Bringing Order to Legal Documents
An Issue-based Recommendation System Via Cluster Association

Qiang Lu[1]* and Jack G. Conrad[2]*
[1]*Kore Federal, 7600A Leesburg Pike, Falls Church, VA 22043, U.S.A.*
[2]*Thomson Reuters Global Resources, Catalyst Lab, Neuhofstrasse 1, Baar, ZG 6340, Switzerland*

Keywords: Unsupervised Learning, Recommendation, Clustering, Labeling, Document-cluster Associations.

Abstract: The task of recommending content to professionals (such as attorneys or brokers) differs greatly from the task of recommending news to casual readers. A casual reader may be satisfied with a couple of good recommendations, whereas an attorney will demand precise and comprehensive recommendations from various content sources when conducting legal research. Legal documents are intrinsically complex and multi-topical, contain carefully crafted, professional, domain specific language, and possess a broad and unevenly distributed coverage of issues. Consequently, a high quality content recommendation system for legal documents requires the ability to detect significant topics from a document and recommend high quality content accordingly. Moreover, a litigation attorney preparing for a case needs to be thoroughly familiar the principal arguments associated with various supporting opinions, but also with the secondary and tertiary arguments as well. This paper introduces an issue-based content recommendation system with a built-in topic detection/segmentation algorithm for the legal domain. The system leverages existing legal document metadata such as topical classifications, document citations, and click stream data from user behavior databases, to produce an accurate topic detection algorithm. It then links each individual topic to a comprehensive pre-defined topic (cluster) repository via an association process. A cluster labeling algorithm is designed and applied to provide a precise, meaningful label for each of the clusters in the repository, where each cluster is also populated with member documents from across different content types. This system has been applied successfully to very large collections of legal documents, O(100M), which include judicial opinions, statutes, regulations, court briefs, and analytical documents. Extensive evaluations were conducted to determine the efficiency and effectiveness of the algorithms in topic detection, cluster association, and cluster labeling. Subsequent evaluations conducted by legal domain experts have demonstrated that the quality of the resulting recommendations across different content types is close to those created by human experts.

1 INTRODUCTION

The goal of a recommendation system is to suggest items of interest to a user based on the user's own historical behavior or behaviors of a community of other users. Practical applications have been widely adopted in different areas, such as recommending books, music, and other products from online shopping sites (Linden et al., 2003), movies at Netflix (Bennett and Lanning, 2007), and news from Google and Yahoo! (Liu et al., 2010; Li et al., 2010).

Depending on the targeted user, however, the recommendation task may differ greatly. A casual news reader or an online shopper may be satisfied with a few good recommendations, regardless of other poor suggestions. Professionals such as attorneys or brokers, by contrast, may demand more precise and comprehensive recommendations. For example, when an attorney is performing legal research, she has to comb through a huge amount of material to identify the most authoritative documents across different sources of the law, such as judicial opinions, statutes, and court briefs, to name just a few. In the field of law, it is well known how important high recall is for attorneys who may be preparing for a trial or related litigation proceeding. They cannot afford to miss a key case that may have significant bearing on their current legal and logical strategy. Furthermore, being presented relevant documents may not be enough; that is, they may not be sufficiently granular since it is the specific sub-document-level topics or arguments

*Both authors worked for Thomson Reuters Corporate Research & Development when this work was conducted.

Lu Q. and Conrad J..
Bringing Order to Legal Documents - An Issue-based Recommendation System Via Cluster Association.
DOI: 10.5220/0004136600760088
In *Proceedings of the International Conference on Knowledge Engineering and Ontology Development* (KEOD-2012), pages 76-88
ISBN: 978-989-8565-30-3

that are critical. Legal practitioners must familiarize themselves not only with the primary arguments that may be marshaled against them; they must also anticipate secondary or alternative arguments associated with the legal issue as well. While ineffective legal research wastes time and money, inaccurate and incomplete research can lead to claims of malpractice (Cohen and Olsen, 2007).

In this paper, we present a robust and comprehensive legal recommendation system; it delivers to its users the top-level legal issues underlying a case along with secondary and supplemental issues as well. So for a given a document, e.g., in a search result, a user is presented with additional documents that are closely related, and these recommended documents are grouped together based on issues that are discussed in the original document. This issue-based recommendation approach, relative to those arguments in the original document, is essential to the effectiveness of a legal research system due to the complete coverage provided. Legal documents are complex and multi-topical in nature. For example, a judicial opinion may deal with a customer suing a hotel for its negligence due to an injury-causing slip and fall in the shower, who later is awarded an amount of compensation by summary judgment in the court. At least three legal issues are presented in this case, namely negligence, summary judgment procedures and appropriateness of compensation, and each could lead to a collection of other important documents specifically addressing such a topic. By providing such an issue-oriented recommendation tool, users are able to explore legal topics within any document at much greater depth rather than simply surveying them.[1] Furthermore, this examination is enabled without explicitly summarizing the issues and constructing different search queries.

