Multi-dimension Diversification in Legal Information Retrieval

Marios Koniaris[1(✉)], Ioannis Anagnostopoulos[2], and Yannis Vassiliou[1]

[1] KDBS Lab, School of ECE,
National Technical University of Athens, Athens, Greece
mkoniari@dblab.ece.ntua.gr
[2] Department of Computer Science and Biomedical Informatics,
University of Thessaly, Lamia, Greece

Abstract. The number of freely available legal data sets is increasing at high speed. Citizens can easily access a lot of information about regulations, court orders, statutes, opinions and analytical documents. Such openness brings undeniable benefits in terms of transparency, participation and availability of new services. However, legal information overload poses new challenges, especially in the field of Legal Information Retrieval. Search result diversification has gained attention as a way to increase user satisfaction in web search. We hypothesize that such a strategy will also be beneficial for search on legal data sets. We address diversification of results in legal search by introducing legal domain specific diversification criteria and adopting several state of the art methods from the web search, network analysis and text summarization domains. We evaluate our diversification framework using a real data set from the Common Law domain that we subjectively annotated with relevance judgments for this purpose. Our findings reveal that web search diversification techniques outperform other approaches (e.g. summarization-based, graph-based methods) in the context of legal diversification, as well as that the diversity criteria we introduce provide distinctively diverse subsets of resulting documents, thus differentiating our proposal in respect to traditional diversification techniques.

1 Introduction

Over the last years, as a result of the momentum of Open data initiatives, there has been a vast increase on the number of freely available legal data sets. Portals that allow users to search for legislation, using keywords, titles, etc. are now a common place. In such portals, legal documents are not stored as plain text, but in a more structured format with a rich set of meta data. Thus, it is possible for the end users to navigate to a specific section of a document or to inquiry information about the documents, such as date of enactment, date of repeal, jurisdiction, etc. Furthermore, with the advent of methods for the semantic indexing of Legal documents [31], several orthogonal categorization schemes can help users to find the information they need via navigation. To alleviate the

W. Cellary et al. (Eds.): WISE 2016, Part I, LNCS 10041, pp. 174–189, 2016.
DOI: 10.1007/978-3-319-48740-3_12

data overload problem, in this paper we propose a novel way to efficiently and effectively diversify legal documents.

Legal text retrieval, in contrary to web retrieval, is primarily based upon concepts and not the explicit wording in documents texts. Earlier works essentially focus on classifing sources of law according to legal concepts. A complementary issue, over-looked in the legal text retrieval literature, is the diversification of the search results, i.e., covering different intents of the query in the top-ranked results. Consider, for example, a lawyer preparing his/her arguments for a given case who submits a user query to retrieve information. He/she has to iteratively browse an enormous number of judgments selecting, through knowledge and experience, relative documents in order to acquire a broad and in-depth context understanding. A diverse result, i.e. a result covering a wide range of possible legal interpretations is intuitively more informative and helpful than a set of homogeneous results that contain only relevant cases with similar features.

In order to satisfy a wide range of users, query results diversification has attracted a lot of attention in the field of text mining. IR systems attempt to diversify search results, so that they cover a wide range of possible interpretations (aspects, intents or subtopics) of a query. In consequence, the number of redundant items in a search result list should decrease, while the likelihood that a user will be satisfied with any of the displayed results should increase. There has been extensive work on query results diversification, see related work Sect. 2, where the key idea is to select a small set of results that are sufficiently dissimilar, according to an appropriate similarity metric.

In this work we address result diversification in the legal IR. To this end, we adopt various methods from the literature that are introduced for text summarization (LexRank [6] and Biased LexRank [27]), graph-based ranking (GrassHopper [37] and DivRank [22]) and web search result diversification (MMR [3], Max-Sum [13], Max-Min [13] and MonoObjective [13]). While investigating the performance of these approaches, we analyze the impact of various features in computing the query-document relevance and document-document similarity scores. We evaluate the performance of the above methods on a legal corpus subjectively annotated with relevance judgments using metrics employed in TREC Diversity Tasks. To the best of our knowledge none of these methods were employed in the context of diversification in legal IR and evaluated using diversity-aware evaluation metrics.

Our findings reveal that (i) web search diversification techniques outperform other evaluated approaches (e.g. summarization-based, graph-based methods) in the context of providing diversified results in the legal domain, and (ii) the diversification criteria we introduce provide distinctively diverse subsets of resulting documents, as opposed to other approaches that are based only on textual similarity.

The remainder of this paper is organized as follows: Sect. 2 reviews previous work in query result diversification, diversified ranking on graphs and in the field of legal text retrieval. Section 3 introduces the concepts of search diversification and presents diversification algorithms, while Sect. 4 describes our experimental

framework and evaluation results. Finally, we draw our conclusions and future work aspects in Sect. 5.

2 Related Work

We first present related work on query result diversification, afterwards on diversified ranking on graphs and then on legal text retrieval techniques.

