Book Review

Simulating Information Retrieval Test Collections

David Hawking
Bodo Billerbeck
Paul Thomas
Nick Craswell
Synthesis Lectures on Information, Concepts, Retrieval and Services
Morgan & Claypool
2020
www.morganclaypool.com
ISBN: 9781681739571; 9781681739588; 9781681739595

Testing and experimenting with real corpus versus simulated col lection have wide variances. Simulated collection can able to yield results quickly, lead to change the population characteristics, validation, performance prediction and so on. This exercise can also lead to scale up corpora and would result in the predictive accuracy and efficiency of the data.

This interesting book has twelve well-knitted chapters. In the first chapter on Introduction, the authors have presented comprehensively the basic concepts and description about simulated text bed, characteristics, methods used, scope of simulation and its importance. This chapter besides also serve as summary of the book content.

In the second chapter on Evaluation Approaches, the authors distinguished the task based and nontask based evaluation. The measures for evaluation are treated briefly in this chapter.

The length of the documents has impact on the efficiency of the test results. Measuring the document length models are described in the third chapter on Modelling Document Length. Word frequencies play a role in IR system models. Those words that occur with high frequency particularly among few used words, have significance. The word frequency distribution and the behaviour of retrieval systems need to be addressed more to understand the process. These are explained in the fourth chapter on Modeling Word Frequencies, assuming independence. The baseline algorithm could model the word frequencies from text which is detailed in this unit.

In the next chapter on Modeling Term Dependence explains the patterns of word associations. It is useful if the distribution of words follows uniform way. The word frequencies, word co-occurrence and ngram are more descripted in this unit.

In the sixth chapter on Modeling Word Strings, the authors study the characteristics of word strings and enlist the methods for its generation. The descriptions of word strings are limited in this chapter as the authors claimed. This chapter is supported with plenty of illustrations and examples.

The size of the text corpus grows in size and other characteristics and hence the IR algorithms written should be scalable. The scalability of the algorithms versions is tested with Markov and numerical generators in the seventh chapter. In the eighth chapter on Generation of Compatible Queries, the authors have explained the methods of generating pseudo queries from annotations.

In the ninth chapter on Proof of Simulation Pudding, evaluated the simulated corpus collection constructed using several methods. The intention in this chapter is to find whether experiments conducted with simulations an able to predict the actual results. In the tenth chapter on Speed of Operation, the authors studied the speed of operation of many components of the synthesis corpus suite. The differential privacy issues of emulation methods are addressed in the eleventh chapter on Leaking Confidential Information. The last chapter brings the summary and discussion with a direction of future research, This book particularly the middle chapters address many technical issues and conceptual issues.

Reading this book will lead to gain an understanding the IR concepts.

Hathairat Ketmaneechairat
King Mongkut's University of Technology, North Bangkok
Thailand