One additional means of ensuring an understanding of the scope of a particular issue in a cluster of documents is to provide a meaningful label. A simple phrase may be sufficient for many traditional documents, such as those in a news collection. By contrast, an hierarchically structured labeling approach may be more suitable for complex legal documents. Such an approach can provide a more precise and detailed description of many complicated legal issues at a fine-grained level yet still maintain a coherent and broad topical overview at the root level. Legal information providers often deploy editorial resources to organize and index content to support specific information needs, and topical taxonomies and encyclopedias are two examples of such editorial tools for content navigation purposes. Nodes in taxonomies can offer well-crafted descriptions to form a solid foundation for a labeling algorithm. It is up to the research scientist to devise the algorithmic means of generating such informative, multi-tiered labeling captions. Although the labeling component of the system represents a post-cluster generation process involving the characterization rather than the generation of the clusters, it is arguably as essential as any of the other components of the recommendation system. Stated simply, if users cannot effectively understand and interpret the meaning of the labels, the quality of the system's clusters and document-cluster associations may not matter; the system may still fail to meet its intended performance objectives.

In the interest of clarity, it is also instructive to emphasize the fundamental underlying use case that motivates this work. It is to provide a robust and effective *complement* to search. Along with search, navigation and personalization, the recommendation approach we implement can greatly improve the efficiency and effectiveness of an otherwise comprehensive information retrieval system for legal research. Given a non-trivial legal domain-specific query, when a user performs a professional search, in addition to receiving ranked lists of search results delivered across multiple content types, the recommendation component of the system presents a supplemental list of ranked clusters, one for each of the key issues present in the selected document from the search results. In short, this cluster recommendation system is enabled by the document-to-cluster association process described in Section 4. The process by which the clusters have been topically defined and populated is described in an earlier work (Lu et al., 2011).

The recommendation system presented in the paper can thus be described as follows:

1. It identifies major issues discussed in individual documents from different content types;
2. It links each issue to the back-end repository of important legal issues via an association process;[2]
3. For every legal issue, it identifies, populates, and updates a set of the most important documents for that issue;
4. It provides a meaningful and sufficiently encompassing label for every issue in the universe.

By defining such a collective framework, we provide a powerful, coherent, and comprehensive approach to supplementing legal research results with additional relevant yet important documents in our issue-based recommendation system. Just as impor-

[1] In the interest of simplification, the expressions *legal issues* and *legal topics* will generally be used interchangeably.

[2] This back-end repository is sometimes referred to as the "universe" of legal issues.

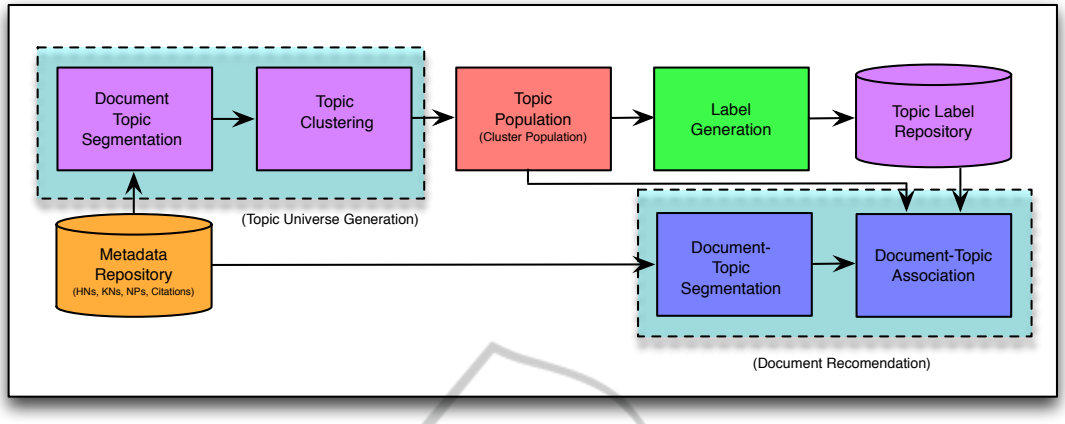

Figure 1: Workflow of document recommendation system.

tantly, the system will produce a more comprehensive coverage of the essential arguments associated with these legal issues.

Figure 1 illustrates the overall workflow of this recommendation system. At the highest level, the primary activities are represented by the two key components of the system: (1) define and generate the legal topic universe, and (2) perform document recommendations via associations (shown in the two light blue boxes delineated by dotted lines). As the subcomponents indicate, both the generation of the universe and the association processes rely on the document topic segmentation results that precede them.[3] Other significant components of the workflow include: (a) populating the clusters with the most important documents, depending on the topics they contain, and (b) labeling the clusters, which is an important piece of any outward rendering of the clusters. Each of these components will be discussed in the remainder of the paper in order to clarify their roles.

The paper is organized as follows. Section 2 briefly surveys related work. Section 3 gives a short description of the metadata available across the different content types in the legal domain. Section 4 presents the overall issue-based recommendation system, which includes a document topic detection algorithm and a document association algorithm. The labeling algorithm is covered in Section 5, followed by a description of the evaluation results and system performance in Section 6. Finally, Section 7 concludes the work while Section 8 discusses future research directions.

[3]The two "Document Topic Segmentation" boxes in Figure 1 are functionally the same.