2.1 Query Result Diversification

Users of (Web) search engines typically employ keyword-based queries to express their information needs. These queries are often underspecified or ambiguous to some extent [5]. Different users who pose exactly the same query may have very different query intents. Simultaneously the documents retrieved by an IR system may reflect superfluous information. Search result diversification aims to solve this problem, by returning diverse results that can fulfill as many different information needs as possible. The published literature on search result diversification is reviewed in [28]. One of the earliest works on diversification is the maximal marginal relevance [3]. It envolves re-ranking search results as the combination of two metrics, one measuring the similarity among documents and the other the similarity between documents and the query. [13] introduced a general framework for result diversification with a set of diversification axioms and three diversification objectives, which we utilize in our work. Other researchers [33] utilized the correlation between documents as a measure of their similarity in the pursuit of diversification and risk minimization in document ranking. Diverfication heuristics that explicitly leverage external information, computed through probabilistic methods also have been proposed in [1,16,29]. In contrary to the above methods, given the fact that these methods utilize proprietary information, we do rely only on implicit knowledge of the legal corpus.

2.2 Diversified Ranking on Graphs

Many network-based ranking approaches have been proposed to rank objects according to different criteria [19] and recently diversification of the results has attracted attention. Research is currently focused on two directions: a greedy vertex selection procedure and a vertex reinforced random walk. The greedy vertex selection procedure, at each iteration, selects and removes from the graph the vertex with maximum random walk based ranking score. One of the earlier algorithms that address diversified ranking on graphs by vertex selection with absorbing random walks is Grasshopper [37]. A diversity-focused ranking methodology, based on reinforced random walks, was introduced in [22]. Their proposed model, DivRank, incorporates the rich-gets-richer mechanism to PageRank with reinforcements on transition probabilities between vertices. We utilize these approaches in our diversification framework considering the connectivity matrix of the citation network between documents that are relevant for a given user query.

2.3 Legal Text Retrieval

Legal text retrieval traditionally relies on external knowledge sources, such as thesauri and classification schemes. [25] presents various techniques used in legal text retrieval. Several supervised learning methods have been proposed to classify sources of law according to legal concepts [2,14,23]. Ontologies and thesaurus have been employed to facilitate information retrieval [12,17,30,32] or to enable the interchange of knowledge between existing legal knowledge systems [15]. Legal document summarization [7,8,24] has been used as a way to make the content of the legal documents, notably cases, more easily accessible. We also utilize state of the art summarizations algorithms but under a different objective: we aim to maximize diversity of the result set for a given query.

In another line of work citation analysis has been used in the field of law to construct case law citation networks [21][1]. Case law citation networks contain valuable information, capable of measuring legal authority [26], identifying authoritative precedent[2] [10], evaluating the relevance of court decisions [9] or even assisting summarizing legal cases [11], thus showing the effectiveness of citation analysis in the Case law domain. While the American legal system has been the one that has undergone the widest series of studies in this direction, recently various researchers applied network analysis in the Civil law domain as well. The authors of [18] propose a network-based approach to model the law. Network analysis techniques where also employed in [34] to identify context networks in dutch legislation and in [35] to recommend relevant sources of law given a focus document. In this work we also utilize citation analysis techniques and construct the Legislation Network, as to cover a wide range of possible aspects of a query.

3 Legal Document Diversification

At first, we define the problem addressed in this paper and provide an overview of the diversification process. Afterwards, legal document's features relevant for our work are introduced and distance functions are defined. Finally, we describe the diversification algorithms used in this work.

3.1 Problem Formulation

Result diversification is a trade-off between finding relevant to the user query documents and diverse documents in the result set. Given a set of legal documents and a query, our aim is to find a set of relevant and representative documents and to select these documents in such a way that the diversity of the set is maximized. More specifically, the problem is formalized as follows:

[1] Case documents usually cite previous cases, which in turn may have cited other cases and thus a network is formed over time with these citations between cases.

[2] Legal norm inherited from English common law that encourages judges to follow precedent by letting the past decision stand.

Definition 1 (Legal document diversification). *Let q be a user query and N a set of documents relevant to the user query. Find a subset $S \subseteq N$ of documents that maximize an objective function f that quantifies the diversity of documents in S.*

$$S = \underset{\substack{|S|=k \\ S \subseteq N}}{\operatorname{argmax}} f(N) \tag{1}$$

3.2 Diversfication Overview

Figure 1, illustrates the overall workflow of the diversification process. At the highest level, the user express his/her information need, the user query. Relevant, with the information need, documents are retrieved. Diversification aims to find a subset of those documents that maximize an objective function that quantifies the diversity of documents. Significant components of the process include:

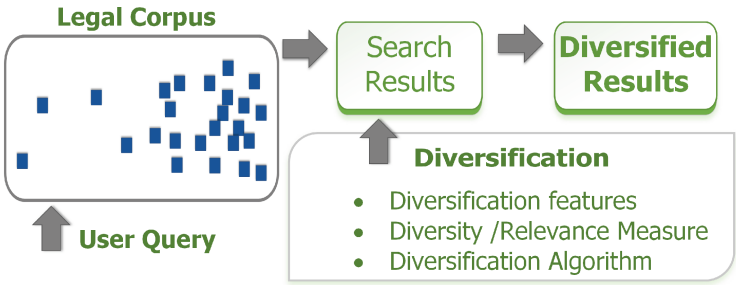

Fig. 1. Diversification overview

- *Ranking Features*, features of legal documents that will be used in the ranking process.
- *Distance Measures*, functions to measure the similarity between two legal documents and the relevance of a query to a given document.
- *Diversification Heuristics*, heuristics to produce a subset of diverse results.

