Parallel Tag Clouds to Explore and Analyze Faceted Text Corpora

Christopher Collins* University of Toronto

Fernanda B. Viégas, and Martin Wattenberg† IBM Research

ABSTRACT

Do court cases differ from place to place? What kind of picture do we get by looking at a country's collection of law cases? We introduce Parallel Tag Clouds: a new way to visualize differences amongst facets of very large metadata-rich text corpora. We have pointed Parallel Tag Clouds at a collection of over 600,000 US Circuit Court decisions spanning a period of 50 years and have discovered regional as well as linguistic differences between courts. The visualization technique combines graphical elements from parallel coordinates and traditional tag clouds to provide rich overviews of a document collection while acting as an entry point for exploration of individual texts. We augment basic parallel tag clouds with a details-in-context display and an option to visualize changes over a second facet of the data, such as time. We also address text mining challenges such as selecting the best words to visualize, and how to do so in reasonable time periods to maintain interactivity.

Keywords: Text visualization, corpus visualization, information retrieval, text mining, tag clouds.

1 INTRODUCTION

Academics spend entire careers deeply analyzing important texts, such as classical literature, poetry, and political documents. The study of the language of the law takes a similar 'deep reading' approach [29]. Deep knowledge of a domain helps experts understand how one author's word choice and grammatical constructs differ from another, or how the themes in texts vary. While we may never replace such careful expert analysis of texts, and we likely will never want to, there are statistical tools that can provide overviews and insights into large text corpora in relatively little time. This sort of 'distant reading' on a large scale, advocated by Moretti [21], is the focus of this work. Statistical tools alone are not sufficient for 'distant reading' analysis: methods to aid in the analysis and exploration of the results of automated text processing are needed, and visualization is one approach that may help.

Of particular interest are corpora that are faceted — scholars often try to understand how the contents differ across the facets. Facets can be understood as orthogonal, non-exclusive categories that describe multiple aspects of information sources. For example, how does the language of Shakespeare's comedies compare to his tragedies? With rich data for faceted subdivision, we could also explore the same data by length of the text, year of first performance, *etc.* Documents often contain rich meta-data that can be used to define facets: for example publication date, author name, or topic classification. Text features useful for faceted navigation can also be automatically inferred during text pre-processing, such as geographic locations extracted from the text [5], or the emotional leaning of the content [9].

In the legal domain, a question often asked is whether different court districts tend to hear different sorts of cases. This question is of particular interest to legal scholars investigating 'forum shopping' (the tendency to bring a case in a district considered to have a higher likelihood to rule favorably), and this was the initial motivation for this investigation. Our research question, then, is whether we can discover distinguishing differences in the cases heard by different courts. We address this question through examination of the written decisions of judges. The decisions of US Courts are officially in the public domain, but only recently have high-quality machine-readable bulk downloads been made freely available [19]. Providing tools to augment our understanding of the history and regional variance of legal decision making is an important societal goal as well as an interesting research challenge. Beyond our specific case study in legal data, we are interested in broader issues such as easing the barriers to overview and analysis of large text corpora by non-experts, and providing quick access to interesting documents within text collections.

Our solution combines text mining to discover the *distinguishing* terms for a facet, and a new visualization technique we call Parallel Tag Clouds (PTCs) to display and interact with the results (see Fig. 1). PTCs blend the visual techniques of parallel coordinate plots [15] and tag clouds. Rich interaction and a coordinated document browsing visualization allow PTCs to become an entry point into deeper analysis. In the remainder of this paper we will describe PTCs in comparison to existing methods of corpus visualization, the interaction and coordinated views provided to support analytics, our text mining and data parsing approach, and some example scenarios of discovery within the legal corpus.

2 BACKGROUND

2.1 Exploring Text Corpora

For the purposes of our work, we define facets in a corpus as data dimensions along which a data set can be subdivided. Facets have a name, such as 'year of publication' and data values such as '1999' which can be used to divide data items. Attention to faceted information has generally been focused on designing search interfaces to support navigation and filtering within large databases (*e. g.*, [11]). In faceted browsing and navigation, such as the familiar interfaces of Amazon.com and Ebay.com, information seekers can divide data along a facet, select a value to isolate a data subset, then further divide along another facet. For our purposes, we divide a document collection along a selected facet, and visualize how the aggregate contents of the documents in each subset differ.

While there are many interfaces for visualizing individual documents and overviews of entire text corpora *e. g.*, [3, 10, 33, 35], there are relatively few attempts to provide overviews to differentiate among facets within a corpus. Notable exceptions include comparison tag clouds [13] for comparing two documents, and the radial, space-filling visualization of [26] for comparing essays in a collection. Neither of these comparative visualizations focus on both visualization and appropriate text mining as a holistic analytic system, but rather use simple word counts to illustrate differences among documents. The work most related to PTCs is Themail [30], a system for extracting significant words from email conversations using statistical measures and visualizing them using parallel columns of words along a timeline. The visualization approach of PTCs shares the focus on discovering differentiating words within subsets of a corpus, and visualizes text along parallel columns of words. However, PTCs can reveal significant *absence*, or underuse of a word, as well as significant *presence*, or overuse. We augment the Themail approach with connections between related data subsets. PTCs are also visu-

*e-mail: ccollins@cs.utoronto.ca

†e-mail: {viegasf,mwatten}@us.ibm.com

IEEE Symposium on Visual Analytics Science and Technology
October 12 - 13, Atlantic City, New Jersey, USA
978-1-4244-5283-5/09/$25.00 ©2009 IEEE