2 RELATED WORK

Topic detection in documents can be applied at two different levels: at the individual document level, and at the document collection level. At the individual level, topic detection is often referred to as topic segmentation, which is to identify and partition a document into topically coherent segments. Algorithms in this category often exploit lexical cohesion information based on the fact that related or similar words and phrases tend to be repeated in coherent segments and segment boundaries often correspond to a change in vocabulary (Choi, 2000; Hearst, 1997; Utiyama and Isahara, 2001). Complementary semantic knowledge extracted from dictionaries and thesauruses, or additional domain knowledge such as the use of hyponyms or synonyms can been incorporated to improve the segmentation (Choi et al., 2001; Beeferman et al., 1997). Topic detection at the document collection level, on the other hand, is to identify underlying common topics among different documents. Many studies derive topics by document clustering, either using entire documents (Aggarwal and Yu, 2006) or sentences (Bun and Ishizuka, 2002; Chen et al., 2007). Topic modeling approaches, such as Probabilistic Latent Semantic Analysis (PLSA) (Hofmann, 1999), Latent Dirichlet Allocation (LDA) (Blei et al., 2003), and Non-negative Matrix Factorization (NMF) (Lee and Seung, 1999), take a more intuitive notion of topics in a collection by characterizing a topic as a cluster of words rather than documents. For example, Prasad, et al. (Prasad et al., 2011) applied a technology in dictionary learning known as NMF to decompose a document collection matrix represented by the vector space model into two non-negative matrixes, one of which was a term-topic matrix with each entry being the "confidence" of a term in a particular topic.

Due to the non-negative nature of the topical matrix, an original document can be reconstructed approximately by the combination of topics using only the allowed additive operation.

Labeling topics is another significant yet often somewhat neglected issue in topic detection research. Topics have traditionally been interpreted via top ranked terms based on either marginal probability in LDA or some statistical measurements such as tf.idf or tf.pdf (Bun and Ishizuka, 2002). A good overview of cluster labeling topics can be found in Jain (Jain et al., 1999). Historically less work has focused on clustering labeling than cluster generation itself. Popescul and Ungar obtained impressive results by combining $\mathbf{X}^2$ and the collection frequency of a term (Popescul and Ungar, 2000). Glover, et al. labeled clusters of Web pages by relying on the information gain associated with those pages (Glover et al., 2002a). Whereas Stein and zu Eissen use an ontology-based approach (Stein and zu Eissen, 2004), the more challenging problem of labeling nodes in a hierarchy (and the related general vs. specific problem) is addressed by Glover, et al. (Glover et al., 2002b) and Treeratpitu and Callan (Treeratpituk and Callan, 2006). More recently Fukumoto and Suzuki performed cluster labeling by relying on concepts in a machine readable dictionary (Fukumoto and Suzuki, 2011) with positive results. In another distinct recent work, Malik, et al., focused on finding patterns (i.e., labels) and clusters simultaneously as an alternative to explicitly identifying labels for existing clusters (Malik et al., 2010).

A comprehensive review of document recommendation systems, as part of broad-based recommendation systems, can be found in (Adomavicius and Tuzhilin, 2005). Such systems are usually classified into three categories, based on how recommendations are made: content-based recommendations, collaborative filtering recommendations, and hybrid methods. Content-based recommendations suggest similar documents based on the preferred ones by users in the past, and this type of approach has its roots in information retrieval and information filtering research. Often, both users and documents are represented by sets of features. Documents that best match the user profile and the profile of previously viewed documents get recommended. A news recommendation system recently described in (Li et al., 2010) is an example of such system. In addition, the authors also introduced an exploitation factor into the algorithm to bring new content to users' attention. The basic concept behind collaborative filtering is to utilize the preferences and evaluations of a peer group to predict the interests of other users. It has been successfully applied to suggest products in shopping sites, such as the one used in Amazon (Linden et al., 2003), but it was rarely used independently among document recommendation systems. Instead, collaborative filtering approaches often combined with content-based approaches into hybrid systems. Google news (Liu et al., 2010) adopted this hybrid approach, with heavy emphasis on the click log analysis on its rich history of usage data, to recommend personalized news to different users. Document recommendation technology has been successfully adapted to legal documents in recently years. Al-Kofahi et al. (Al-Kofahi and et al., 2007) described a legal document recommendation system blending retrieval and categorization technologies together. They designed a two-step process. In the first step a ranked list of suggestions was generated based on content-based similarity using a CaRE indexing system [4], a meta-classifier consisting of Vector Space, Bayesian, and KNN modules. In the second step, recommendations were re-ranked based on user behavior and document usage data. The recommendation system described in this paper differs from theirs in several respects: (1) it recommends documents based on important topics discussed in the original document and groups them as such, (2) it links the topics in the document to topics in a universe instead of each individual document, and (3) each topic in the document is presented with a precise and hierarchically structured label.

3 LEGAL DOMAIN CHARACTERISTICS

Documents in the legal domain possess some noteworthy 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.

3.1 Data and Metadata Resources

Legal documents in the U.S. tend to be complex in nature because they are the product of a highly analytical and often adversarial process that involves determining relevant law, interpreting such law and applying it to the dispute to which it pertains. 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. 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).

3.1.2 Case Law 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.

3.1.3 Headnote Classification, Key Number System

Headnotes are further classified to a legal taxonomy known as the West Key Number System,[4] a 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 a descriptive name, together with a hierarchically structured descriptor, known as a catchline. An example of a headnote and its key numbers (including catchlines) is shown in Figure 2.

3.1.4 Citation System

Legal documents contain rich citation information just as documents from other domains do, such as scientific publications and patents. These are the legal domain's equivalent to URLs. 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.

3.1.5 Statutes' Notes of Decisions

Another primary law source is statutes, which are laws enacted by a state legislature or Congress. For some statutes, both in federal and state levels, Notes of Decisions (NODs) are created for them. NODs are the editorially chosen case law documents that construe or apply the statute. In particular, NODs are tied to specific headnotes in the applying cases. An example of NODs is shown in Figure 2, with the blue hyperlink at the end of the headnote text (i.e., 8 U.S.C.A....).