3.3 Ranking Features

Under the Vector Space model, which we employ in this work, each document u can be represented as a term vector $U = (is_{w1u}, is_{w2u}, ..., is_{wmu})^T$, where $w_1, w_2, ..., w_m$ are all the available terms, and is can be any popular indexing schema e.g. $tf, tf - idf, logtf - idf$. User queries are represented in the same manner as documents.

Typically diversification techniques measure diversity in terms of content, where only textual similarity between items is used in order to quantify information similarity. In this work, we extend the notion of diversity on supplementary

features/dimensions, besides textual similarity. In order to identify these features we examine the unique characteristics of the legal documents. Documents in the legal domain possess some noteworthy characteristics, such as being intrinsically multi-topical, relying on well crafted, domain-specific language, and possessing a broad and unevenly distributed coverage of legal issues. [20].

– **Content**. Various well-known functions from the literature (e.g. Jaccard, cosine similarity etc.) can be employed at computing the textual similarity of legal documents. In this work, we choose cosine similarity as a similarity measure, thus the textual similarity between documents u and v, with term vectors U and V is:

$$S_c(u, v) = \cos(u, v) = \frac{U \cdot V}{\| U \| \| V \|} \tag{2}$$

– **Topical Taxonomies**. We consider the selection of categories that cover many different interpretations in respect to legal users' information needs. Topical similarity of two documents having topical sets X_u and X_v is calculated using the Jacard similarity

$$S_x(u, v) = \frac{|X_u \cap X_v|}{|X_u \cup X_v|} \tag{3}$$

– **Time**. Time is a valuable diversification dimension, since in many cases, subtopics associated to queries in the legal domain are temporally ambiguous due to dynamic evolution and dependencies across the legislation system. Time similarity, between documents u and v, having timestamps t_u and t_v is calculated on the difference of their normalized timestamps with Min-Max Normalization.

$$S_t(u, v) = 1 - |t_{norm}(u) - t_{norm}(v)| \tag{4}$$

– **Readability**. A document's writing quality is a diversification factor, since it expresses comprehensibility of the document itself. The most influential quantitative measure of text quality is the Flesch Reading Ease Score[3], which produces a numerical score, with higher numbers indicating easier texts. Readability similarity, between documents u and v, having readability scores r_u and r_v, is calculated on the difference of their normalized scores with Min-Max Normalization.

$$S_r(u, v) = 1 - |r_{norm}(u) - r_{norm}(v)| \tag{5}$$

Following diversification features formalization we define:

– **Document Similarity**. The final similarity score of two documents u, v is calculated as a linear weighted function of the Content, Topical Taxonomies, Time and Readability score

[3] http://en.wikipedia.org/wiki/Readability.

$$sim(u,v) = \sum_{i=1}^{|4|} w_i \; feat_i(u,v) = w_1 \; S_c(u,v) \; + \; w_2 \; S_x(u,v) \; + \; w_3 \; S_t(u,v) \; + \; w_4 \; S_r(u,v) \quad (6)$$

with weights $\sum_{i=1}^{|4|} w_i = 1$.

– **Document Distance**. The distance of two documents is

$$d(u,v) = 1 - sim(u,v) \quad (7)$$

– **Query Document Similarity**. The relevance of a query q to a given document u can be assigned as the initial ranking score obtained from the IR system, or calculated using the similarity measure e.g. cosine similarity on the corresponding term vectors

$$r(q,u) = S_c(q,u) \quad (8)$$

3.4 Diversification Heuristics

Most of existing diversification methods first retrieve a set of documents based on their relevance scores, and then re-rank the documents so that the top-ranked documents are diversified to cover more query subtopics. Since the problem of finding an optimum set of diversified documents is NP-hard, a greedy algorithm is often used to iteratively select the diversified document. Let N the document set, $u, v \in N$, $r(q,u)$ the relevance of u to the query q, $d(u,v)$ the distance of u and v, $S \subseteq N$ with $|S| = k$ the number of documents to be collected and $\lambda \in [0..1]$ a parameter used for setting trade-off between relevance and similarity. In this paper, we focus on the following representative diversification methods discussed in the previous section.