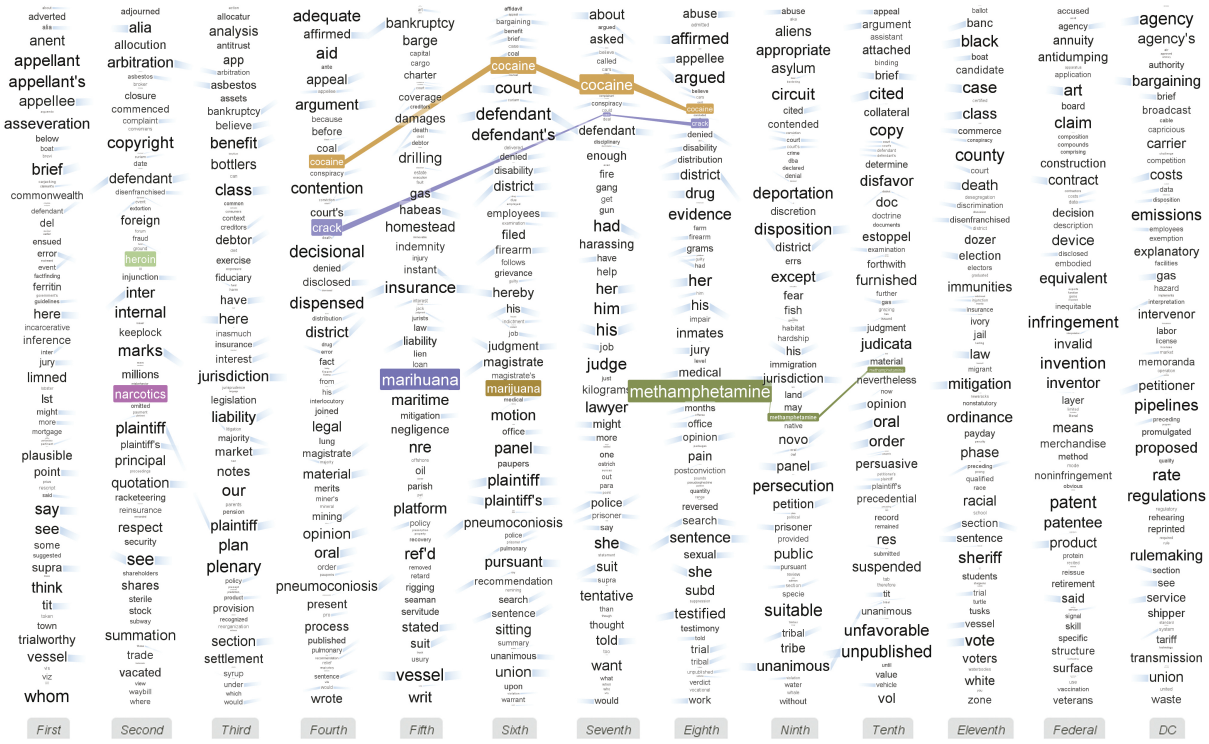

Figure 1: A PTC revealing the differences in drug prevalence amongst the circuits.

ally similar to the connected lists view of Jigsaw [28], however PTCs use size-weighting of words in the display.

Shneiderman and Aris [27] have previously explored the contents of a faceted legal document databases using matrix-based visualizations to reveal the number and type of data items matching each facet value. Our work differs in that we seek to aggregate and visualize the contents of the data items, not only their presence or absence. A matrix visualization approach would not be appropriate as our word-selection method, described later, seeks to maximize the differences between corpus subsets. Rather than the single vertical column of words that a words $\times$ facets matrix would contain, our approach allows the entire space to be filled with a wide variety of words. VisGets, or visualization widgets, have been used to explore faceted collections of web-based streaming data [5]. Facets are filtered using scented visual widgets [34] appropriate for the data type, providing both an overview of the available data items and a method to drill down along several facets simultaneously. A tag cloud VisGet consists of a traditional tag cloud summarizing all available documents — text differentiation along a facet is only achieved through interactive brushing. The goal of VisGets is to provide coordinated overview and navigation tools in a faceted information space, where our work is customized to providing meaningful differentiating overviews across facets within large amounts of textual data.

Finally, the Authorlines visualization [31] provides an overview of individual messages using arrays of circles, sized according to message length. We borrow this visual encoding and extend it to small multiples of bar charts in the document browser coordinated view, linked to the PTC.

2.2 U.S. Circuit Court Decisions

> "Jargon serves lawyers as a bond of union: it serves them, at every word, to remind them of that common interest, by which they are made friends to one another, enemies to the rest of mankind." *Jeremy Bentham* [2, 292]

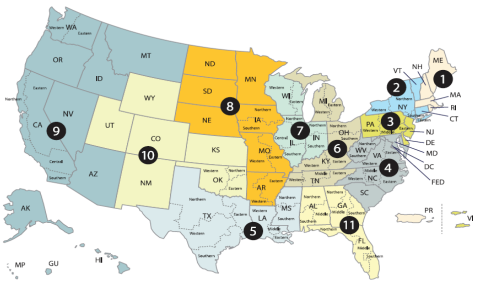

Figure 2: US Court *Circuits* are multi-state regions.

The words of the iconoclast Bentham were not the last written on the topic of legal language. Law and language meet in many academic ways: forensic linguists help solve crimes, judges make semantic rulings on unclear contract wording, and social scholars take a high-level view, studying the language of lawyers and judges [29]. By analyzing the written decisions of the US Circuit Courts of Appeal, we hope to shed light on thematic and potentially linguistic differences between subsets of the data. Differences in word usage between courts has been previously studied using legal databases as a source for historical lexicography [8]. However, in that work, text-based searches provided information on particular words of interest. Through text mining and visualization, we select words of interest and provide a broad overview as an entry point to deeper analysis.

The US Circuit Courts of Appeal are made up of 12 regionally-based court divisions (numbered First through Eleventh, plus the DC Circuit) and the Federal Circuit, which hears cases of national relevance, such as patent-related appeals (see Fig. 2). This data contains of 628,000 court decisions, each labeled by circuit. The judgments are faceted, because they can be organized along several dimensions, such as the lead authoring judge, the decision length, the date of the decision, or whether the lower court was upheld or overturned. For

our purposes, we parse the raw data and divide it into subsets by circuit, but we could equally well subdivide along other facets.

Each court decision is made up of several parts: the court name, the parties involved in the case, the date of the hearing, the date of the decision, the authoring and concurring judges, the main decision, optional concurring and dissenting opinions, and optional footnotes. In the data we obtained, most sections were pre-labeled in XML, but there were many errors, such as the date coded as the court name. The breaks between main opinion and consenting/dissenting opinions were not labeled. We cleaned the data by re-parsing the full text and labeling each section using regular-expression matching. Our visualization supports viewing PTCs by dividing the data along the 'court' facet and loading cases from a selected time period. Comparisons can be made between different courts across any of the textual parts of case data: the entire case, party names, main opinions, concurring opinions, dissenting opinions, and footnotes.