4 RECOMMENDATION VIA CLUSTER ASSOCIATION

Given the complexity and multi-topical nature of so many legal documents, recommendation algorithms that treat documents as single atomic units tend to produce results that are topically unbalanced in the sense that recommended documents with different topical focuses are blended together in a single list, and it is left up to users to decipher which section of the original document is the target of the recommendation. To tackle this deficiency, the recommendation approach we have developed has a built-in topic detection algorithm that explicitly segments each document into coherent legal topics. Recommendations are thus tailored toward each individual topic. The algorithm consists of two steps: (1) segment documents into topics, and (2) associate topics with clusters of documents containing the most relevant like topics. The document segmentation algorithm is designed to accommodate various content types, and the association step is based on the similarity between topics and clusters.

As alluded to in the aforementioned second step, prior to the association process, a universe of topic clusters (i.e. topics) that has a comprehensive topical coverage in the legal domain needs to be defined such that topics detected in each individual document can be linked to this universe. Without this universal coverage, one could not envision issue-based recommendations, similar to "more like this" for the Open Web, where legal researchers could discover topics related to a document they are examining or probe

[4] http://store.westlaw.com/westlaw/advantage/keynumbers/

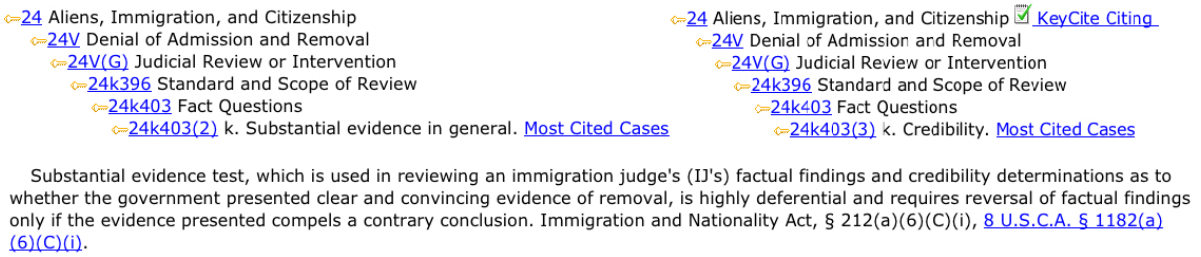

Figure 2: A example of a headnote with its assigned key number.

deeper into a topic of interest. It is also critical that most if not all legal documents (regardless of their type) be linked to these clusters. The clusters, described in prior work (Lu et al., 2011), are meant to contain the most important case law documents on a legal topic. Yet they are also populated with other types of legal documents such as statutes, regulations, and court briefs, to ensure comprehensive coverage. In short, the utility of the 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 coverage. As such, the cluster universe generation process is not part of the recommendation algorithm, but we mention it here because it is the foundation upon which our recommendation algorithm is built.

4.1 Topic Segmentation

The topic segmentation algorithm leverages available metadata from different document types. We found this approach to provide better quality results over traditional topic segmentation algorithms that rely upon lexical cohesion and utilize document text alone.

The segmentation algorithm differs depending on the availability of metadata in different document types. For the purposes of illustrating variations of the segmentation algorithm, we will subsequently have a brief discussion of case law documents (possessing headnotes), statutes, and court briefs.

4.1.1 Case Law Topic Segmentation

As stated before, headnotes are short summaries of points of law in case law documents. Collectively, they provide near-complete coverage of the main legal issues within a case. By grouping headnotes within a case law document based on their "similarities," it is possible to identify the main legal topics within a document. We use the vector-space model to represent headnotes in a case. A headnote is depicted in terms of four types of features: text, key numbers, KeyCite citations, and noun phrases. 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.

The similarity functions for the component vectors are defined using cosine similarity or one of its variations, and an agglomerative clustering algorithm grouping similar headnotes to generate the topics for a case law document. The algorithm merges two headnotes together while maximizing the following equations,

$$F = maximize \frac{\tau}{\varepsilon} \tag{4.1}$$

where,

$$\tau = maximize \sum_{r=1}^{k} \sum_{h_i \in T_r} sim(h_i, \overline{T_r}) \tag{4.2}$$

and

$$\varepsilon = minimize \sum_{r=1}^{k} n_r sim(\overline{T_r}, \overline{T}) \tag{4.3}$$

$$\overline{T_r} = \frac{\sum_{hinT_r} h}{n_r} \tag{4.4}$$

$$\overline{T} = \frac{\sum_{rinT} \overline{T_r}}{k} \tag{4.5}$$

where τ is the intra-cluster similarity and ε is the inter-cluster similarity, k denotes the total number of topics in a document, T denotes the topics for a document, T_r denotes an individual topic, and n_r is the number of headnotes in the topic T_r. Also, $\overline{T_r}$ and $\overline{T}$ 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.

4.1.2 Statutes Topic Segmentation

Unlike case law, federal or state statutes do not have headnotes; however, some of them contain NODs as stated in 3.1.5, and these NODs are tied to specific headnotes that construe or apply the statute. By grouping these headnotes in the same manner as stated in 4.1.1, legal topics inside the statutes can be identified.

4.1.3 Court Briefs Topic Segmentation

A majority of the court briefs do not contain headnotes, however they have rich citation links, both in-links (other documents citing this brief) and out-links (this brief cites other documents, mostly case law). Utilizing the headnotes available from these directly cited documents and grouping them in the same manner as stated in 4.1.1 can help identify main legal topics within a brief.

