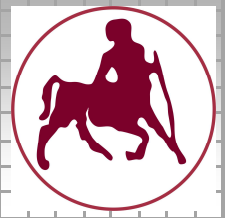
National Technical University of Athens



University of Thessaly

#mhdw2016

Diversifying the Legal Order

Marios Koniaris, Ioannis Anagnostopoulos, Yannis Vassiliou

5th Mining Humanistic Data Workshop
Thessaloniki, Greece 16-18 September 2016

Overview

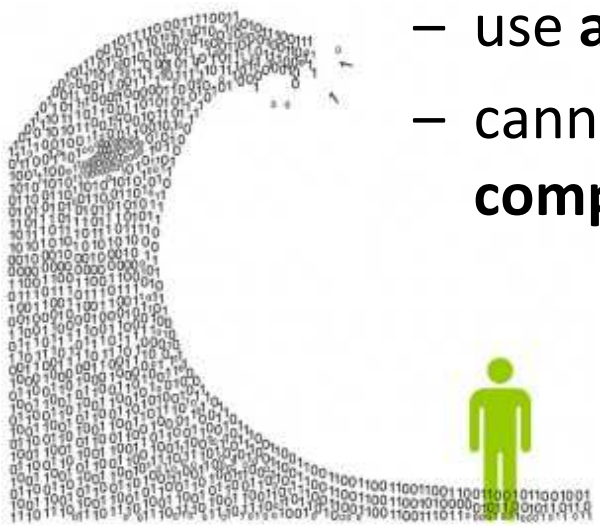
- The Problem
- Existing Approaches
- Proposed Model
- Evaluation
- Summary – Future Work

Diversifying the Legal Order

motivation & definition

Diversifying – Problem

- The number of freely available legal data sets is increasing at high speed
- Legal information overload
- Legal stakeholders:
  - find legislation **hard to read** /have **limited time**
  - use **ambiguous** and **short queries**
  - cannot find **information** that is both **relevant** & **comprehensive** for their **needs**