– **MMR:** Maximal Marginal Relevance [3], a greedy method to combine query relevance and information novelty, iteratively constructs the result set S by selecting documents that maximizes the following objective function

$$f_{MMR}(u,q) = (1-\lambda) \;\; r(u,q) + \lambda \sum_{v \in S} d(u,v) \quad (9)$$

MMR incrementally computes the standard relevance-ranked list when the parameter $\lambda = 0$, and computes a maximal diversity ranking among the documents in N when $\lambda = 1$. For intermediate values of $\lambda \in [0..1]$, a linear combination of both criteria is optimized. The set S is usually initialized with the document that has the highest relevance to the query. Since the selection of the first element has a high impact on the quality of the result, MMR often fails to achieve optimum results.

– **MaxSum:** The Max-sum diversification objective function [13] aims at maximizing the sum of the relevance and diversity in the final result set. This is achieved by a greedy approximation algorithm that selects a pair of documents that maximizes Eq. 10 in each iteration.

$$f_{MAXSUM}(u,v,q) = (1-\lambda) \; (r(u,q) + r(v,q)) + 2\lambda \; d(u,v) \quad (10)$$

where (u, v) is a pair of documents, since this objective considers document pairs for insertion. When $|S|$ is odd, in the final phase of the algorithm an arbitrary element in N is chosen to be inserted in the result set S.

- **MaxMin:** The Max-Min diversification objective function [13] aims at maximizing the minimum relevance and dissimilarity of the selected set. This is achieved by a greedy approximation algorithm that select a document that maximizes Eq. 11 in each iteration.

$$f_{MAXMIN}(u, q) = (1 - \lambda)\ r(u, q) + \lambda \min_{v \in S} d(u, v) \tag{11}$$

where $\min_{v \in S} d(u, v)$ is the minimum distance of u to the already selected documents in S.

- **MonoObjective:** MonoObjective [13] combines the relevance and the similarity values into a single value for each document. It is defined as:

$$f_{MONO}(u, q) = r(u, q) + \frac{\lambda}{|N| - 1} \sum_{v\, \epsilon\, N} d(u, v) \tag{12}$$

- **LexRank:** LexRank [6], is a stochastic graph-based method for computing relative importance of textual units. A document is represented as a network of inter-related sentences, and a connectivity matrix based on intra-sentence similarity is used as the adjacency matrix of the graph representation of sentences. In LexRank scoring formula 13, Matrix B captures pairwise similarities of the sentences and square matrix A, which represents the probability of jumping to a random node in the graph, has all elements set to $1 = M$, where M is the number of sentences.

$$p = [\lambda\ A + (1 - \lambda)\ B]^T p \tag{13}$$

In our setting, instead of sentences, we use documents that are in the initial retrieval set N for a given query and thus set Matrix B as the connectivity matrix based on document similarity.

- **Biased LexRank:** Biased LexRank [27] provides for a LexRank extension that takes into account a prior document probability distribution e.g. the relevance of documents to a given query.

$$p = [\lambda\ A + (1 - \lambda)\ B]^T p \tag{14}$$

In Biased LexRank scoring formula 14, we set Matrix B as the connectivity matrix based on document similarity for all documents that are in the initial retrieval set N for a given query and Matrix A elements proportional to the query document relevance.

- **DivRank:** DivRank balances popularity and diversity in ranking, based on a time-variant random walk. In contrast to PageRank which is based on stationary probabilities, DivRank assumes that transition probabilities change over time, they are reinforced by the number of previous visits to the target vertex. If $p_T(u, v)$ is the transition probability from any vertex u to vertex v

at time T, $p^*(d_j)$ is the prior distribution that determines the preference of visiting vertex d_j, and $p_0(u, v)$ is the transition probability from u to v prior to any reinforcement then,

$$p_T(d_i, d_j) = (1 - \lambda).p^*(d_j) + \lambda.\frac{p_0(d_i, d_j).N_T(d_j)}{D_T(d_i)} \tag{15}$$

where $N_T(d_j)$ is the number of times the walk has visited d_j up to time T and,

$$D_T(d_i) = \sum_{d_j \in V} p_0(d_i, d_j) N_T(d_j) \tag{16}$$

Since DivRank is a query independent ranking model, we introduce a query dependent prior and thus utilize DivRank into a query dependent ranking schema. In our setting, we use documents that are in the initial retrieval set N for a given query q, create the citation network between those documents and apply DivRank algorithm to select top-k divers documents in S.

– **Grasshopper:** A similar with DivRank ranking algorithm, is described in [37]. This model starts with a regular time-homogeneous random walk and in each step the vertex with the highest weight is set as an absorbing state.

$$p_T(d_i, d_j) = (1 - \lambda).p^*(d_j) + \lambda.\frac{p_0(d_i, d_j).N_T(d_j)}{D_T(d_i)} \tag{17}$$

where $N_T(d_j)$ is the number of times the walk has visited d_j up to time T and,

Since Grasshopper and DivRank utilize a similar approach and will ultimately present similar results we utilized Grasshopper distinctively from DivRank. In particularly, instead of creating the citation network of documents belonging to the initial result set, we form the adjacency matrix based on document similarity.