One distinction between the court briefs and case law documents, however, makes further topic segmentation refinement for briefs necessary. Case law opinions in general contain complete legal facts regarding a law suit, while court briefs have a much narrower focus. Most often they only deal with certain aspects of the complete set of facts. To help identify these specific aspects, a generic document summarization algorithm (Schilder and Kondadadi, 2008) is first applied to the brief text to extract the most important sentences. Then these sentences are sent to a key number classification system built with CaRE (Al-Kofahi et al., 2001) to retrieve major key numbers from it. By analyzing the commonalities among topics from the retrieved key numbers and the aforementioned grouped topics (that have associated key numbers as well) from cited case law documents, irrelevant legal topics can be filtered in a straightforward manner and the remaining ones will form the legal topics thereafter.

4.2 Recommendation Via Topic Association

The above topic segmentation algorithm produces legal topics for each document regardless its content type, and each topic serves as an anchor point (i.e., a source topic) for the recommendation algorithm to retrieve the most similar clusters in the universe. As stated earlier, each cluster has been pre-populated with the most important documents from different content types, namely, case law, briefs, statutes, regulations, and administrative decisions, jury verdicts, trial court orders, expert witness reports, pleadings, motions & memoranda and analytical documents. By assembling and organizing these documents for each individual segmented topic, issue-based recommendations can then be generated for that document.

We treat the retrieval of the most similar cluster of an anchor topic as a ranking problem such that candidates are first generated and sorted; then the top ranked cluster is picked at the completion of the process.

To generate candidate clusters, a document classification engine, CaRE, is used, although other indexing engines (e.g., Lucene) could be used. For each cluster, three classifiers are trained one per feature type headnote text, key numbers, and citation patterns.

Given an anchor topic, we use CaRE to retrieve a pool of candidate clusters. We then use a ranker SVM to determine which cluster in the candidate pool is most similar to the anchor topic. To do this, we represent each anchor-candidate pair in terms of a feature vector. The features include CaRE scores, and variations of the noun phrases and key number features described in 4.1.1. These similarity features are then used to train a ranker SVM to rank clusters in the candidate pool, and the top ranked cluster is selected as the recommendation for the anchor topic in the document. The same process is performed for each segmented topic in the document. A series of evaluations of this process is presented in Section 6.

5 CLUSTER LABELING

Considerable research and development effort was dedicated to the labeling algorithm for these clusters. The breadth and depth of this research – generating distinct labels for approximately 360K clusters – is beyond the scope of this paper and merits its own research report. Nonetheless, some of the key features explored for the labels in question are briefly discussed here. We harness a portion of a heading from an existing taxonomy [Key Number System (See Section 3)], express a bias towards shorter over longer labels, and strive to avoid any redundancy. These last two features were the direct result of early evaluation rounds which employed attorney domain experts and the feedback we received from the business unit supporting this research.

Our current labels generally consist of three segments. These segments range from general headings to more specific ones. The most effective number of segments was determined empirically. The current algorithm evolved from our experiments and represents a hybrid of two earlier versions. The first stage

consists of identifying the first two more general segments of the label. As stated, it leverages an existing taxonomy of case law topical classifications (multi-level classifications coming from the Key Number System). The second stage relies on a ranking process that orders noun phrases that were identified in the headnotes in the cases in the given cluster. The resulting algorithm uses both information from a seed case for a cluster and the case law documents with which the cluster itself has been populated. The basic idea is to pick the most representative noun phrases from the headnotes in the top N cases in the given cluster for which its key number(s) are the same as the clustered headnotes in the seed case. These are appended to the first two segments from the first stage (the ones consisting of the topic titles from the most popular key numbers in the top N cases). Earlier experiments revealed that noun phrases from headnotes classified to specific key numbers were more accurate than those from more generally assembled headnotes. Furthermore, the optimum number of top key numbers to use was determined empirically (i.e., more than 5).

The process can be summarized as follows:

- [Segments 1 and 2] Select the top two headings from the most popular key numbers;
- Get the headnote text from the top N cases with the same key number(s) as in the clustered headnotes in the seed case;
- Add back the clustered headnotes from the seed case;
- Use a noun phrase chunker on these text strings to identify optimal noun phrases (NPs); then rank the extracted NPs based on several simple features, such as the length of the NP (how many words), the combined within headnote term frequency (TF) for each of the word in the NP, the TF of the NP itself, the cross-collection document frequency (DF) of the NP, its inverse (IDF), etc.;
- [Segment 3] Pick the top NP after ranking, and append it to the top two headings from the key numbers above.

Based on an iterative manual assessment and error analysis process, we identified a set of potential enhancements that were worth investigating. These include:

1. suppressing redundancies across label segments to foster clarity and the absence of duplication within a label;
2. enforcing normalization across all term and doc statistics used (tf, idf, tf.idf, ...) to ensure consistency across the universe of clusters;
3. promoting label segments that contain more frequently occurring n-grams (frequently occurring substantive noun phrases);

Figure 3: A example of a set of clusters associated with a case on adoption. (Note the 2-3 tier layout).

4. creating a bias towards shorter vs. substantially longer labels to encourage clearer, more readable labels;
5. adding a labeling duplication detection step, and a priority-based process by which duplicate labels are methodically avoided.

The baselines used in our experiments include: (1) the titles from the most popular key numbers in the top N cases (termed 'original baseline'), and (2) the most popular (frequent) NPs in the top N cases. See Figure 3 for an example of a set of four cluster labels for a legal case on "adoption" where religion plays a pivotal role. Also see Table 3 for a comparative set of these results. Evaluations performed on the labeling process, which examines the potential contribution of several feature enhancements, is presented in the next section.

