ENHANCED NEWS RETRIEVAL: PASSAGES LEAD THE WAY!

Summary:

Catena et al. point out that most relevant terms are primarily concentrated at the beginning and the end of news articles. Using this idea, they try to improve the effectiveness of a retrieval systems in the news articles. They propose a new version of the BM25 weighting model, called BM25 Passage (BM25P). BM25P uses different weights for the different passages in the news articles. The relevance score is computed according to linear combination of term frequencies of each passage in the article. They experiment with Aquaint, Signal, and RCV1 news datasets using different settings. They show that their method gives better results than BM25 in term of both NDCG and MRR.

Contributions:

- Their method is the first contribution in the direction of exploiting in-document terms distributions within news documents.
- Signal and RCV1 datasets do not provide any evaluation data. Hence, they use a different methodology for evaluation. They assume that there is only one relevant article for each query. It is the article that the title belongs to. Using this idea, they successfully evaluate their system.

Positive Points:

- Their system indexes the unstructured body of news articles. This has a positive and a negative effect. It is positive because their system is able to work on unstructured news articles. Also, this type of index allows their system to differently weight the contribution of key terms.
- Their method is clearly better than BM25. Also, they make significant tests in order to tell whether their results are better.
- They claim that news articles are concentrated towards the beginning and the end of the documents. In order to prove this, they split each article into 10 passages and compute the distributions of the occurrences of top k terms with the highest IDF values. They show the results in figure 1.
- Instead of using a single dataset, they experiment with three different news datasets; Aquaint, Signal, and RCV1.
- Instead of using a single evaluation method, they evaluate their system with two different metrics. They use NDCG for Aquaint dataset; MRR for Signal and RCV1 datasets.

Negative Points:

- Their system indexes the unstructured body of news articles. This has both a positive and a negative effect. It is negative because their system ignores all collection-specific fields such as; titles, source, category, media type, and publishing date. Since, most news articles on web are in the structured form, they lose all of these information.
- They do not compare their results with a very similar weighting model BM25F.
- They get different results for different datasets. While, the best setting for Aquaint is $BM25P_{10}$, the best setting for Signal and RCV1 is $BM25P_5$. This creates questions in mind. They conclude that Signal and RCV1 benefit from more skewed probability distribution and these datasets are better when larger importance is given to the first and last passages.