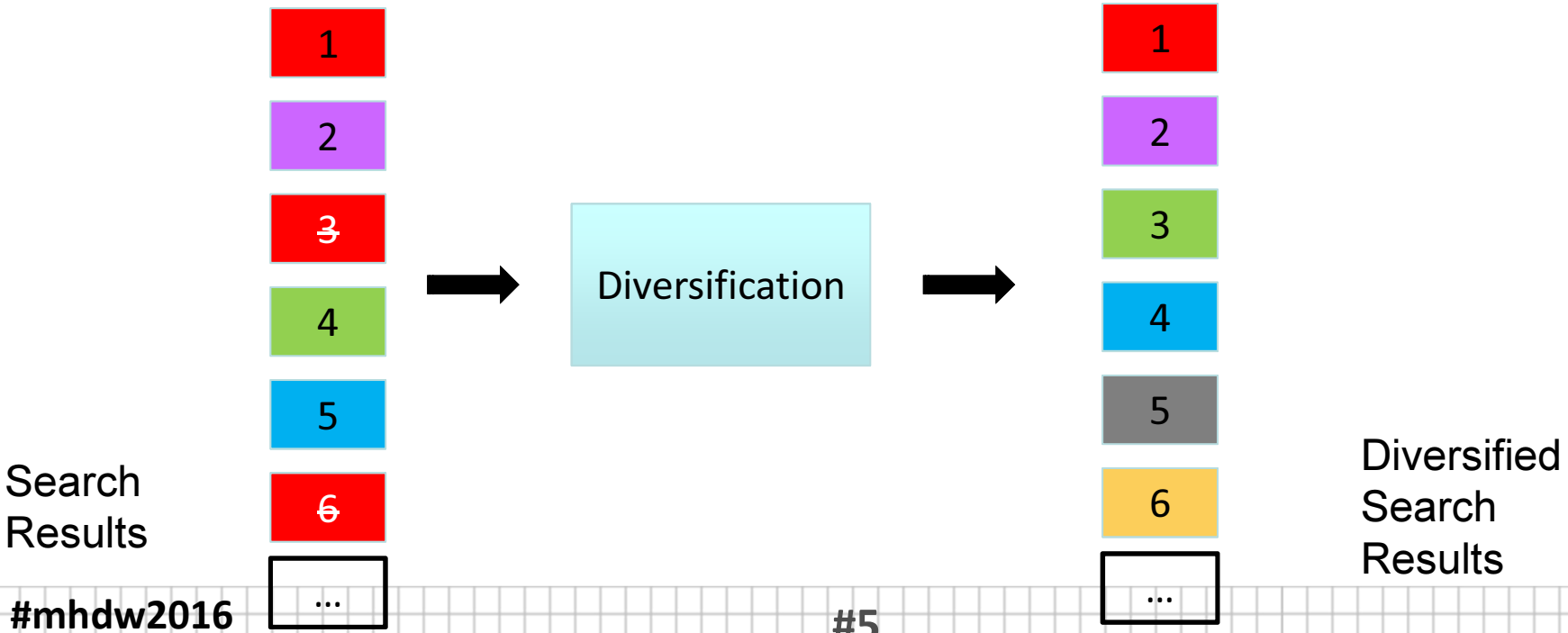


filter out redundant data while maximizing diversity among different aspects of a topic

Diversifying – Motivation

- ⊘ Outliers/ Duplicate Information
- ⊘ Users not finding a relevant result in top positions

- ✔ Cover all interpretations of the query in first results
- ✔ Maximize probability of showing an interpretation relevant to the user information need



Diversifying – Motivation

Public legal information from all countries and international institutions is part of the common heritage of humanity. **Maximizing** access to this information promotes justice and the rule of law

Declaration on Free Access to Law by Legal information institutes of the world

Diversifying – Applications

Goal: is to define /evaluate the potential of results diversification in the legal information retrieval.

- Affects
  - Simple users, law issuers, other legal stakeholders
- Improves
  - the effectiveness of legal IR systems
  - quality of search results

Related Work

Search Results Diversification

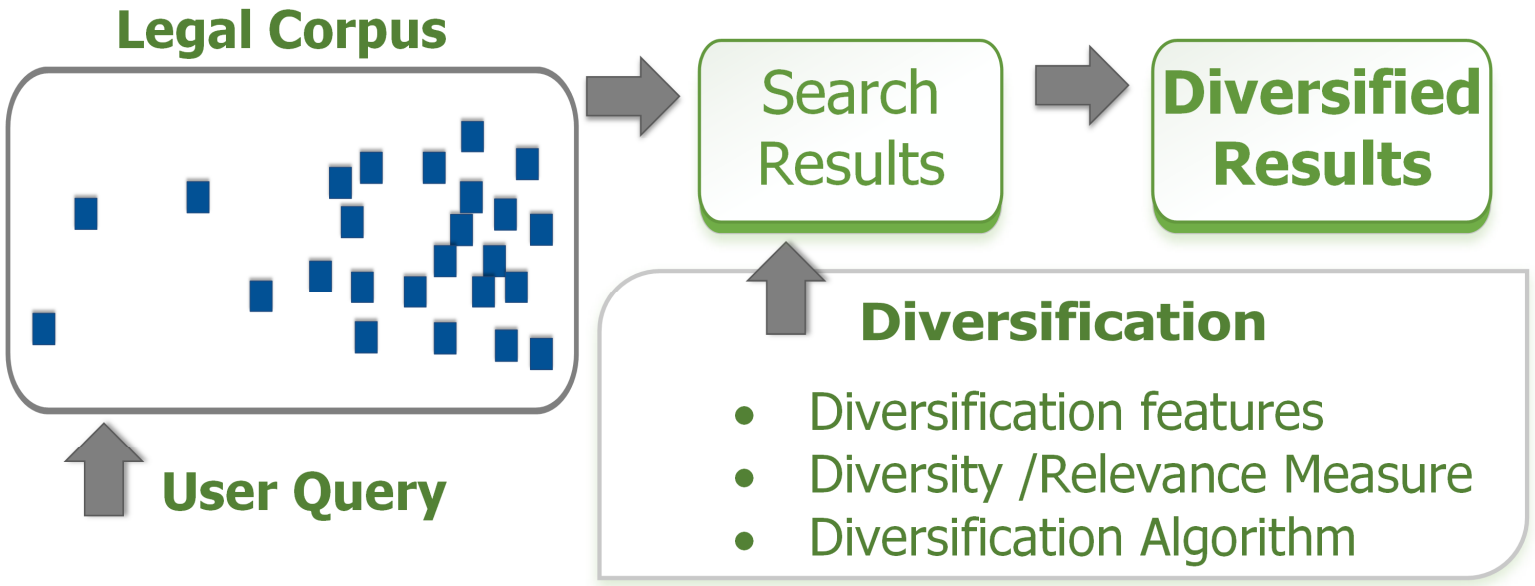
- Diversity + other ranking criterion
  - *relevance* to the user's query
- Maximal marginal relevance (MMR) (Carbonell 98)
  - linear combination of relevance and diversity
- Max-sum / Max-Min/ Mono-objective diversification objectives (Gollapudi 09)
- Explicit knowledge
  - Diversifying Search Results (Agrawal 09)
    - explicit use of taxonomy
  - Probabilistic framework xQuAD (Santos 10)