6 PERFORMANCE AND EVALUATION

To access the performance of our recommendation algorithm, we generated over 360,000 clusters from about 7 million U.S. case law documents, which collectively contain more than 22 million headnotes classified to approximately 100,000 key numbers. Each cluster is populated with the most important case law documents, as well as statutes, regulations, administrative decisions, and 6 other additional previously mentioned content types as its members. Over 100 million different documents have gone through the association process to generate recommendations. Over one thousand documents of different document types have been manually reviewed by legal experts in order to determine the quality of the recommendation, along with the labels for the associated clusters. The following results are based on these reviewed samples.

6.1 Evaluation Design

We report three different evaluations as follows:

6.1.1 Evaluation I: Association and Recommendation Quality

In Evaluation I, the quality of the clusters associated with input documents for different content types was graded by legal experts using a five-point Likert scale from A (high quality associated cluster) to F (low quality associated cluster). In addition, the top ranked recommended documents of different content types within those clusters have been scored using a coherence measurement, which is defined as the extent to which the documents in a given cluster address the same specific legal issue with respect to the anchor topic extracted via segmentation from the original source document. Furthermore, the recommended case law documents have been given a utility score, which is defined as the usefulness of the documents in the given cluster to a legal researcher. The rationale for assessing utility in addition to coherence is because it would be possible to have a cluster with a high coherence score which is not very useful to a legal researcher. For both metrics, the reviewers used a five-point scale ranging from 5 (high coherence with the current cluster's central topic or high utility to a legal researcher) to 1 (low coherence with the current cluster's central topic or low utility to a legal researcher).

6.1.2 Evaluation II: Label Quality

For the label evaluations performed, human domain experts (attorneys) with many months of clustering assessment experience provided the judgments. A five-point Likert scale was again used from A (highest quality) to F (lowest quality) based on accuracy and clarity. Specifically

A = on point, completely captures the topic .... (5)
B = captures the main topic in a general sense (4)
C = captures some of the main topical focus ... (3)
D = captures secondary or minor topical focus (2)
F = misses the topical focus of the cluster ..... (1)

6.1.3 Evaluation III: Composite Legal Report and Issue 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 domain experts. Further, they found that each of the 10 reports in this study had 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.

The algorithm was applied to the same set of reports to detect clusters associated with these documents. The objective of this assessment was to evaluate two things: (1) is the recommendation algorithm able to discover all the legal topics in these documents and (2) is the recommendation algorithm able to find the common legal issue in each of the reports.

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

- Precision, P, is defined as the number of correctly identified topics of a document (compared to number of manually identified topics) divided by the total number of topics, and
- Recall, R, is defined as the number of correctly identified topics of a document divided by the total number of manually identified topics of a document (the ground truth).

For the second objective, we used true positive rate, TPR, 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 document in all reports by experts.

6.2 Performance

6.2.1 Evaluation I

In Evaluation 1, legal experts were asked to grade the quality of 105 clusters associated with the topics identified in 25 source documents. During this exercise, 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 clusters receiving an A or B grade, and the "success rate" as the ratio of clusters receiving an A, B or C grade, to the total associated clusters, respectively.

Table 1 shows the association quality for case law documents, federal statues, and court briefs in ranked order. The precision and recall rate decline based on rank which is not unexpected. This behavior justifies our decision of picking the top ranked cluster as the output of the association process. In the court briefs category, we also show the quality difference between using and not using the summarization tool (which picked the first N-ranked sentences (e.g. N=10) from the document). As demonstrated by these results, summarization noticeably improved the quality of the top-ranked clusters. Table 1 also illustrates that in or-

Table 1: Association Quality on Rank.

	Expert Assessment							
	Case Law		Federal Statutes		Court Briefs w/ Summarization		Court Briefs w/o Summarization	
Rank	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
1	60%	93%	51%	86%	74%	95%	61%	87%
2	55%	91%	45%	86%	69%	95%	63%	88%
3	51%	91%	48%	85%	68%	94%	59%	87%
4	50%	90%	42%	85%	66%	90%	56%	88%
5	48%	89%	41%	84%	66%	90%	56%	88%

Table 2: Recommendation Quality Assessment.

Grade	Expert Assessment					
	Case Utility	Case Coherence	Analytical Coherence	Court Brief Coherence	Overall Recommendation per Topic	Overall Recommendation per Document
5	49	41	39	47	67	17
4	39	23	37	29	1	0
3	12	29	23	17	30	8
2	5	7	3	8	2	0
1	0	5	3	4	5	0
Average	4.2	3.8	4.0	4.0	4.2	4.4

der to achieve high recall, i.e., in the upper 80% range and above, it is a challenge to obtain comparable precision rates. These percentages are noticeably lower, with summary-aided briefs highest, followed by case law documents, and finally statutes.

Table 2 shows the quality assessment of 105 recommended topics from 25 documents represented by case law opinions, U.S. secondary law materials (a.k.a. analytical materials), and court briefs, in terms of coherence and utility, and the overall recommendation quality per topic level and per document level (additional scores to assess the overall recommendation quality across all document types). As one can observe from Table 2, the recommendation quality at the topic-level across document types remains high, in the 4.0 range. For case law documents, this is true both in terms of utility as well as coherence. Further, the overall recommendation quality at the topic-level and the document-level similarly remains high, above 4.0, in both instances.