In the following description and accompanying illustrations we will discuss examples pertaining to discovering *distinguishing words* within the written decisions of each circuit court. While the method for discovering words of interest is discussed in detail in Sec. 4, it is sufficient for the following explanation to think of the selected words as characteristic or representative of the court in which they appear, when compared to the remainder of the corpus.

3 Parallel Tag Clouds

PTCs combine layout techniques from parallel coordinates with word-sizing techniques from tag clouds to provide a visualization for comparing large amounts of text. The basis of the visualization is the arrangement of words of interest into parallel columns, one for each distinct subset of the data across a facet of interest (see Fig. 1). Of several visual encodings tested by Bateman *et al.* [1], font size was the best way to convey importance, so we use it to encode a pre-assigned score. We scale by font size rather than the area of a word, as area gives undue visual prominence to short words. Words that are common across columns are connected by nearest-neighbor edges. Edges are drawn with varying width, sized relative to the words at the endpoints to reinforce relative size differences across columns. An important distinction between PTCs and parallel coordinates is that in PTCs edges may bypass a column (parallel coordinates axis) when a word is not present.

Through informal trials, we have found that the edges provide useful information about the degree of overlap among columns in general, but also tend to increase the complexity of the visualization and reduce the legibility of the words. To reduce the problem, all edges are drawn as 'stubs' that connect to each endpoint word and fade to transparency between the words. These edge stubs indicate the presence and direction of a connection, while not further cluttering the display and disrupting legibility. We clarify edges through interaction: when the pointer hovers over or selects a term, all occurrences of that term, and the connecting edges between them, are fully drawn and highlighted. Additionally, an entire column may be selected, making edges attached to this column visible and revealing all terms it shares with other columns. This provides the ability to *drill across* the corpus: by finding a term of interest in one column, one can easily discover others in which it is present and expand exploration laterally.

We experimented with two arrangements of words: alphabetical and ordered by size. Alphabetical arrangement was preferable for several reasons. While ordering by size offers the ability to identify the most significant words in each column (by reading across the tops), the layout is not space-efficient. Inter-column spacing must be wider as all the largest words cluster at the top. Alphabetical ordering distributes words of different sizes throughout the vertical space, allowing columns to be closer together. That is, we can place columns so close that two words at the largest size will overlap, because two words at the largest size are unlikely to be adjacent. Additionally, alphabetical ordering supports visual scanning to determine the presence of words of interest.

Figure 3: Sizing by score reveals that the Federal Circuit (far right) is the most different from other courts, and that the word 'patent' is overall the most differentiating word in the selected time period of the corpus.

3.1 Sizing by Rank and Score

In a dense visualization which has the special requirement that nodes are words which must be legible, maximal use of space is crucial. In order to maximize space usage, the default view of PTCs sizes words by their *rank* in the column. The result is that each column, assuming it has the same number of words, will have the same distribution of sizes, thus the same length. This provides for efficient usage of the vertical space and maximizes the average font size for all columns. However, information is lost about the relative magnitude of the underlying scores which generate the ranks. As we use a significance threshold to select words for the visualization, every word in the view is significant, so arguably maximizing size over precision may be preferable. Sizing proportional the maximum score across all columns may result in some columns becoming very small if their relative scores are small. This can be informative, and we provide this view as an option. For example, using our scoring function of how distinguishing a word is for a particular court, a shorter column would represent a court which is overall less different from the entire corpus than a court with a long column (see Fig. 3).

Outliers can be distracting to an analyst, and detrimental to the ability to visually to discriminate among other data items which are closer in value. Analysts often desire the ability to interactively remove distracting data items from view [32]. By right-clicking a word in the PTC, it is removed from the view, causing remaining items to rescale to fill the space.

3.2 Coordinated Views

The initial population of the PTC occurs by selecting a facet of interest to subdivide the corpus. In our implementation this is fixed: we subdivide by court. However, we only visualize courts of interest. For example, the Federal Circuit is quite different from the others, as it hears mainly patent-related cases, so it may be omitted. Data is also filtered by selecting a time period. We use scented widgets [34] to simultaneously allow for courts of interest to be selected while encoding how many cases are in that court for the selected time-frame. After selecting a data range, the tag cloud is populated with the top N words in each column, where N is pre-set to be maximal to allow for readable font sizes given the current window height, but may be adjusted. Larger values of N will introduce the need to scroll the visualization to explore the entire tag cloud.

Text overview visualizations are generally most useful if an analyst can interactively obtain additional information about specific words. As our implementation of PTCs draws on a large collection of documents, we provide the ability to select terms of interest and explore