Legal Text Retrieval

- External knowledge sources
  - thesauri, ontologies, classification schemes
- Supervised learning methods
  - classify sources of law legal according to legal concepts
- Legal document summarization techniques
- Bill outlier detection

Approach Description

Method Outline



Problem Formulation

Let q be a user query and N a set of documents relevant to the user query. Find a subset $S \subseteq N$, , with $|S| = k$ *of documents that maximize an objective function f that quantifies the diversity of documents in S.*

Diversification Process

1. Items are ranked by relevance

2. Diversity is measured

3. Combine measures for final ranking

Diversity Measures

- Vector Space model
  - document u represented as a term vector V
  - Query q (the same)
  - indexing schema e.g. tf; tf-idf; logtf-idf.
- Document Similarity/ Distance
  - Jaccard, cosine similarity, …
- Query Document Similarity
  - IR system ranking score, similarity measure

Diversification Heuristics

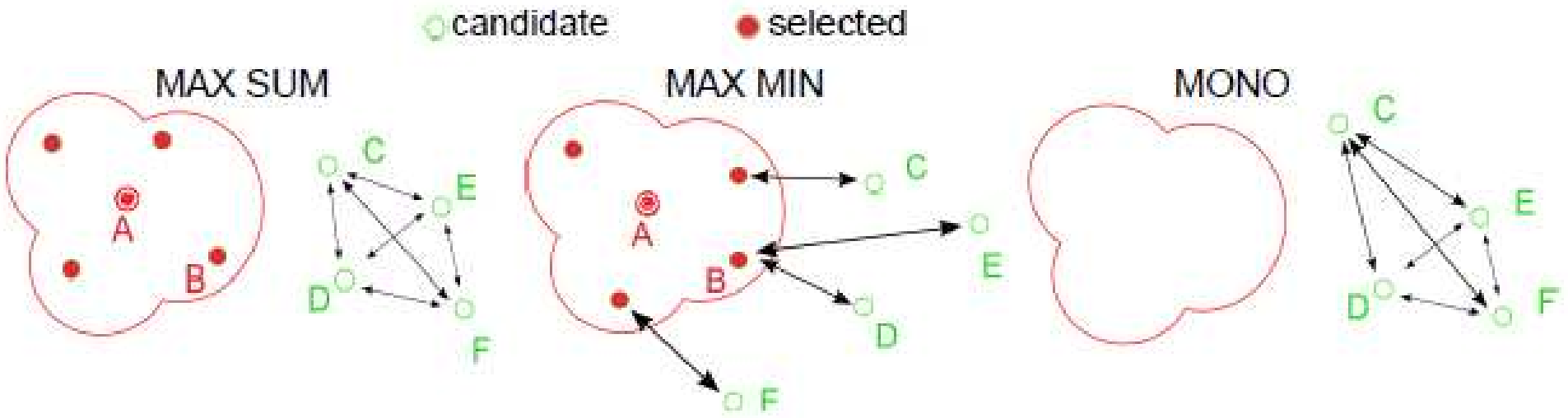
- NP-hard problem
  - a greedy algorithm is often used
- MMR

