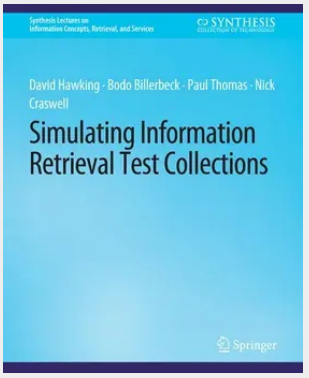
Hawking / Billerbeck / Thomas

Simulating Information Retrieval Test Collections



Simulated test collections may find application in situations where real datasets cannot easily be accessed due to confidentiality concerns or practical inconvenience. They can potentially support Information Retrieval (IR) experimentation, tuning, validation, performance prediction, and hardware sizing. Naturally, the accuracy and usefulness of results obtained from a simulation depend upon the fidelity and generality of the models which underpin it. The fidelity of emulation of a real corpus is likely to be limited by the requirement that confidential information in the real corpus should not be able to be extracted from the emulated version. We present a range of methods exploring trade-offs between emulation fidelity and degree of preservation of privacy.

We present three different simple types of text generator which work at a micro level: Markov models, neural net models, and substitution ciphers. We also describe macro level methods where we can engineer macro properties of a corpus, giving a range of models for each of the salient properties: document length distribution, word frequency distribution (for independent and non-independent cases), word length and textual representation, and corpus growth.

We present results of emulating existing corpora and for scaling up corpora by two orders of magnitude. We show that simulated collections generated with relatively simple methods are suitable for some purposes and can be generated very quickly. Indeed it may sometimes be feasible to embed a simple lightweight corpus generator into an indexer for the purpose of efficiency studies.

Naturally, a corpus of artificial text cannot support IR experimentation in the absence of a set of compatible queries. We discuss and experiment with published methods for query generation and query log emulation.

We present a proof-of-the-pudding study in which we observe the predictive accuracy of efficiency and effectiveness results obtained on emulated versions of TREC corpora. The study includes three open-source retrieval systems and several TREC datasets. There is a trade-off between confidentiality and prediction accuracy and there are interesting interactions between retrieval systems and datasets. Our tentative conclusion is that there are emulation methods which achieve useful prediction accuracy while providing a level of confidentiality adequate for many applications.

Many of the methods described here have been implemented in the open source project SynthaCorpus, accessible at: https://bitbucket.org/davidhawking/synthacorpus/

Simulated test collections may find application in situations where real datasets cannot easily be accessed due to confidentiality concerns or practical inconvenience. They can potentially support Information Retrieval (IR) experimentation, tuning, validation, performance prediction, and hardware sizing. Naturally, the accuracy and usefulness of results obtained from a simulation depend upon the fidelity and generality of the models which underpin it. The fidelity of emulation of a real corpus is likely to be limited by the requirement that confidential information in the real corpus should not be able to be extracted from the emulated version. We present a range of methods exploring trade-offs between emulation fidelity and degree of preservation of privacy.We present three different simple types of text generator which work at a micro level: Markov models, neural net models, and substitution ciphers. We also describe macro level methods where we can engineer macro properties of a corpus, giving a range of models for each of the salient properties: document length distribution, word frequency distribution (for independent and non-independent cases), word length and textual representation, and corpus growth.We present results of emulating existing corpora and for scaling up corpora by two orders of magnitude. We show that simulated collections generated with relatively simple methods are suitable for some purposes and can be generated very quickly. Indeed it may sometimes be feasible to embed a simple lightweight corpus

64,19 €
59,99 € (zzgl. MwSt.)

Lieferfrist: bis zu 10 Tage

Artikelnummer: 9783031011955
Medium: Buch
ISBN: 978-3-031-01195-5
Verlag: Springer
Erscheinungstermin: 04.09.2020
Sprache(n): Englisch
Auflage: Erscheinungsjahr 2020
Serie: Synthesis Lectures on Information Concepts, Retrieval, and Services
Produktform: Kartoniert
Gewicht: 363 g
Seiten: 162
Format (B x H): 191 x 235 mm

Kundenservice Verlag Dr. Otto Schmidt KG
Neumannstraße 10, 40235 Düsseldorf | kundenservice@otto-schmidt.de | 0221-93738-997 (Inland)