4 Experimental Setup

In this section, we describe the legal corpus we use, the set of query topics, the respective methodology for subjectively annotating our corpus with relevance judgments for each query, as well as the metrics employed for the evaluation assessment. Finally, we provide our diversification results along with a short discussion.

4.1 Legal Corpus

Our corpus contains 63,742 precedential legal cases from the Supreme Court of the United States[4]. The cases were originally downloaded from CourtListener[5].

[4] http://www.supremecourt.gov/.

[5] http://www.courtlistener.com, a free legal research website containing legal opinions from federal and state courts.

The legal corpus contains all cases from the Supreme Court of the United States, covering more than two centuries of legal history, spanning from 1754 up to 2015. We extracted from the cases text all the necessary information for our feature selection framework e.g. relationships to other documents, date of Judgment. Since our corpus was initially unclassified, we acquired topical taxonomies from the Supreme Court Database[6] using commonly shared unique identification variable SCDB Case ID. Topical taxonomies within Supreme Court Database are the outcome of a manual analysis and interpretation of the legal provisions considered in each case. Our text pre-processing step involved standard stop word removal and porter stemming. Finally our index, build with log based $tf - idf$ indexing technique contains a total of 63,742 documents, 174,370 unique terms and 54,243,977 terms in total. Overall we believe that the corpus is of size to demonstrate the effectiveness of our proposed approach.

4.2 Evaluation Metrics

We evaluate diversification methods using metrics employed in TREC Diversity Tasks[7]. In particular we report

- **a-nDCG:** a-Normalized Discounted Cumulative Gain [4] metric quantifies the amount of unique aspects of the query q that are covered by the $top - k$ ranked documents. We use $a = 0.5$, as typical in TREC evaluation.
- **Precision-IA:**. Precision-Intent Aware [1] accounts for the ratio of relevant documents for different subtopics within the $top - k$ items.
- **Subtopic-Recall:** Subtopic-Recall [36] quantifies the amount of unique aspects of the query q that are covered by the $top - k$ ranked documents

4.3 Relevance Judgements

One of the difficulties in evaluating methods designed to introduce diversity in the legal document ranking process is the lack of standard testing data. Evaluating diversification requires a data corpus, a set of query topics and a set of relevance judgments, preferably made by human assessors for each query. While TREC added a diversity task to the Web track in 2009, this dataset was designed assuming a general web search, and so it not possible to adapt it to our setting. In the absence of a standard dataset specifically tailored for this purpose and since it was not feasible to involve legal experts in this sort of exploratory study, we looked for an subjective way to evaluate and assess the performances of various diversification methods on our corpus. We do acknowledge the fact that the process of automatic query generation is at best an imperfect approximation of what a real person would do. To this end we employed the following method:

User Profiles/Queries. We used West Law Digest Topics[8] as candidates user queries. Each topic was issued as candidate query to our retrieval system.

[6] http://scdb.wustl.edu.

[7] http://trec.nist.gov/data/web10.html.

[8] A taxonomy of identifying points of law from reported cases and organizing them by topic and key number. It is used to organize the entire body of American law.

Table 1. West Law Digest Topics as user queries

31:	Antitrust and Trade Regulation	61:	Breach of Marriage Promise
84:	Commodity Futures Trading Regulation	199:	Implied and Constructive Contracts
376:	Unemployment Compensation	398:	Merit Systems Protection

Outlier queries, whether too specific/rare or too general, where removed using the interquartile range, below or above values $Q1$ and $Q3$, sequentially in terms of number of hits in the result set and score distribution for the hits, demanding in parallel a minimum cover of $min|N|$ results. In total, we kept 330 queries The following Table 1 provides a sample of the topics we further consider as user queries.

Query assessments and ground-truth. For each topic/query we kept the $top - n$ results. An LDA topic model, using an open source implementation[9], was trained on the $top - n$ results for each query. From the resulting topic distributions for each document, with an acceptance threshold of 15 %, we consider relevance judgments for each query/ document and subtopic. In other words, we consider the topics created from LDA as aspects of each query, and based on the topic/ document distribution we can infer whether a document is relevant for an aspect. In total, we acquired 1,650 subtopics for all the 330 queries. We have made available[10] our complete dataset, ground-truth data, queries and relevance assessments in standard qrel format, as to encourage progress on the diversification in legal IR.

4.4 Results

As a baseline to compare diversification methods, we consider the simple ranking produced from an IR system using cosine similarity and log based $tf - idf$ indexing schema. For each query, our initial set N contains the $top - n$ query results. For all variations that apply diversity, we set a fixed weight for the diversity score to $\lambda = 0.5$ and, thus, the weight for query-to-document similarity is $1 - \lambda = 0.5$. We present the evaluation results for the methods employed, using the aforementioned evaluation metrics, at cut-off values of 5, 10 and 20, as typical in TREC evaluations. Note that each of the diversification variations, is applied in combination with each of the diversification algorithms and for each user query. Table 2 summarizes testing parameters and their corresponding ranges.