Interpolation parameter λ

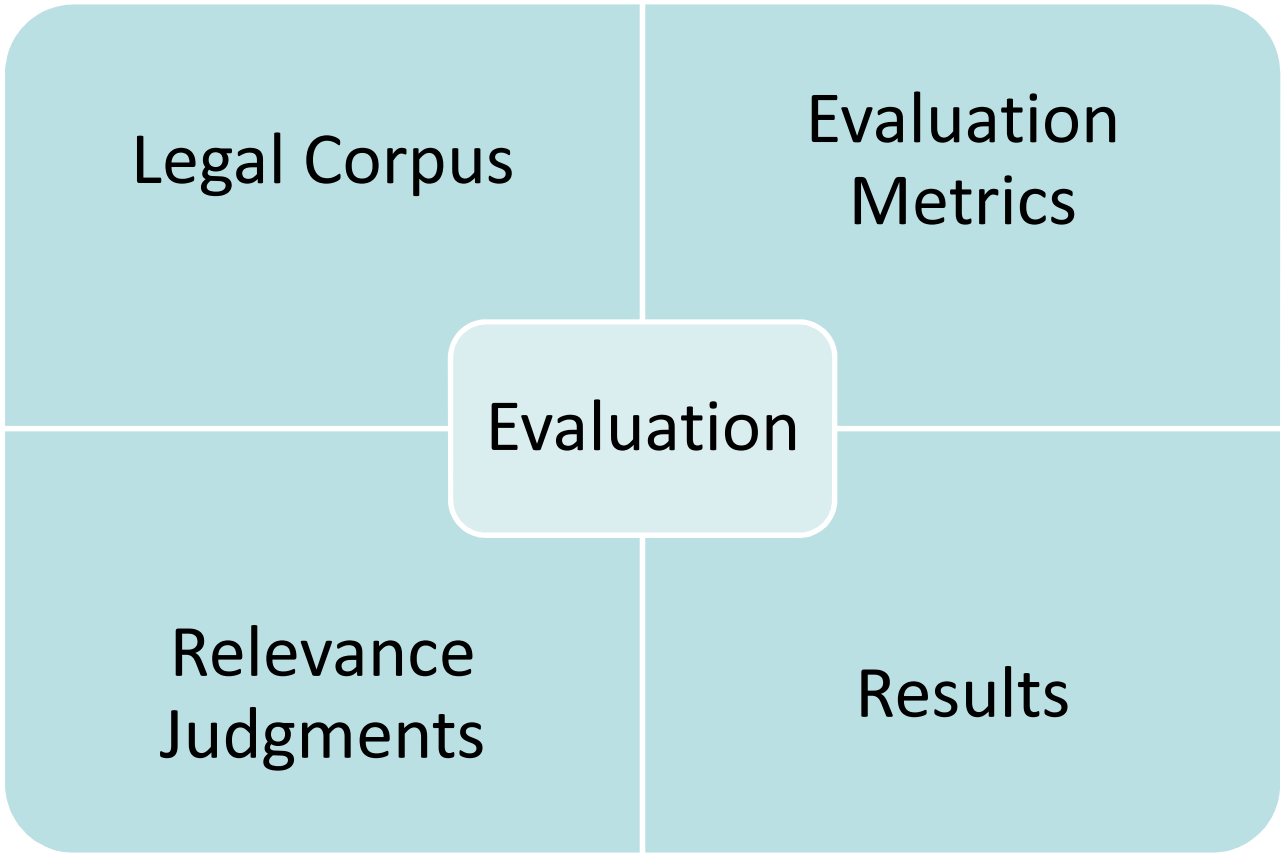
relevance

diversity

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



Evaluation


Legal Corpus
Evaluation Metrics
Evaluation
Relevance Judgments
Results

Evaluation / standard datasets ?