6.2.2 Evaluation II

The following cluster labeling results were based on 100 final clusters which included 60 from an initial set of clusters associated with case law documents and another set of 40 from clusters associated with annotated statutes. They were produced while run in our newly developed operational environment. The label sets assessed included our original taxonomy-based baseline label set, our NP-based baseline label set, a set where redundancies across the label segments were eliminated, a set where features such as df were comprehensively normalized, a set where labels with more frequently occurring n-grams were promoted, and lastly, a set where a bias towards shorter labels was used.

As is evident from Table 3, performance of the original baseline label set was solid and in fact difficult to outperform. Although the initial baseline received some of the highest scores from the assessors, for being understandable and closer to human-like quality, it is also worth noting that these labels did not harness the computational effort that subsequent versions did in attempting to produce a readable and precise (i.e., more granular) three-segment label, and thus even though of high quality readability, they arguably do not contain the same degree of information as their more mature and labored counterparts in subsequent versions. That said, given other presentation-related factors such as a bias towards shorter labels, other results such as those produced for version 5 may be preferable. The other features tested in the label experiments proved not to be as consistently effective as anticipated.

6.2.3 Evaluation III

Regarding the first objective the algorithm's ability to discover legal topics within a document set the precision and recall of the system on 10 reports across different document types in shown in the Figure 4. Overall, the algorithm achieves reasonably high precision, but the recall in this instance was quite low, especially for case law documents. The main reason for this performance is the aggressive filtering, i.e., by

Table 3: Quality Assessment of Baseline and Hybrid Labels.

Grade	Original Baseline 1	Version 1 Baseline 2	Version 2 Redundundancy Suppression	Version 3 Normalized Features	Version 4 N-Gram Promotion.	Version 5 Short Lgth Bias
Average	3.68	3.65	3.65	3.08	3.52	3.66
A	18	16	16	18	19	19
B	47	46	45	20	38	41
C	23	27	29	55	26	29
D	9	9	8	18	14	9
F	3	2	2	2	4	2
Total	100	100	100	100	100	100

adopting much higher thresholds, in the post processing of the system in order to achieve high

For the second objective the algorithm's ability to identify a common legal issue within the document set Table 4 shows the performance of the system in each individual report, as well as overall.

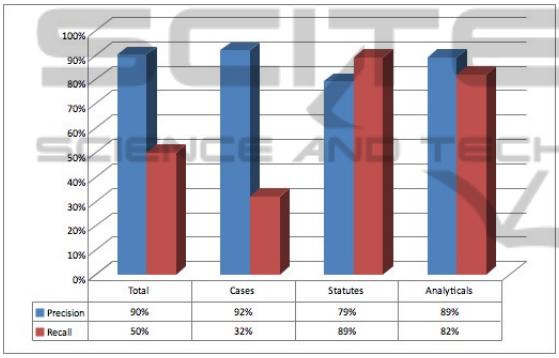

Figure 4: Precision and Recall of Evaluation III.

Table 4: Precision of common legal issues of Evaluation III.

Rpt IDs	Total No. Docs	Docs w/ Same Theme Ed. Gen. Assns	Docs w/ Same Theme Alg. Gen. Assns
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%

In this analysis, each of the 10 reports has a common legal issue running through it. However, this 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 (unshaded rows); our system identified a common thread through all documents in 6 of these 7 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 topics 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 58 documents (87.9%), representing a small but still appreciable 2% drop in TPR.

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

7 CONCLUSIONS

Document recommendation remains an active area of research containing a spectrum of challenging research problems as different applications have different needs. It is particularly challenging in the legal domain where documents are intrinsically complex, multi-topical, and contain carefully crafted, professional, domain specific language. Recommendations made in the legal domain require not only high precision but also high recall, since legal researchers cannot afford to miss important documents when preparing for a trial or related litigation proceedings. In addition, legal practitioners must familiarize themselves not only with primary arguments but with secondary or tertiary arguments associated with the legal issue as well. To this end, we describe in this paper an effective legal document recommendation system that relies on a built-in topic segmentation algorithm. The system is capable of high quality recommendations of important documents among different document types, recommendations that are specifically tailored to each of the individual legal issues discussed in the source document. The performance of the system is encouraging, especially given its validation by human legal experts through a series of different test assessments. Given these results, this paper makes three contributions to the field. First, it demonstrates how one can expand the set of original relevant legal documents using "more like this" functionality, one that

does not require explicitly defining legal issues and constructing queries. Second, by providing meaningful, hierarchically structured labels by way of our labeling algorithm for legal issues, we show that users can effectively identify interesting and useful topics. And third, the system is highly scalable and flexible, as it has been applied to on the order of 100 million associations across different document types.

Based on our studies, users, especially legal researchers, often prefer to have the ability to drill down and focus on key issues common 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 topically homogeneous but content-type heterogeneous, high quality document clusters, not to mention scalability are the chief characteristics of this issue-based recommendation system. It represents a powerful research tool for the legal community.

8 FUTURE WORK

Our future work will focus on improvements in the existing topic segmentation algorithm for documents which contain little metadata information. We have been experimenting with topic modeling algorithms, such as Latent Dirichlet Allocation (LDA) (Blei et al., 2003) and Non-negative Matrix Factorization (NMF) (Lee and Seung, 1999), in other related projects, and have witnessed very promising outcomes. Human quality labels remain a challenge since up until now, substantial manual reviews by human experts have been required to ensure quality. We are pursuing this subject as another future research direction.