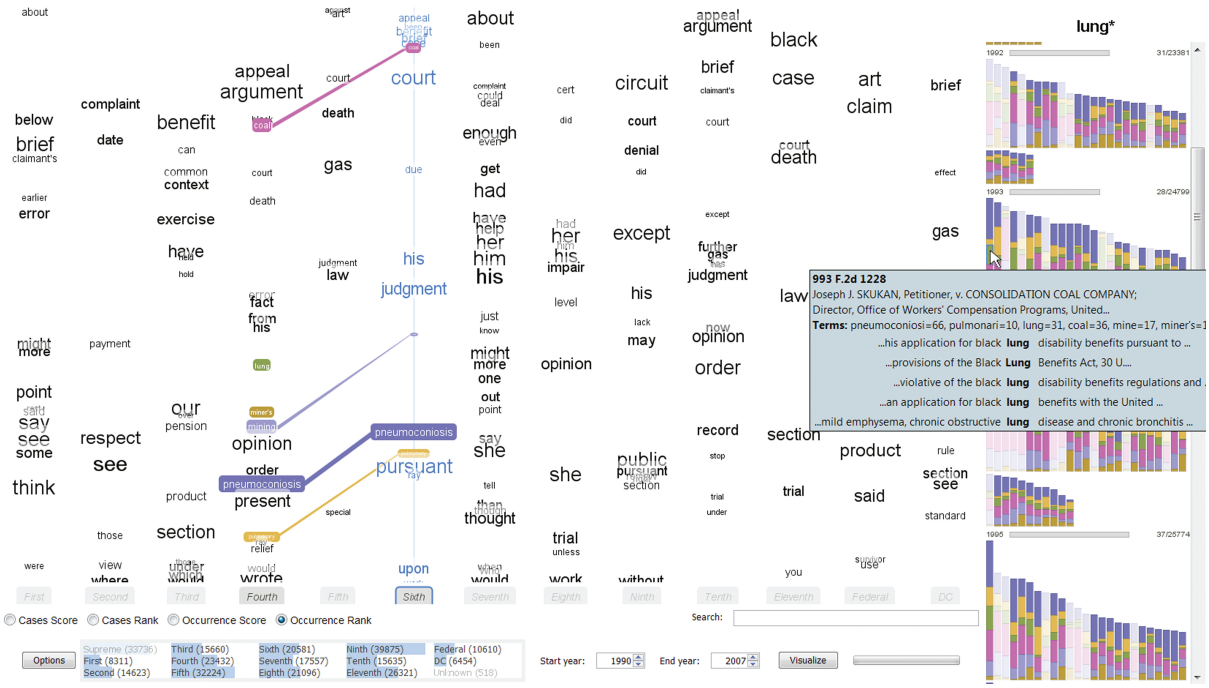

Figure 4: Selected terms in the PTC reveal litigation related to the coal mining-related disease pneumoconiosis in both the Fourth and Sixth Circuits. The document browser at right reveals the distribution of selected terms in individual cases over time. The Fourth and Sixth Circuits are selected in the PTC, causing documents from circuits other than these to become semi-transparent in the document browser. The mouse points at a section of a stacked document bar. The words in that document are enlarged on the PTC and all other words are removed. A tooltip shows examples of the hovered word used in context.

their distribution throughout the document collection. When a term of interest is selected, a coordinated document browser visualization is populated with bar charts representing the individual documents in which that term occurs, organized in rows by a second facet in the data, such as by year. The height of the bar is proportional to the number of occurrences of the term in that document. When multiple terms are selected, each is assigned a unique highlight color on the tag cloud, and the document glyphs become stacked bar charts. Multiple selections are treated as an AND query, preventing an overload of document results. Results are grouped by year and ordered largest to smallest. A maximum of 100 results per year are shown. To provide a complete picture of the results, horizontal 'distribution bars' beside the year labels show the relative number of documents matching the search terms and what portion of these are hidden.

Views are interactively linked: brushing across a document icon in the document browser highlights all the terms occurring in that document which are also in the PTC (see Fig. 4). Words are highlighted by increasing the font size and fading out words that are not in the document. Additionally, we highlight which corpus subset contains the document by drawing that column in blue. This interaction provides a lightweight form of document content overview, although only words which are already in the PTC are shown. Tooltips in the document browser reveal detailed case data, including the citation, parties, authoring judge, and a keyword-in-context table showing examples of the selected word in use [18].

We provide filtering of items in the document browser by selecting columns of interest in the PTC. When any column is selected, documents from non-selected columns become partially transparent (see Fig. 4, right). We retain the presence of faded document glyphs to give an indication of what proportion of total documents containing the selected terms come from the selected corpus subsets.

Finally, an analyst may wish to read a particular document in detail. By double-clicking a document glyph, the source document is opened in a web browser. Additionally, the full text of the document is visualized in a separate tab using a Many Eyes tag cloud.

Figure 5: Data changes are highlighted in orange. Here we see the emergence of 'methamphetamine' (second column from right) as we move from 1990–1995 to 1995–2000. 'Marijuana' is present in both time periods.

3.3 Revealing Change

As we provide interactive ways to filter the data backing the visualization, such as selecting a time range, we provide for visual highlighting of changes in the visualization when new data is loaded. New words can appear, for example, by selecting a different time period to extract from a large corpus, or by adjusting the method by which words are selected. When the data filters are adjusted, some words may be removed from view, while others are added. We visually highlight all deleted words and animate them out of view by increasing their size while simultaneously fading them to transparent. This provides a hint at what has been removed. In a second stage of animation, we reveal words that have been added. These remain highlighted until the analyst cancels the highlights through clicking an icon on the interface (see Fig. 5).

4 Mining Faceted Corpora

The most common approach to visualizing text as tag clouds is to count the word frequencies (*e. g.*, [7]). While this provides a relatively meaningful overview of a single text, or even a collection of texts treated as one, word frequency does not have sufficient distinguishing power to expose how subsets of a text collection differ. While one could compare multiple frequency-based tag clouds for subsets of a document collection, it is likely that these tag clouds will highlight similar words. If there is enough text in each subset, on the order of millions of words, each frequency-based tag cloud will start to approximate the distribution of terms in the domain generally. That is, the most common words will be similar in all data subsets. We would be unlikely, for example, to find much to distinguish among different court districts, where the legal language common among them will dominate. Such an approach may be appropriate for comparing text collections where dramatic differences in common terms were expected, or when similarities are desired.

The information retrieval community has long been interested in discovering words that make a document or collection of documents distinct from the background noise of a large corpus. These distinguishing terms are often given higher weight as index terms for a document, for example. Distinguishing terms have other uses, such as comparing corpora for similarity and homogeneity [17], or subdividing text automatically based on changes in distinguishing terms [12]. While there have been many uses for discovering distinguishing terms in a corpus in applications such as information retrieval and automatic summarization, interactive analysis tools for investigating distinguishing terms in a corpus have not been reported. In fact, Rayson and Garside [25] explicitly call for analyst intervention, claiming that simply identifying terms is not enough: human expertise is needed to understand why terms may be identified and if the reason is truly meaningful in the analysis context. They suggest 'the researcher should investigate occurrences of the significant terms using standard corpus techniques such as KWIC (key-word in context)'. Interactive visualization, such as PTCs, can offer more powerful analytic avenues for deeper investigation over standard corpus techniques.