We firstly employed the diversification methods using only content similarity as used in most works handling diversification, e.g. in web search results diversification. That is, weights on features time, readability and topical categories were set to zero. Table 3 presents results of the diversification methods.

[9] http://mallet.cs.umass.edu/.
[10] https://github.com/mkoniari/MultiLegalDiv.

Table 2. Parameters tested in the experiments

Parameter	Range
Algorithms tested	MMR, MaxMin, MaxSum, Mono, LexRank, BiasedLexRank, DivRank, GrassHopper
Tradeoff λ values	0.5
Candidate set size n = $\|N\|$	100
Result set size k = $\|S\|$	5, 10, 20
# of sample queries	330
Exp. 1 Feature weights	Content 1.0, Time, 0 Readability 0, Topical Taxonomies, 0
Exp. 2 Feature weights	Content 0.6, Time, 0.13, Readability 0.13, Topical Taxonomies, 0.14

Statistically significant values, using the paired two-sided t-test with $p_{value} < 0.05$ are denoted with $^{\circ}$ and with $p_{value} < 0.01$ with *.

MMR and DivRank are the best diversification strategies for different evaluation metrics for $N = 100$ and $k = 30$. In particular, MMR outperforms all other methods in terms of the nDCG and Subtopic-Recall metrics, whereas DivRank achieves the highest score for the Precision IA metric. Interestingly, text summarization methods (LexRank, Biased LexRank and GrassHopper, as it was utilized without a network citation graph) failed to improve the baseline ranking. They actually constantly perform lower than the baseline ranking at all levels across all metrics. From web search result diversification methods, MMR almost constantly achieves better results in respect to the rest methods for all metrics, with the exception of nDCG@5 where MaxMin performs better. Graph diversification method, DivRank, outerperforms other methods in Precision IA metric at all levels, but generally fails to improve over the baseline ranking for nDCG and Subtopic-Recall metrics.

As a second experiment, we incorporate all ranking features into the diversification methods while computing the similarity scores for the documents pairs, except DivRank where the citation network between documents in the result set for each query is utilized. In particular we set the following weights on ranking features: Content 0.6, Time 0.13, Readability 0.13 and Topical Taxonomies 0.14. In Table 4 we present results of the second experiment, alongside with indicators for statistically significant values.

It is clear that with the incorporation of the suggested ranking features all of the approaches tend to perform better than using only content similarity. We also notice a similar trending behavior with the one discussed for Table 3. MMR and DivRank are the best diversification strategies for different evaluation metrics. Text summarization methods, although with better scores, once again fail to improve over the baseline ranking. MMR almost constantly achieves better results in respect to the rest methods for all metrics, with the exception of Precision IA where MaxMin and DivRank perform better.

Table 3. Retrieval Performance of the diversification algorithms using only content similarity for $N = 100$ and $k = 30$. Highest scores are shown in bold. Statistically significant values, using the paired two-sided t-test with $p_{value} < 0.05$ are denoted with $^{\circ}$ and with $p_{value} < 0.01$ with *

	a-nDCG			Precision IA			ST recall		
Method	@5	@10	@20	@5	@10	@20	@5	@10	@20
IR	0,532	0,595	0,656	0,314	0,313	0,314	0,688	0,833	0,948
MMR	**0,571***	**0,643***	**0,695***	0,315°	0,321	0,322*	**0,783***	**0,923***	**0,977***
MaxSum	0,549	0,620*	0,675*	0,300*	0,305°	0,303*	0,744*	0,880*	0,969°
MaxMin	0,568*	0,633*	0,686*	0,319	0,319*	0,319*	0,777*	0,907*	0,976*
MonoObjective	0,541°	0,602°	0,664*	0,313	0,310	0,312	0,713*	0,844	0,960°
LexRank	0,487	0,532*	0,586*	0,308*	0,313	0,320	0,584*	0,705*	0,820*
BiasedLexRank	0,488*	0,533*	0,587*	0,309	0,314	0,320	0,585*	0,708*	0,821*
DivRank	0,533	0,589	0,635	**0,320**	**0,326**$^{\circ}$	**0,326**$^{\circ}$	0,667	0,803	0,888*
GrassHopper	0,492*	0,542*	0,598*	0,310	0,316	0,322°	0,592*	0,725*	0,846*

