayyad. Clustering very large databases using em mixture models. In *Proceedings of ICPR*, volume 2, pages 2076–2080, 2000.
- [7] F. Choi. Advances in domain independent linear text segmentation. In *Proceedings of the Association for Computational Linguistics*, pages 26–33, 2000.
- [8] F. Choi, P. Wiemer-Hastings, and J. Moore. Latent semantic analysis for text segmentation. In *Proceedings of EMNLP*, pages 109–117, 2001.
- [9] J. Conrad, K. Al-Kofahi, Y. Zhao, and G. Karypis. Effective document clustering for large heterogeneous law firm collections. In *Proceedings of the 10th International Conference on Artificial Intelligence and Law (ICAIL05)*, pages 177–187, 2005.
- [10] J. Dean and S. Ghemawat. Mapreduce: Simplified data processing on large clusters. In *Proceedings of the Sixth Symposium on Operating System Design and Implementation (OSDI04)*, 2004.
- [11] S. Deerwester, et al. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6):391–407, 1990.
- [12] Apache hadoop. <http://hadoop.apache.org/>, 2010.
- [13] Marti Hearst. Texttiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23:33–64, 1997.
- [14] T. Hofmann. Probabilistic latent semantic indexing. In *Proceedings of 22rd Annual International SIGIR Conference*, 1999.
- [15] M.C. Hung and D.L. Yang. An efficient fuzzy c-means clustering algorithm. In *Proceedings of the IEEE International Conference on Data Mining*, pages 225–232, 2001.
- [16] R. Kondadadi and R. Kozma. A modified fuzzy art for soft document clustering. In *Proc. of International Joint Conference on Neural Networks IJCNN*, pages 2545–2549, 2002.
- [17] H. Kozima. Text segmentation based on similarity between words full text. In *Proc. of the ACL*, pages 286–288, 1993.
- [18] H. Kozima and T. Furugori. Similarity between words computed by spreading activation on an english dictionary. In *Proceedings of the ACL*, pages 232–239, 1993.
- [19] J. Lin and C. Dyer. Data-intensive text processing with mapreduce. *Synthesis Lectures on Human Language Technologies*, 2010.
- [20] Apache mahout overview. <http://lucene.apache.org/mahout/>, 2010.
- [21] A. McCallum, K. Nigam, and L. Ungar. Efficient clustering of high-dimensional data sets with application to reference matching. In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD00)*, pages 169–178, 2000.
- [22] D. Merkl and E. Schweighofer. En route to data mining in legal text corpora: Clustering, neural computation, and international treaties. In *Proceedings of the 8th International Workshop on Database and Expert Systems Applications (DEXA '97)*, 1997.
- [23] C. Ordonez and E. Omiecinski. Frem: Fast and robust em clustering for large data sets. In *Proceedings of CIKM*, pages 590–599, 2002.
- [24] M. Shafiei and E. Milios. A statistical model for topic segmentation and clustering. *Lecture Notes in Computer Science*, 5032, 2008.
- [25] Svm light. <http://svmlight.joachims.org/>, 2010.
- [26] A. Tagarelli and G. Karypis. A segment-based approach to clustering multi-topic documents. In *Proceedings of the Text Mining Workshop, SIAM Data Mining Conference*, 2008.
- [27] M. Utiyama and H. Isahara. A statistical model for domain-independent text segmentation. In *Proceedings of the ACL*, pages 499–506, 2001.
- [28] N. Vaughn and D. Boley. Automated clustering and extraction of distinctive words in legal documents. Dept. of computer science and engineering report, University of Minnesota, 2001.
- [29] O. Zamir and O. Etzioni. Web document clustering: A feasibility demonstration. In *Proceedings of the 21st Int'l ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 46–54, 1998.