There are a multitude of measures reported in the NLP community for scoring and ranking distinguishing terms, and indeed much argument about their relative quality (*e. g.*, [6, 17, 22, 25]). Measures such as TF-IDF [16] are commonly used to select distinguishing terms for a paragraph, document, or collection of documents. The Themail visualization [30] uses a variant of TF-IDF to collect distinguishing terms from a corpus of emails. While TF-IDF is an appropriate measure for detecting distinguishing words in a text sample against a reference corpus, it cannot highlight *significant absence*, nor do the scores it returns reflect a measure of significance for which there are reliable thresholds. A common word that does not appear in a document has a TF-IDF score of zero, the same as a rare word that does not appear.

Often, multiple metrics are applied in weighted combination or in sequence by re-ranking term lists. While multi-statistic methods may return improved results, the numerical scores are difficult to interpret. Indeed, the common practice is to heuristically choose a threshold and discard everything below it [14]. We choose to follow [25] and use a G^2 statistic, which is able to approximate χ^2 for words occurring 5 times or more. The G^2 metric can be interpreted as a measure of significance: higher G^2 corresponds a smaller p value. Or, to simplify: G^2 tells us the probability that the frequency of occurrence of a word in one corpus differs significantly from another.

For low frequency words, Dunning [6] shows that p values obtained using a G^2 statistic to lookup from a χ^2 tables can be off by several orders of magnitude. However, Moore [22] suggests a method to approximate p-values for low frequency events using the linear relationship between the negative of the natural logarithm of p-values computed from Fisher's exact test and log likelihood ratio scores.

Some [17, 25] have argued that applying the statistic in hypothesis testing is not appropriate given the non-random nature of text: some significant differences among texts is always expected, making the null hypothesis non-interesting. While this is certainly true for any two random documents, our texts are subsets of a larger corpus in the same domain, and each subset of text we compare consists of millions of words. With the increased sample size, the expectation that the subsets will converge on the same domain-specific overall word distribution grows. Thus, differences found may be significant. While hypothesis testing may be theoretically arguable for judging significance of G^2 scores, we follow [23] and use a $p < 0.01$ threshold of significance when visualizing distinguishing terms. This allows us to reduce the number of identified terms, as we cannot visualize all words, and to provide useful hints to an analyst comparing the relevance of terms identified by our statistical tests. The G^2 statistic is calculated using the following contingency table and equations:

	Target Subset	**Remainder of Corpus**	**Total**
$C(word)$	a	b	$a+b$
$C(other\ words)$	$c-a$	$d-b$	$c+d-a-b$
Total	c	d	$c+d$

$$E_1 = c*(a+b)/(c+d) \quad (1)$$

$$E_2 = d*(a+b)/(c+d) \quad (2)$$

$$G^2 = 2*(a*ln(a/E_1)+b*ln(b/E_2)) \quad (3)$$

where $C(word)$ is the count of the target word, and E_1 and E_2 are the expectation values for the word frequency in the target subset and the remainder of the corpus respectively. To find a significance level of $p < 0.01$, we use Moore's conservative approach, without assuming the > 5 word occurrences needed for reliable approximation by χ^2 tables:

$$G^2 \approx -2*ln(p)+2.30 \quad (4)$$

which gives us a G^2 threshold of 11.15. We employ a Sidak correction for repeated testing to adjust the significance levels. We assume $50,000$ repeated trials (the approximate number of word forms compared on a typical run of our system) and adjust p as follows:

$$p\prime = 1-(1-p)^{1/k} \quad (5)$$

where $p\prime$ is the adjusted level of significance, and k is the number of trials. This gives us an adjusted $p\prime$ of $2.01*10^{-7}$, which has a corresponding G^2 cutoff of 33.13, which we use as the threshold in our significance testing. If $a < E_1$, we know the statistic represents a lower than expected frequency of occurrence, otherwise the actual occurrence is higher than expected.

While our prototype of PTCs uses the G^2 statistic, our visualization is neutral to the scoring method applied to the terms: the visual techniques would work equally well for a frequency-based metric as for the frequency-profiling techniques we have described.

4.1 Occurrence and Case-Based Scoring

Experiences with Themail [30] revealed that techniques for identifying distinguishing words are prone to identifying words which are highly occurring in a particularly long document, but may not be distributed throughout the corpus subset under investigation. For example, in our analysis, 'voters' was identified as a distinguishing term for the Fifth Circuit, however, further investigation revealed a single very lengthy decision on an election-related class action which used the word 'voters' extensively. While this may be of interest to an analyst, it is important to support easy discovery of terms which have

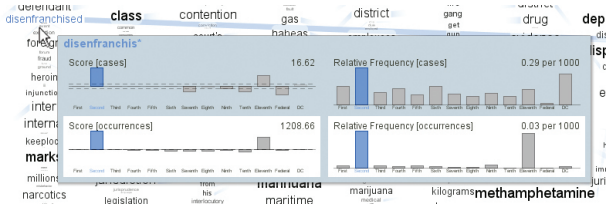

Figure 6: The bar chart tooltip provides word score details.

high occurrence but low distribution within the corpus subset. To address this, we measure two G^2 scores for each word: an occurrence-based score, and a case-based score. In the case-based measure, we populate the G^2 contingency table by counting how many individual documents (court cases) the word appears in at least once. As we will demonstrate in the analysis, the case-based measure identifies terms which occur in a larger than expected *number of cases* in a corpus subset, rather than an absolute number of occurrences. Both measures have analytical significance and reveal complementary information about a corpus. We provide for viewing PTCs based on either measure but we also provide for interactive tools to allow for the two forms of score to be compared for a particular word of interest. Additionally, the document browser can quickly reveal the distribution of a selected word within the corpus.

4.2 Data-Rich Tooltips

While our visualization can only reveal a limited number of words per parallel column, our word scoring measures assign values for all words for all corpus subsets. For example, our measure of the distinguishing nature of a term can identify words which occur more often than expected, or less often than expected. Due to space considerations, we choose to only show words which occur more often than expected. We also calculate occurrence and case-based measures, but can size the tag cloud based on only one. We provide for data-rich graphical tooltips which use bar charts to reveal the score and the normalized frequency of occurrence for a term across all subsets of the corpus, for both occurrence- and case-based measures. The column in which the word under the mouse appears is highlighted in blue to provide easy reference. Threshold lines reveal the G^2 significance threshold, and bars below the threshold are faded out. These tooltip graphs can quickly reveal where a word which is distinguishing in a particular corpus subset is unusually unpopular in another, and whether a term identified using occurrence-based scoring also appears in a significantly high number of cases in the selected court.