Table 4. Retrieval Performance of the diversification algorithms using all ranking features for $N = 100$ and $k = 30$. Highest scores are shown in bold. Statistically significant values, using the paired two-sided t-test with $p_{value} < 0.05$ are denoted with $^{\circ}$ and with $p_{value} < 0.01$ with *

	a-nDCG			Precision IA			ST recall		
Method	@5	@10	@20	@5	@10	@20	@5	@10	@20
IR	0,532	0,595	0,656	0,314	0,313	0,314	0,688	0,833	0,948
MMR	**0,586***	**0,657***	**0,709***	0,321	0,321°	0,325*	**0,815***	**0,939***	**0,989***
MaxSum	0,564*	0,636*	0,689*	0,306	0,306°	0,308°	0,779*	0,913*	0,977*
MaxMin	0,581*	0,650*	0,702*	**0,322**	0,322	0,321*	0,793*	0,931*	0,983*
MonoObjective	0,550*	0,612*	0,673*	0,321°	0,313	0,314	0,716*	0,857°	0,968*
LexRank	0,484*	0,532	0,587*	0,304*	0,306	0,316	0,604*	0,724*	0,839*
BiasedLexRank	0,488*	0,537*	0,592*	0,304	0,308	0,316	0,607*	0,731*	0,845*
DivRank	0,533	0,589	0,635	0,320	**0,326***	**0,326**$^{\circ}$	0,667	0,803	0,888*
GrassHopper	0,504°	0,555*	0,612*	0,306	0,308	0,317	0,649	0,760*	0,880*

Overall it is demonstrated that more refined criteria than plain content similarity can improve the effectiveness of the diversification process. Furthermore web search diversification techniques outperform other approaches (e.g. summarization-based, graph-based methods) in the context of legal search diversification. Graph based diversification, DivRank generally fails to improve over the baseline ranking but outerperforms other methods in terms of Precision IA metric. We do plan to further examine the performance of graph based diversification heuristics, in terms of citation network criteria and ranking features, as to enrich search results with otherwise hidden aspects of the legal query space.

5 Conclusions

In this paper, we studied the novel problem of diversifying legal documents by incorporating diversity in four dimensions: content, time, topical taxonomies and readability. We adopted and compared the performance of several state of the art methods from the web search, network analysis and text summarization domains as to handle the problems' challenges. We evaluated all the methods/ dimensions using a real data set from the Common Law domain that we subjectively annotated with relevance judgments for this purpose. Our findings demonstrate the effectiveness of our proposed method, as opposed to applying plain content diversity on legal search results.

A challenge we faced in this work was the lack of ground-truth. We hope on an increase of the size of truth-labeled data set in the future, which would enable us to draw further conclusions about the diversification techniques. In the future we plan to perform an exhaustive evaluation of all the methods as to provide insights for legal IR systems between reinforcing relevant documents, result set similarity, or sampling the information space around the legal query, result set diversity.

References

1. Agrawal, R., Gollapudi, S., Halverson, A., Ieong, S.: Diversifying search results. In: Proceedings of WSDM 2009, pp. 5–14 (2009)
2. Biagioli, C., Francesconi, E., Passerini, A., Montemagni, S., Soria, C.: Automatic semantics extraction in law documents. In: Proceedings of ICAIL 2005 (2005)
3. Carbonell, J., Goldstein, J.: The use of mmr, diversity-based reranking for reordering documents and producing summaries. In: Proceedings of SIGIR 1998, pp. 335–336 (1998)
4. Clarke, C.L.A., Kolla, M., Cormack, G.V., Vechtomova, O., Ashkan, A., Büttcher, S., MacKinnon, I.: Novelty and diversity in information retrieval evaluation. In: Proceedings of SIGIR 2008 (2008)
5. Cronen-Townsend, S., Croft, W.B.: Quantifying query ambiguity. In: Proceedings of Human Language Technology Research 2002 (2002)
6. Erkan, G., Radev, D.R.: LexRank: graph-based lexical centrality as salience in text summarization. J. Artif. Int. Res. **22**(1), 457–479 (2004)
7. Farzindar, A., Lapalme, G.: Legal text summarization by exploration of the thematic structures and argumentative roles. In: Text Summarization Branches Out Workshop Held in Conjunction with ACL, pp. 27–34 (2004)
8. Farzindar, A., Lapalme, G.: Letsum, an automatic legal text summarizing system. In: Proceedings of JURIX 2004, pp. 11–18 (2004)
9. Fowler, J.H., Johnson, T.R., Spriggs, J.F., Jeon, S., Wahlbeck, P.J.: Network analysis and the law: measuring the legal importance of precedents at the U.S. Supreme Court. Polit. Anal. **15**(3), 324–346 (2006)
10. Fowler, J.H., Jeon, S.: The authority of Supreme Court precedent. Soc. Netw. **30**(1), 16–30 (2008)
11. Galgani, F., Compton, P., Hoffmann, A.: Citation based summarisation of legal texts. In: Anthony, P., Ishizuka, M., Lukose, D. (eds.) PRICAI 2012. LNCS (LNAI), vol. 7458, pp. 40–52. Springer, Heidelberg (2012). doi:10.1007/978-3-642-32695-0_6

