Computing Reviews

Association for Computing Machinery
ThinkLoud

Review

A nonparametric term weighting method for information retrieval based on measuring the divergence from independence
Kocaba$ İ., Dinçer B., Karaoğlan B. Information Retrieval 17(2): 153-176, 2014. Type: Article

Date Reviewed: 10/28/14

In information retrieval (IR), it is important to use effective term-weighting schemes. Existing works focus on the divergence from randomness (DFR) scheme, which assumes precise shape of the actual distribution of term frequency in the dataset. To overcome this issue, the authors propose a nonparametric approach, called divergence from independence (DFI), where a specialty word occurs in its semantically related content with a frequency that is different from that of a non-specialty word.

Based on DFI, the authors introduce three models: saturated model, standardization, and normalized chi-squared distance. Subsequently, they compare the retrieval accuracy that is obtained by these models, their variants, and five DFR-based approaches. The experimental results indicate that the DFI model that is based on standardization using inverse document frequency (IDF) works well for well-structured documents (such as newspapers and congressional records), while the DFI model that is based on normalized chi-squared distance works well for non-controlled documents (such as web pages).

Although the experiments reported in the paper are interesting, additional experiments would help to provide a better picture of the quality of the proposed models. For example, it has been reported in other works that log entropy is effective in the case of scholarly papers [1], which are a type of well-structured document. Hence, the authors should have compared their approach with log entropy in the context of scholarly papers. In addition, it is important to highly rank relevant documents in IR systems. Therefore, the authors should have evaluated their approach in terms of how many relevant documents are ranked in the top 1,000 results (1,000 precision) rather than the total number of relevant documents retrieved. Recently, social media content, which consists of short text, has emerged. Some works address this point by improving classical term-weighting methods [2,3]. It would have been nice if the authors had extended their experiments to cover social media content.

This paper will be helpful for IR researchers, especially those developing IR models.

1) Dumais, S. T. Improving the retrieval of information retrieval from external sources. *Behavior Research Methods, Instruments & Computers* 23, 2(1991), 229–236.

2) Phan, X.-H.; Nguyen, L.-M.; Horiguchi, S. Learning to classify short and sparse text & web with hidden topics from large-scale data collections . *In Proc. of the 17th International Conference on World Wide Web (WWW ’08).* ACM, (2008), 91–100.

3) Naveed, N. ; Gottron, T.; Kunegis, J. ; Alhadi, A. C. Searching microblogs: coping with sparsity and document quality. *In Proc. of the 20th ACM International Conference on Information and Knowledge Management (CIKM ‘11).* ACM, (2011), 183–188.

Reviewer: Kazunari Sugiyama Review #: CR142865