Abstract

The problem of set intersection counting appears as a subroutine in many techniques used in natural language processing, in which similarity is often measured as a function of document cooccurence counts between pairs of noun phrases or entities. Such techniques include clustering of text phrases and named entities, topic labeling, entity disambiguation, sentiment analysis, and search for synonyms. These techniques can have real-time constraints that require very fast computation of thousands of set intersection counting queries with little space overhead and minimal error. On one hand, while sketching techniques for approximating intersection counting exist and have very fast query time, many have issues with accuracy, especially for lists that have low Jaccard similarity. On the other hand, space-efficient computation of *exact* intersection sizes is particularly challenging in real-time. In this paper, we show how an efficient space-time trade-off can be achieved for set intersection counting, by combining state-of-the-art algorithms with precomputation and judicious use of compression. In addition, we show that the performance can be further improved by combining the best aspects of these algorithms. We present experimental evidence that real-time computation of exact intersection sizes is feasible with low memory overhead: we improve the mean query time of baseline approaches by over a factor of 100 using a data structure that takes merely twice the size of an inverted index. Overall, in our experiments, we achieve running times within the same order of magnitude as well-known approximation techniques.