12. Gangemi, A., Sagri, M.T., Tiscornia, D.: Metadata for content description in legal information. In: Proceedings of LegOnt Workshop on Legal Ontologies (2003)
13. Gollapudi, S., Sharma, A.: An axiomatic approach for result diversification. In: Proceedings of WWW 2009, pp. 381–390 (2009)
14. Grabmair, M., Ashley, K.D., Chen, R., Sureshkumar, P., Wang, C., Nyberg, E., Walker, V.R.: Introducing LUIMA. In: Proceedings of ICAIL 2015 (2015)
15. Hoekstra, R., Breuker, J., di Bello, M., Boer, A.: The lkif core ontology of basic legal concepts. In: Proceedings of the Workshop on Legal Ontologies and Artificial Intelligence Techniques (LOAIT 2007) (2007)
16. Hu, S., Dou, Z., Wang, X., Sakai, T., Wen, J.R.: Search result diversification based on hierarchical intents. In: Proceedings of CIKM 2015, pp. 63–72 (2015)
17. Klein, M.C., Van Steenbergen, W., Uijttenbroek, E.M., Lodder, A.R., van Harmelen, F.: Thesaurus-based retrieval of case law. In: Proceedings of JURIX 2006, vol. 152, p. 61 (2006)
18. Koniaris, M., Anagnostopoulos, I., Vassiliou, Y.: Network analysis in the legal domain: a complex model for european union legal sources. In: Physics and Society, Cornell University Library, arXiv (2015). http://arxiv.org/abs/1501.05237
19. Langville, A.N., Meyer, C.D.: A survey of eigenvector methods for web information retrieval. SIAM Rev. **47**(1), 135–161 (2005)
20. Lu, Q., Conrad, J.G., Al-Kofahi, K., Keenan, W.: Legal document clustering with built-in topic segmentation. In: Proceedings of CIKM 2011, p. 383 (2011)
21. Marx, S.M.: Citation networks in the law. Jurimetrics J. **10**(4), 121–137 (1970)
22. Mei, Q., Guo, J., Radev, D.: Divrank: the interplay of prestige and diversity in information networks. In: Proceedings of KDD 2010, pp. 1009–1018 (2010)
23. Loza Mencía, E., Fürnkranz, J.: Efficient pairwise multilabel classification for large-scale problems in the legal domain. In: Daelemans, W., Goethals, B., Morik, K. (eds.) ECML PKDD 2008. LNCS, vol. 5212, pp. 50–65. Springer, Heidelberg (2008). doi:10.1007/978-3-540-87481-2_4
24. Moens, M.F.: Summarizing court decisions. Inf. Process. Manage. **43**(6), 1748–1764 (2007)
25. Moens, M.: Innovative techniques for legal text retrieval. Artif. Intell. Law **9**(1), 29–57 (2001)
26. van Opijnen, M.: Citation analysis and beyond: in search of indicators measuring case law importance. In: Proceedings of JURIX 2012, pp. 95–104 (2012)
27. Otterbacher, J., Erkan, G., Radev, D.R.: Biased LexRank: passage retrieval using random walks with question-based priors. Inf. Process. Manage. **45**(1), 42–54 (2009)
28. Santos, R.L.T., Macdonald, C., Ounis, I.: Search result diversification. Found. Trends Inf. Retrieval **9**(1), 1–90 (2015)
29. Santos, R.L., Macdonald, C., Ounis, I.: Exploiting query reformulations for web search result diversification. In: Proceedings of WWW 2010, pp. 881–890 (2010)
30. Saravanan, M., Ravindran, B., Raman, S.: Improving legal information retrieval using an ontological framework. Artif. Intell. Law **17**(2), 101–124 (2009)
31. Schweighofer, E.: Semantic indexing of legal documents. In: Francesconi, E., Montemagni, S., Peters, W., Tiscornia, D. (eds.) Semantic Processing of Legal Texts. LNCS, vol. 6036, pp. 157–169. Springer, Heidelberg (2010). doi:10.1007/978-3-642-12837-0_9
32. Schweighofer, E., Liebwald, D.: Advanced lexical ontologies and hybrid knowledge based systems: first steps to a dynamic legal electronic commentary. Artif. Intell. Law **15**(2), 103–115 (2007)

33. Wang, J., Zhu, J.: Portfolio theory of information retrieval. In: Proceedings of SIGIR 2009 (2009)
34. Winkels, R., Boer, A., Plantevin, I.: Creating context networks in dutch legislation. In: Proceedings of JURIX 2013, vol. 259, p. 155 (2013)
35. Winkels, R., Boer, A., Vredebregt, B., van Someren, A.: Towards a legal recommender system. In: Proceedings of JURIX 2014, pp. 169–178 (2014)
36. Zhai, C.X., Cohen, W.W., Lafferty, J.: Beyond independent relevance. In: Proceedings of SIGIR 2003 (2003)
37. Zhu, X., Goldberg, A.B., Van Gael, J., Andrzejewski, D.: Improving diversity in ranking using absorbing random walks. In: HLT-NAACL, pp. 97–104 (2007)