# Can Corpus Similarity-Based Self-Annotation Assist Information Retrieval?

Vinay Deolalikar

Groupon, Inc. Research (Work done when author was at Hewlett-Packard Labs)

## Enterprise internal corpora do not correspond to external taxonomies

- Enterprises generate their own internal unstructured text corpora
- R&D teams write documents, conduct experiments, present results, create prototypes, etc.
- Language and taxonomies are internal

## Problem Statement

Can we somehow extract information from the corpus itself (without recourse to any external means) that could be used *in lieu* of external annotations?

## Research Question

Can we extract information-theoretic measures of importance of terms from a corpus, and use these to self-annotate?

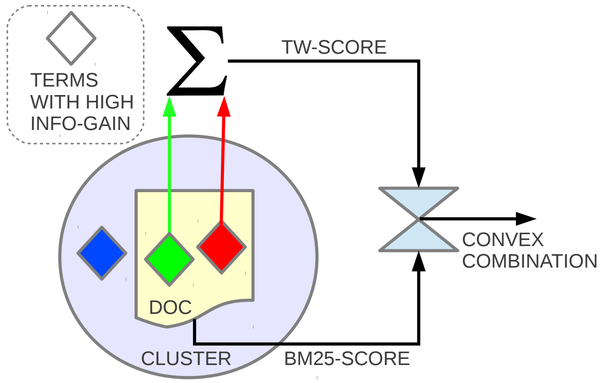
## Approach

- When we cluster a corpus, terms that are "important" in the corpus lead to the formation of clusters around them
- Terms that are informative with respect to cluster membership are important
- Using this importance in our retrieval system would constitute a form of "self-annotation"
- **BM25** does not use the similarity structure of a corpus in its term weighting We will augment **BM25** by weighting a term higher if it is more informative in the clustering

## Contrast to other approaches in IR

- In our scheme, documents within a cluster will not necessarily be ranked close to each other
- We do not use a cluster as a unit of retrieval
- Do not directly use cluster hypothesis

## Overview of document scoring scheme




**BM25** score is augmented by a score using the information-gain of terms relative to the cluster in which the document falls.

## Augmenting document scores

### Computing document score from clustering

Define the score $\mathbf{score_{tw}}(D)$ of a document $D$ as follows. $I[i,j]$ as the mutual information between the random variables $D_i$ and $t_{j,D}$. $N(*,*)$ is term frequency.

$$\mathbf{score_{tw}}(D) = \sum_{t_j \in D} I[i,j](1 + \log(N(t_j, D))). \tag{1}$$

Next, we restrict the scoring to contributions from terms that appear in the query $Q$.

$$\mathbf{score_{twq}}(D) = \sum_{t_j \in D \wedge t_j \in Q} I[i,j](1 + \log(N(t_j, D))). \tag{2}$$

### Combining into convex combination

Define two parametric families of convex combinations. $A$ is parameter.

1. combine $\mathbf{score_{BM25}}(\cdot)$ with $\mathbf{score_{tw}}(\cdot)$: $\mathbf{score_{tw,BM25}}(D, A) = A \cdot \mathbf{score_{BM25}}(D) + (1 - A) \cdot \mathbf{score_{tw}}(D)$
2. combine $\mathbf{score_{BM25}}(\cdot)$ with $\mathbf{score_{twq}}(\cdot)$: $\mathbf{score_{twq,BM25}}(D, A) = A \cdot \mathbf{score_{BM25}}(D) + (1 - A) \cdot \mathbf{score_{twq}}(D)$

## Datasets

- 125 queries from the topic distillation category of TREC 2003 and 2004
- total of 110,229 documents (about 1.85G) occur in the top-1000 **BM25** lists

## Design

- static clustering ($k = 200, 500$)
- query-specific clustering ($k = 20$)
- use top 20 terms (in information gain) for each cluster
- $A$ increased from 0 to 1 in steps of 0.1

## Results and Conclusions

Two patterns emerge

- For both $\mathbf{score_{tw,BM25}}(\cdot, A)$ and $\mathbf{score_{twq,BM25}}(\cdot, A)$, the precision falls monotonically with the proportion $A$ of the information gain that is used in document scoring.
- The precision of $\mathbf{score_{tw}}(\cdot)$ is consistently higher than that of $\mathbf{score_{twq}}(\cdot)$.

## Takeaway and Future Work

- Loss of precision seen in our information-gain augmented IR schemes should be seen in light of previous studies, which too showed negative results for cluster-based retrieval.
- In cluster-based retrieval, while studies have shown some evidence for the cluster hypothesis, finding the clusters that have many relevant documents is very hard to do automatically.
- (analogously) Is there a set of terms such that annotating relative to these terms will increase precision; yet finding the set of terms automatically is hard?
- (suggested future work by reviewer) Investigation of this technique in use cases other than plain search, for example, interactive query negotiation or navigation browsing, or other tasks for which a coarse grained clustering or classification structure is helpful.