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. We thank Khalid Al-Kofahi, Bill Keenan and Peter Jackson as well for their on-going feedback and support.

REFERENCES

Adomavicius, G. and Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17(6):734–749.

Aggarwal, C. C. and Yu, P. S. (2006). A framework for clustering massive text and categorical data streams. In *Proceedings of the Sixth SIAM International Conference on Data Mining (SDM 2006)*.

Al-Kofahi, K. and et al. (2007). A document recommendation system blending retrieval and categorization technologies. In *Proceedings of AAAI Workshop on Recommender Systems in e-Commerce*, pages 9–16.

Al-Kofahi, K., Tyrrell, A., Vachher, A., Travers, T., and Jackson, P. (2001). Combining multiple classifiers for text categorization. In *Proceedings of the 10th International Conference on Information and Knowledge Management (CIKM01)*, pages 97–104. ACM Press.

Beeferman, D., Berger, A. L., and Lafferty, J. D. (1997). A model of lexical attraction and repulsion. In *Proceedings of the 35th Annual Meeting of the Association of Computational Linguistics (ACL97)*, pages 373–380.

Bennett, J. and Lanning, S. (2007). The netflix prize. In *Proceedings of KDD Cup and Workshop*.

Blei, D. M., Ng, J. A., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Bun, K. K. and Ishizuka, M. (2002). Topic extraction from news archive using tf*pdf algorithm. In *Proceedings of the Third International Conference on Web Information Systems Engineering (WISE02)*, pages 73–82.

Chen, K.-Y., Luesukprasert, L., and cho Timothy Chou, S. (2007). Hot topic extraction based on timeline analysis and multidimensional sentence modeling. *IEEE Transactions on Knowledge and Data Engineering*, 19(8):1016–1025.

Choi, F. Y. (2000). Advances in domain independent linear text segmentation. In *Proceedings of the Applied Natural Language Processing Conference (ANLP00)*, pages 26–33.

Choi, F. Y., Wiemer-Hastings, P. M., and Moore, J. (2001). Latent semantic analysis for text segmentation. In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing (EMNLP01)*, pages 109–117.

Cohen, M. L. and Olsen, K. C. (2007). *Legal Research in a Nutshell*. Thomson West, Saint Paul, MN, 9th edition.

Fukumoto, F. and Suzuki, Y. (2011). Cluster labeling based on concepts in a machine-readable dictionary. In *Proceedings of the 5th International Joint Conference on Natural Language Processing*, pages 1371–1375. AFNLP.

Glover, E. J., Kostas, T., Lawrence, S., Pennock, D. M., and Flake, G. W. (2002a). Using web structure for classifying and describing web pages. In *Proc. of the World Wide Web*, pages 562–569. ACM Press.

Glover, E. J., Pennock, D. M., Lawrence, S., and Krovetz, R. (2002b). Inferring hierarchical descriptions. In *Proceedings of the 11th International Conference on Information and Knowledge Management (CIKM02)*, pages 507–514. ACM Press.

Hearst, M. (1997). Texttiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23:33–64.

Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proc. of 22nd Annual International SIGIR Conference*, pages 50–57. ACM Press.

Jain, A., Narasimha, M., and Flynn, P. (1999). Data clustering: A review. *ACM Computing Surveys*, 31(3):264–332.

Lee, D. D. and Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791.

Li, L., Chu, W., Langford, J., and Schapire, R. E. (2010). A contextual-bandit approach to personalized news article recommendation. In *Proceeding of World Wide Web Conference (WWW10)*, pages 661–671.

Linden, G., Smith, B., and York, J. (2003). Amazon.com recommendations: Item-to-item collaborative filtering. *IEEE Internet Computing*, 7(1):76–80.

Liu, J., Dolan, P., and Pedersen, E. R. (2010). Personalized news recommendation based on click behavior. In *Proceedings of the 15th International Conference on Intelligent User Interfaces (IUI10)*, pages 31–40.

Lu, Q., Conrad, J. G., Al-Kofahi, K., and Keenan, W. (2011). Legal document clustering with built-in topic segmentation. In *Proceedings of the 20th International Conference on Information and Knowledge Management (CIKM11)*, pages 383–392. ACM Press.

Malik, H. H., Kender, J. R., Fradkin, D., and Mrchen, F. (2010). Hierarchical document clustering using local patterns. *Journal of Data Mining Knowledge Discovery*, 21(1):53–185.

Popescul, A. and Ungar, L. H. (2000). Automatic labeling of document clusters. Unpublished MS Thesis.

Prasad, S., Melville, P., Banerjee, A., and Sindhwani, V. (2011). Emerging topic detection using dictionary learning. In *Proceedings of the 20th International Conference on Information and Knowledge Management (CIKM11)*, pages 745–754. ACM Press.

Schilder, F. and Kondadadi, R. (2008). Fastsum: Fast and accurate query-based multi-document summarization. In *Proceedings of the 46th Association for Computational Linguistics (ACL08)*, pages 205–208.

Stein, B. and zu Eissen, S. M. (2004). Topic identification: Framework and application. In *Proceedings of the 4th International Conference on Knowledge Management (KNOW04)*, pages 353–360.

Treeratpituk, P. and Callan, J. (2006). Automatically labeling hierarchical clusters. In *Proceedings of the 2006 International Conference on Digital Government Research (DG.O 06)*, pages 167–176.

Utiyama, M. and Isahara, H. (2001). A statistical model for domain-independent text segmentation. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL01)*, pages 499–506.