In Fig. 6, we show a tooltip created by hovering on the word 'disenfranchised' in the Second Circuit. We can see that this term has a significantly high score for both the Second and Eleventh Circuits when based on the occurrence count, and occurs less than expected in the Third through Eighth Circuits (bottom left). Note that the significance bars are at the baseline due to the large scale, so are not visible. However, based on the case scores (top left), only the Second Circuit has a significant score. This indicates that the occurrence-based score in the Eleventh Circuit must be due to a few cases with a high number of mentions of this term.

4.3 Data Filtering

PTCs, as with any word-based visualization, cannot reveal all the words in a given corpus given typically limited screen resolutions. Significant filtering is necessary. In order to provide for interactive visualization, we carry out several filtering steps at the pre-processing stage. We optionally remove listed 'stop words' from the data — words like 'the', 'and' that do not often carry meaning. Domain-specific stop words are identified as the top 0.5 percentile by overall number of documents they occur in, and removed. This captures terms such as 'judge', 'court', and 'circuit' in our data. This filtering is optional because in linguistic study these common words can be very informative if they are unevenly distributed across a corpus.

To further reduce the data size, we identify the word frequency at the 40th percentile when words are sorted ascending by overall occurrence count. We then remove all terms with overall frequency below this cut-off. The 40th percentile was selected to remove much of the 'long tail' of terms which are unlikely to be identified as distinguishing — most words removed only occur once or twice in the entire dataset. Our trials have shown that the vast majority of terms with G^2 scores above the significance threshold have frequency > 7. This achieves a vast reduction in the number of terms for which G^2 scores much be calculated at run-time, resulting in a significant speed increase and memory savings with no change to the visualized output.

To reduce the data size further, we also optionally remove words beginning with an upper case letter which do not start a sentence ('initial uppers'). Identifying initial uppers is a quick way to approximate proper noun detection in English. Aside from reducing the data, this technique was necessary to remove place and judge names from the visualization. Initial prototypes revealed that the highest scoring terms were almost exclusively proper nouns. These terms are not informative, as we expect the names of states and cities within a circuit, or the names judges writing decisions in that circuit, to be distinguishing. While this was a useful sanity test on our technique, we removed these terms in the current version. Proper nouns are interesting, however, when viewing the distinguishing terms in the 'parties' section of the case data, as common litigants are identified.

4.4 Reverse Stemming

In order to merge word forms with the same root, such as 'jurisdiction' and 'jurisdictions', we perform stemming using the Lucene Snowball Stemmer, an implementation of the Porter stemming algorithm [24]. However, the stemming algorithm strips all suffixes, leaving, for example 'jurisdic'. While this is acceptable for counting purposes, we discovered with early prototypes that it is surprisingly difficult to read a text visualization consisting of word stems. As a result, during data pre-processing, we count all occurrences of (word,stem) pairs generated by the stemmer, and retain the most common mapping for each stem. Then, as a final pre-processing step, we reverse the stemming on each term vector using the most common mapping. Thus the visualization shows real words.

As an interesting side-effect, the word forms shown in PTC reveal the most common form of each word within the dataset. We were interested to note that most verbs appear in their past tense form, such as 'averted' and 'insisted', but some appear in present tense, such as 'disagree' and 'want'. By selecting these words in the tag cloud and examining KWIC views for the associated documents, we found a separation between discussion of the facts of a case 'the plaintiff averted the problem', 'the district judge erred when she insisted' and the commentary of the judges 'I disagree with my colleagues because', 'We want to reinforce'.

4.5 Visual Variations

The G^2 score used to identify distinguishing terms provides information about significant *lack* of a word, as well as an unusually high presence. Through graphical tooltips, we provide both positive and negative scores for terms which are present in the tag cloud. However, what if a term is unexpectedly low in a circuit, but does not appear on the tag cloud because it is not high in another other circuit? A tooltip will not help. To address this, we provide a view which selects the top N words per column by absolute value. Words are sized and ranked by the absolute value of the score. Negatively scoring terms are distinguished by a red hue. In Fig. 7 we see 'patent' scores significantly low in all but the Federal and DC Circuits. Perhaps more interestingly, we see 'dissenting' in the First Circuit, revealing that dissenting opinions are provided in that circuit significantly less often than expected.

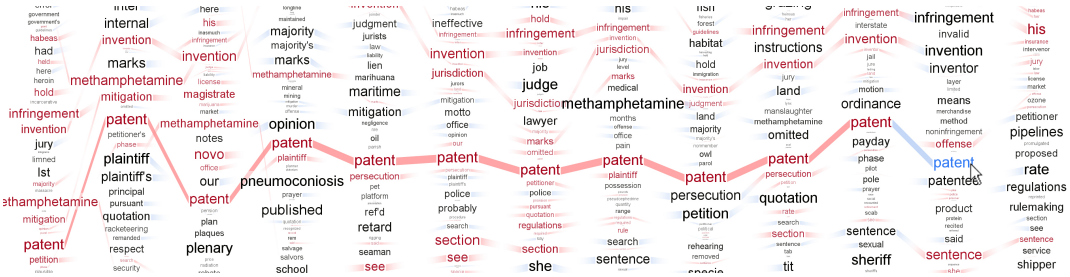

Figure 7: In a variation on the visual encoding, we reveal both significantly high and significantly low scores. Lower than expected occurrences are red, and higher than expected are black (blue on hover).

Extending our approach to two-word phrases brings several challenges. Tracking multi-word phrases results in an exponential growth in the dataset, and we have more data to fit onto the display space while maintaining legibility. However, we have experimented with two-word phrases using the existing PTCs implementation, finding differences in verb usage, such as 'unanimously finds' in the Sixth Circuit compared to 'unanimously agrees' in the Ninth.