- need of **task-specific** standard datasets
  - data corpus
  - set of query topics
  - set of relevance judgments, preferably by human assessors for each query
  - **set of Metrics**

Legal Corpus

- 3.890 Australian legal cases / Federal Court of Australia
- Index: 9.782.911 terms & 53.791 unique terms
- Testing parameters

Parameter	Range
Tradeoff values	0.1, 0.2,0.3 … 0.8,0.9
Candidate set size	100
Result set size	5,10,20
# of queries	298

- two-fold strategy
  - qualitative analysis diversification and precision of each employed method
  - scalability analysis of diversification methods when increasing the query parameters

Relevance Judgments

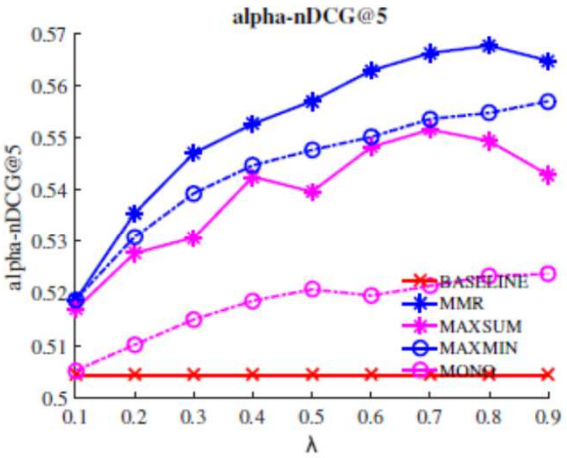
- Query Topics
  - West Law Digest Topics
    - taxonomy / organized by topic and key number
  - Candidate query to retrieval system
  - Exclude interquartile range (Q1 and Q3)
- Query assessments and ground-truth
  - train LDA topic model on top-n results for each query
  - resulting topic distribution -> *infer whether a document is relevant for an aspect*

Execution Notes

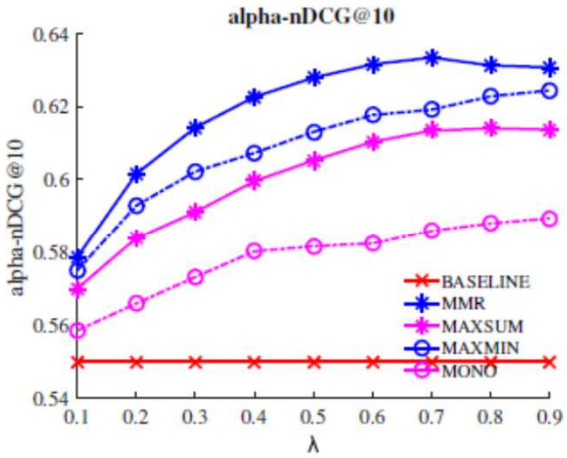
- Baseline
  - cosine similarity and log based tf-idf indexing schema
- Interpolation parameter $\lambda \in [0..1]$ tuned in 0.1 steps
  - separately for each method
- *"there is no evaluation metric that seems to be universally accepted as the best for measuring the performance of algorithms that aim to obtain diverse rankings."* Radlinski *–SIGIR 2009*

