CS 410 Tech Review:
Overview and Applications of Beta-Gamma Thresholding

Fall 2021

Jennifer Jasperse (jbj3)

Introduction

In text recommender systems, our task is to recommend a document to a user. Recommender systems filter retrieved documents in a way that aligns with either properties about the user (collaborative filtering) or properties about similar items (content-based filtering). However, the document space can often be large and the similarity matrices can be sparse, so we encounter a problem where we want to explore the document space to see if the user might be interested in documents not delivered by the earlier filtering methods. We define this dilemma as the exploration-exploitation tradeoff: we must choose whether to explore the document space or exploit the promising areas already identified.

Adaptive filtering allows us to balance the exploration-exploitation tradeoff. One such method is beta-gamma thresholding, an adaptive filtering method used to update thresholds in text recommender systems initially proposed by Zhai (1998). The threshold is inversely proportional to the number of documents in the system, meaning that the system would start to favor exploitation over exploration as the number of documents increases. It accounts for upper-bound bias in the optimal threshold introduced by the inclusion of training samples by setting a parameter β, which pushes the selected threshold closer to the lower-bound. The primary applications of beta-gamma thresholding are in text retrieval problems, but the approach can be generalized to work with any utility-based ranking problem.

Overview

Gamma-beta thresholding was introduced by Zhai (1998) and takes a heuristic-based approach to solve the exploration-exploitation tradeoff by selecting a threshold balanced between the optimal threshold and a cutoff lower-bound threshold. The task is defined below as:

$$\theta = \alpha * \theta_{zero} + (1 - \alpha) * \theta_{opt}$$

where θ is the selected threshold, θ_{zero} is the lower-bound cutoff threshold, and θ_{opt} is the upper-bound optimal threshold. In order to define θ_{zero} and θ_{opt}, we must also define a utility function for scoring document relevance. We then score each document until reaching K documents, where the utility score of the $(k+1)^{th}$ document yields the first negative score. K can be initialized heuristically using one of the following:

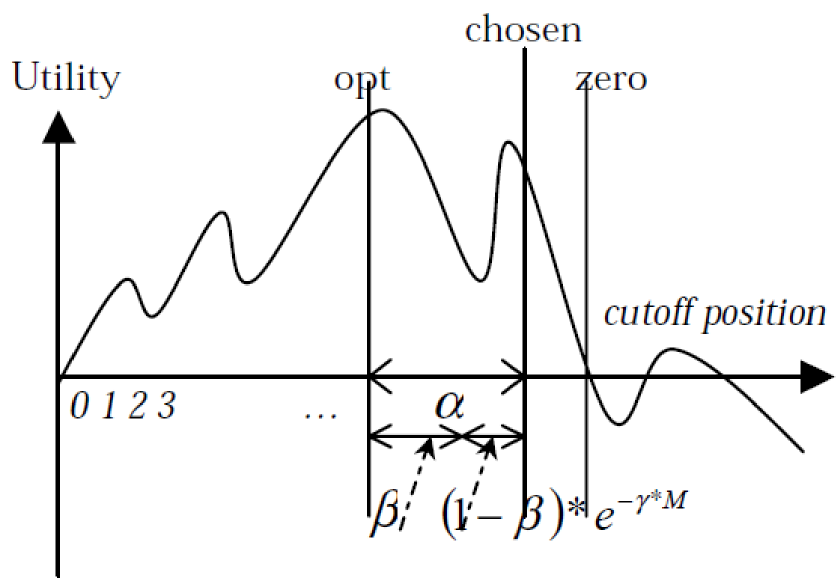
$$k = 10 + 10\log(N + 1)$$

$$k = r * N$$

where N is the number of documents available for training and r is the percentage of documents from the corpus that a user would like returned. The regularization parameter α controls the extent to which the lower- or upper-bound thresholds and is defined as:

$$\alpha = \beta + (1 - \beta) * e^{-\gamma M}$$

where β is a score bias correction factor, γ is the learning convergence speed that approximates the inverse of the number of documents at the midpoint of our range so that more training examples are directly proportional to our confidence in finding the true optimal threshold, and M is the number of documents already processed. β corrects the bias resulting from the incorporation of documents from the training set that have relatively higher scores. The beta-gamma thresholding method is visualized below:



Here, we see the relationship between θ, θ_{zero}, and θ_{opt} and their relative utility scores as the number of documents in the training set increases.

Beta-gamma thresholding balances the exploration-exploitation tradeoff by setting a lower-bound cutoff at the document with zero utility and an upper-bound at the optimal utility document found so far. However, that lower-bound cutoff often tends to be lower than necessary, potentially requiring additional search in the document space among low-utility documents. In the tests conducted in both Zhai (1998) and Zhai (2000) indicate that using this approach to adaptive filtering, while based solely on heuristic, was highly effective in the text recommendation task.

Applications

So far, the primary applications of beta-gamma thresholding have been in the field of information retrieval and text recommendation. One of the most common applications is web search. For example, Hiemstra (2001) applied beta-gamma thresholding integrated with language translation models for cross-language information retrieval, and Kukulenz (2007) applied beta-gamma thresholding to filter continuous query results among web documents for applications in cases where information is continuously and frequently refreshed, like financial stocks or event tickets that sell out quickly. This approach has also been used in polarity analysis. Salvetti (2004) introduces a ranking system based on the likelihood of a statement having positive or negative polarity and uses beta-gamma thresholding to determine the best cutoff point.

To generalize this approach to any utility-based ranking problem, we first define our set of possible items to be ranked as N. We treat each item in the ranking system as a document in the original approach and initialize our first k with the approximation provided. We then use the same definitions for our thresholds θ, θ_{zero}, and θ_{opt} and parameters α, β, and γ to calculate the selected threshold. This approach could be applied as a heuristic for balancing the exploration-exploitation tradeoff in movie or TV show recommendation, targeted online shopping advertisements, or news feed prioritization, among others.

Conclusion

Beta-gamma thresholding is able to heuristically address the exploration-exploitation tradeoff by dynamically updating the selected threshold for the retrieval task. It is able to incorporate a desired target results returned and the parameters defined in the algorithm help reduce scoring bias and allows for more exploration in the document space while the threshold confidence is low until the number of documents explored by the algorithm is large. This technique is most frequently applied in text retrieval, but could be expanded to improve any utility-based ranking problem.

References

Hiemstra, D. (2001). Using language models for information retrieval.

Kukulenz, D., & Ntoulas, A. (2007). Answering bounded continuous search queries in the world wide web. WWW '07.

Salvetti, F., & Reichenbach, C. (2004). Polarity Identification of Movie Reviews.

Zhai, C., Jansen, P., Stoica, E., Grot, N., & Evans, D. (1998). Threshold Calibration in CLARIT Adaptive Filtering. *TREC*.

Zhai, C., Jansen, P., & Evans, D. A. (2000). Exploration of a heuristic approach to threshold learning in adaptive filtering. *SIGIR Forum (ACM Special Interest Group on Information Retrieval)*, 360-362.