5 Implementation

In order to quickly analyze selected subsets of a large corpus such as the history of the US Circuit Courts of Appeal, significant data preparation is necessary. Our implementation, written in Java, makes use of the open-source Lucene search engine, both for its search capabilities, and for the {word,count} *term vectors* it stores to support search. In a data pre-processing step we extract parts of each case and pass them to Lucene for indexing, stemming, and optional initial upper removal. We also collect the document ID, court ID, and date for each document into a Postgresql database. In further pre-processing the term vectors are retrieved from Lucene for each document in the dataset. A module called the Term Vector Composer takes the term vectors for each document, along with the court ID and date, and creates yearly summary vectors of {stem,count[ ]} where count[ ] is an array of counts across each court. The summary vectors are filtered to remove stop words, then written to the disk. At runtime, selected year vectors are further composed into a single term vector representing all words in a time range. This is passed to the scoring module and on to the reverse stemmer and visualization.

The indexing and term vector preprocessing operations take approximately 10 hours using a 2.53GHz dual core processor with 3GB memory. Preprocessing document term vectors into year vectors reduces the number of composition operations at runtime by a factor of 10,000 (but reduces the time resolution to years). Retrieving and composing term vectors from the disk takes approximately 3 seconds per year, with the majority of this time spent on disk operations. As even 30 seconds feels like a long wait for a visualization to be populated with a 10 year span, after the initial view is presented, we use a background thread to pre-cache term vectors for 10 years on either side of a selected time range. When the visualization is closed, the summary vector and the years it contains are saved. If the same time range is later requested, only one load operation is necessary.

6 Analysis

As we developed this visualization, we worked with two legal experts. In this section we describe some of the phenomena that were revealed through usage of the system. We do not claim these as original discoveries, but rather as examples of how PTCs can point to a range of interesting patterns in a real-world data set.

National vs. Regional Issues

Because of the arrangement differences between court districts, case law can reveal geographic cultural (and criminal) variations. For instance, drug-related terms appear in most circuits, revealing the national dimension of this problem. Closer inspection, however, uncovers distinct regional flavors: methamphetamine seems to plague midwestern and western states the most, appearing in the Eighth, Ninth, and Tenth circuits. Cocaine cases afflict the East, emerging in the Fourth, Sixth, Seventh, and Eighth circuits. Heroin cases are concentrated in the Second circuit, which includes New York. These differences might point to either a regional variation in drug use, or perhaps the level of prosecution (see Fig. 1).

Issues challenging a particular jurisdiction are revealed through the data. For instance, 'deportation' shows up in the Fifth Circuit, which includes Texas, the state with largest crossing border in the U.S., 'gun' appears in the Seventh Circuit, whereas 'copyright' shows up in the Second Circuit, which includes New York. The common occurrence of words, shown through edges in the PTC, can reveal similarities as well as differences. For example, in Fig. 4, we see that the Fourth and Sixth Circuits are similar by virtue of common terms: coal, mining, pneumoconiosis. These similarities make sense since the two circuits are adjacent and share some of the largest coal reserves in the country.

Language variation

Court cases can also provide insight into variations in legal vocabulary and linguistic idiosyncrasies of a particular court. For example, we discovered the odd words 'furculum', 'immurement', and 'impuissant', all in the First Circuit. By revealing the cases in the document browser and isolating the First Circuit, we see that almost all occurrences of these terms originate from a single judge, Judge Selya. A follow up web search revealed that Judge Selya is well known for his rich and often obscure vocabulary. One legal expert we consulted was fascinated by the potential to use these distinctive pieces of vocabulary as markers to track the influence of a particular judge. For example, the expert pointed to the presence of the word 'ostrich' in the Seventh Circuit. 'Ostrich' here refers to the 'ostrich instruction', shorthand for a particular directive to juries. This term was used almost — but not entirely — exclusively by that circuit over the past 10 years. Our legal expert pointed to this as meaningfully idiosyncratic piece of vocabulary that might be used to track attitudes toward jury instructions.

Forum Shopping

When bringing a lawsuit, a plaintiff sometimes has a choice between several possible venues. The natural tendency to pick the venue whose judges are historically most favorable to the plaintiff's case is known as 'forum shopping.' This phenomenon stands out clearly using our tools. For example, we can easily see one class of forum shopping by examining data from the years leading up to the creation of the Federal Circuit. The term 'patent' appears for the Seventh Circuit in the period 1970–1980, highly scoring on both occurrence and case-based measures. This is an accurate reflection of legal history: The Federal Circuit was created to combat the varying treatment given to patent rights in the circuit courts; the Seventh Circuit was one of the preferred venues [4].

7 Conclusion

Visual extensions to this technique present interesting future research challenges. The ordering of axes is an important factor when designing a parallel coordinates view. In this work, we took the approach that the data contains a semantic relation (the ordering of the circuits from First to Eleventh). Disrupting semantically meaningful arrangements is potentially problematic for a user [20]. For other data sets, automatic column reordering may be appropriate, or a facility for interactive reordering could be provided. Our instantiation of PTCs includes change highlighting through color. Additional methods to reveal change are needed, particularly to reveal which terms are removed from view when a parameter changes.

PTCs present a method for visualizing differences across facets of a large document corpus. Combined with text mining techniques such as measures of distinguishing terms, this approach can reveal linguistic differences. We have applied the technique to legal data, but many additional application areas exist. For example, PTCs could be used for visualizing homogeneity in a corpus: are linguistic differences discovered where homogeneous language is expected? Other applications include comparisons of individuals, such as collections of academic writing divided by author, or customer service call transcripts divided by employee. Additional lexical filters could also be applied, such as filtering based on part-of-speech or semantics. Finally, the comparison text could be changed: instead of viewing a corpus subset against the whole, we could compare, for example, a blog against the web-as-a-corpus. Along with improved visual encodings, these options are exciting directions for future work.

References

[1] S. Bateman, C. Gutwin, and M. Nacenta. Seeing things in the clouds: The effect of visual features on tag cloud selections. In *Proc. of the ACM Conf. on Hypertext and Hypermedia*, 2008.

[2] J. Bowring, editor. *The Works of Jeremy Bentham*, volume 7, page 282. Thoemmes Continuum, 1843.

[3] C. Collins, S. Carpendale, and G. Penn. Docuburst: Visualizing document content using language structure. *Computer Graphics Forum (Proc. of the Eurographics/IEEE-VGTC Symposium on Visualization (EuroVis))*, 28(3):1039–1046, 2009.