a-Normalized Discounted Cumulative Gain

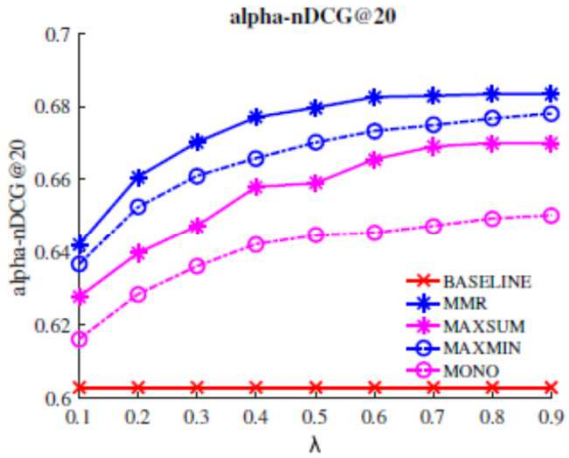
- A document is relevant when it contains a nugget/ aspect needed by the user



(a) alpha-nDCG@5



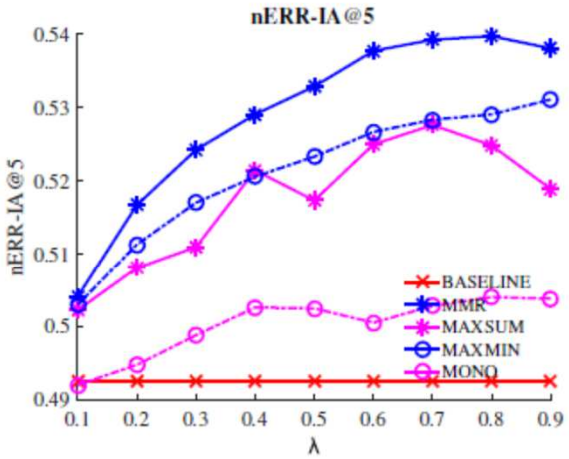
(b) alpha-nDCG@10



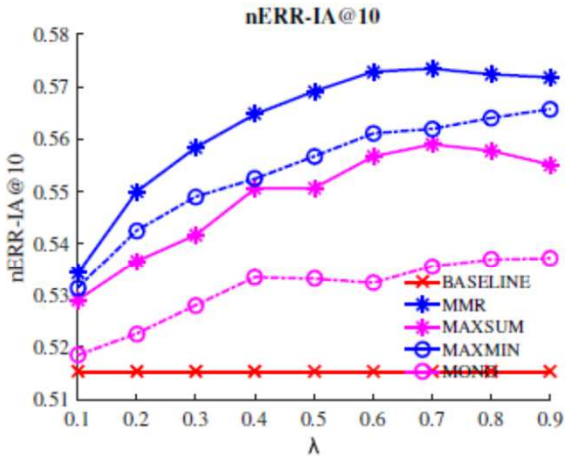
(c) alpha-nDCG@20

Expected Reciprocal Rank - Intent Aware

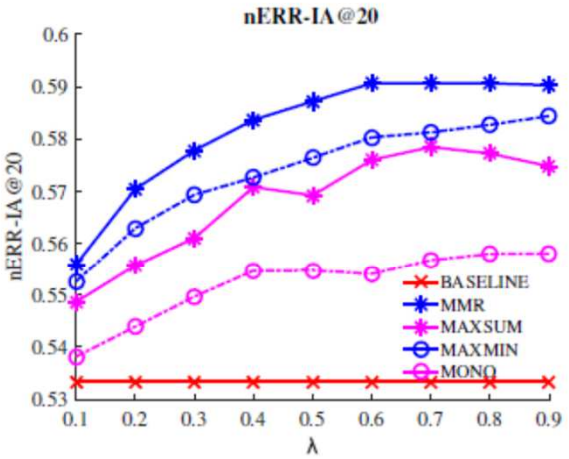
- relevance of documents ranked above



(a) nERR-IA@5



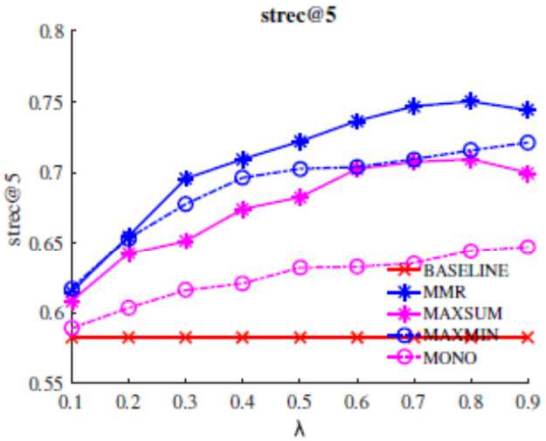
(b) nERR-IA@10



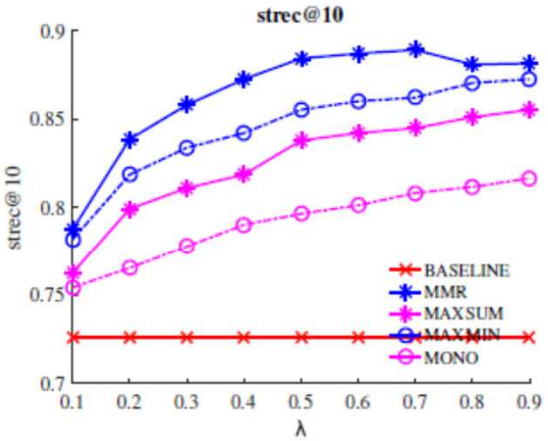
(c) nERR-IA@20

Subtopic recall

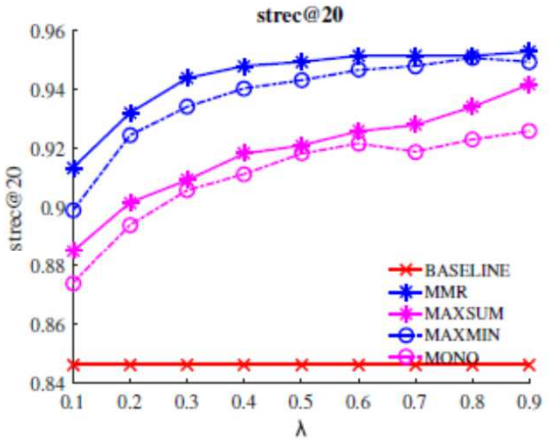
- Is the result set exhaustive?
- s–recall at k = $\frac{number\ of\ subtopics\ covered\ by\ the\ first\ k\ documents}{total\ number\ of\ subtopics}$



(a) S-Recall@5



(b) S-Recall@10



(c) S-Recall@20

Conclusions & Future Work

Conclusions

- studied the novel problem of diversifying legal search results
- adopted & compared the performance of several state of the art methods from web search domain
- performed an exhaustive evaluation of all the methods
  - using a real data set
  - subjectively annotated with relevance judgments
- diversification methods offer
  - notable improvements and enrich search results around the legal query space
  - balance boundaries between reinforcing relevant documents or sampling the information space around the legal query.

Future Work

- incorporate additional features in our legal search result diversification framework
  - features of legal documents that will be used in the ranking/ diversification process.
- investigate the performance of heuristics provided for other domains
  - e.g. for text summarization and graph diversification.

Ευχαριστώ !

Questions?

Key References

- Carbonell, J., Goldstein, J.: The use of mmr, diversity-based reranking for reordering documents and producing summaries SIGIR '98. pp. 335 -336
- Gollapudi, S., Sharma, A.: An axiomatic approach for result diversification. WWW '09 pp. 381-390
- Drosou, M., Pitoura, E.: Search result diversification. ACM SIGMOD Record 39(1), 41 (2010)
- Chapelle, O., Metlzer, D., Zhang, Y., Grinspan, P.: Expected reciprocal rank for graded relevance. CIKM '09. pp. 621 - 630
- Biagioli, C., Francesconi, E., Passerini, A., Montemagni, S., Soria, C.: Automatic semantics extraction in law documents. ICAIL '05
- Moens, M.: Innovative techniques for legal text retrieval. Artificial Intelligence and Law pp. 29-57 (2001)