A Goodness of Fit Test Approach in Information Retrieval

Kostas Fragos[1], Yannis Maistros[2]

1 Department of Computer Engineering, National technical University of Athens, Iroon
Polytexneiou 9 15780 Zografou Athens Greece
http://nts.ece.ntua.gr/nlp_lab
kfragos@ece.ntua.gr
2 Department of Computer Engineering, National technical University of Athens, Iroon
Polytexneiou 9 15780 Zografou Athens Greece
maistros@ece.ntua.gr

Abstract.

In many probabilistic modeling approaches to Information Retrieval e are interested in estimating how well a document model ``fits'' he user's information need (query model). On the other hand in statistics, goodness of fit tests are well established techniques for assessing the assumptions about the underlying distribution of a data set. Supposing that the query terms are randomly distributed in the various documents of the collection, we actually want to know whether the occurrences of the query terms are more frequently distributed by chance in a particular document. This can be quantified by the so-called goodness of fit tests. In this paper, we present a new document ranking technique based on Chi-square goodness of fit tests. Given the null hypothesis that there is no association between the query terms *q* and the document *d* irrespective of any chance occurrences, we perform a Chi-square goodness of fit test for assessing this hypothesis and calculate the corresponding Chi-square values. Our retrieval formula is based on ranking the documents in the collection according to these calculated Chi-square values. The method was evaluated over the entire test collection of TREC data, on disks 4 and 5, using the topics of TREC-7 and TREC-8 (50 topics each) conferences. It performs well, outperforming steadily the classical OKAPI term frequency weighting formula but below that of KL-Divergence from language modeling approach. Despite this, we believe that the technique is an important non-parametric way of thinking of retrieval, offering the possibility to try simple alternative retrieval formulas within *goodness-of-fit* statistical tests' framework, modeling the data in various ways estimating or assigning any arbitrary theoretical distribution in terms.