[4] D. Crouch. Forum shopping in patent cases [online]. 2006. Available from: `http://www.patently.com/patent/2006/07/forum_shopping_.html`.

[5] M. Dörk, S. Carpendale, C. Collins, and C. Williamson. VisGets: Coordinated visualizations for web-based information exploration and discovery. *IEEE Transactions on Visualization and Computer Graphics (Proc. of the IEEE Conf. on Information Visualization)*, 14(6):1205–1213, Nov./Dec. 2008.

[6] T. Dunning. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74, 1993.

[7] J. Feinberg. Wordle: Beautiful word clouds [online]. 2008 [cited 2 December, 2008]. Available from: `http://www.wordle.net`.

[8] S. Fred R. The politically correct US Supreme Court and the motherfucking Texas Court of Criminal Appeals: Using legal databases to trace the origins of words. In *Language and the Law: Proceedings of a Conference*, pages 367–372. William S. Hein & Co., 2003.

[9] M. L. Gregory, N. Chinchor, P. Whitney, R. Carter, E. Hetzler, and A. Turner. User-directed sentiment analysis: Visualizing the affective content of documents. In *Proc. of the Workshop on Semtiment and Subjectivity in Text*, pages 23–30. ACL, 2006.

[10] S. Havre, E. Hetzler, P. Whitney, and L. Nowell. ThemeRiver: visualizing thematic changes in large document collections. *IEEE Transactions on Visualization and Computer Graphics*, 8, Jan. 2002.

[11] C. G. Healey. Perception in visualization. Website, 2007. Available from: `http://www.csc.ncsu.edu/faculty/healey/PP`.

[12] M. A. Hearst and C. Karadi. Cat-a-cone: an interactive interface for specifying searches and viewing retrieval results using a large category hierarchy. In *Proc. of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 246–255. ACM Press, 1997.

[13] IBM Research. Many eyes comparison tag cloud [online]. 2009 [cited 25 March, 2009]. Available from: `http://manyeyes.alphaworks.ibm.com/manyeyes/page/Tag_Cloud.html`.

[14] D. Z. Inkpen and G. Hirst. Acquiring collocations for lexical choice between near-synonyms. In *Unsupervised Lexical Acquisition: Proc. of ACL SIGLEX*, pages 67–76, 2002.

[15] A. Inselberg. The plane with parallel coordinates. *Visual Computer*, 1(4):69–91, 1985.

[16] K. S. Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21, 1972.

[17] A. Kilgariff and T. Rose. Measures for corpus similarity and homogeneity. In *Proc. of the Conf. on Empirical Methods in Natural Language Processing*, pages 46–52, 1998.

[18] H. P. Luhn. Keyword-in-context index for technical literature. *American Documentation*, 11(4):288–295, 1960.

[19] C. Malamud. Us federal reporter 2nd and 3rd ed., bulk download [online]. June 2008. Available from: `http://bulk.resource.org/courts.gov/c`.

[20] K. Misue, P. Eades, W. Lai, and K. Sugiyama. Layout adjustment and the mental map. *Journal of Visual Languages and Computing*, 6:183–210, 1995.

[21] F. Moretti. *Graphs, Maps, Trees*. Verso, 2005.

[22] M. R. Morris, K. Ryall, C. Shen, C. Forlines, and F. Vernier. Beyond "social protocols": Multi-user coordination policies for co-located groupware. In *Proc. of Computer-Supported Cooperative Work*, 2004.

[23] M. Mueller. Comparing word form counts (WordHoard documentation) [online]. 2008 [cited 20 August, 2008]. Available from: `http://wordhoard.northwestern.edu/userman/analysis-comparewords.html`.

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

[25] P. Rayson and R. Garside. Comparing corpora using frequency profiling. In *Proc. of the Annual Meeting of the Association for Computational Linguistics Workshop on Comparing Corpora*, pages 1–6, 2000.

[26] M. Rembold and J. Späth. Graphical visualization of text similarities in essays in a book [online]. 2006 [cited 10 August, 2006]. Available from: `http://www.munterbund.de/visualisierung_textaehnlichkeiten/essay.html`.

[27] B. Shneiderman and A. Aris. Network visualization by semantic substrates. *IEEE Transactions on Visualization and Computer Graphics (Proc. of the IEEE Symp. on Information Visualization)*, 12(5):733–740, Sept.–Oct. 2006.

[28] J. Stasko, C. Görg, Z. Liu, and K. Singhal. Jigsaw: Supporting investigative analysis through interactive visualization. In *Proc. of the IEEE Symp. on Visual Analytics Science and Technology (VAST)*, pages 131–138, 2007.

[29] P. M. Tiersma. *Legal Language*. University of Chicago Press, 1999.

[30] F. B. Viégas, S. Golder, and J. Donath. Visualizing email content: Portraying relationships from conversational histories. In *Proc. of the SIGCHI Conf. on Human Factors in Computing Systems*, 2006.

[31] F. B. Viégas, M. Wattenberg, and K. Dave. Studying cooperation and conflict between authors with history flow visualizations. In *Proc. of the SIGCHI Conf. on Human Factors in Computing Systems*, pages 575–582. ACM Press, 2004.

[32] M. Wattenberg. Visual exploration of multivariate graphs. In *Proc. of the SIGCHI Conf. on Human Factors in Computing Systems*, pages 811–819, 2006.

[33] M. Wattenberg and F. B. Viégas. The word tree, and interactive visual concordance. *IEEE Transactions on Visualization and Computer Graphics (Proc. of the IEEE Conf. on Information Visualization)*, 14(6):1221–1229, Nov/Dec 2008.

[34] W. Willett, J. Heer, and M. Agrawala. Scented widgets: Improving navigation cues with embedded visualizations. *IEEE Transactions on Visualization and Computer Graphics (Proc. of the IEEE Conf. on Information Visualization)*, 2007.

[35] J. A. Wise, J. J. Thomas, K. Pennock, D. Lantrip, M. Pottier, A. Schur, and V. Crow. Visualizing the non-visual: spatial analysis and interaction with information for text documents. In *Proc. of the IEEE Symp. on Information Visualization*, 1995.