AN OPEN SOURCE IMPLEMENTATION OF WEB CLUSTERING ALGORITHMS FOR SELECTIVE SEARCH

Gijs Hendriksen, Djoerd Hiemstra, and Arjen P. de Vries*
Radboud University, The Netherlands

Abstract

In distributed search, a document collection is partitioned across several *shards*, which can be queried independently to speed up query processing. Selective search builds upon this infrastructure, but reduces the required resources further by only querying a small number of the index shards. A resource selection algorithm is used to predict which shards are relevant for a given query. To ensure that this works effectively, the shards are usually created using a topic-driven clustering algorithm, so that different documents that are relevant for the same query are more likely to be assigned to the same shard. As a result, the resource selection step should become easier, and only a few shards need to be searched to obtain a high number of relevant results.

An effective clustering algorithm, *size bounded sampling-based K-means* (SB^2 K-means) [2], uses the lexical similarity of two documents as a proxy for their semantic similarity. A symmetric version of the negative Kullback-Leibler divergence is used to compare the unigram language models of two documents to determine how similar they are to one another. To scale the algorithm to larger document collections like Gov2[1] or ClueWeb09B,[2] the clustering step is performed on a smaller sample (e.g., 1%) of the documents, and the resulting centroids are used to assign the remaining documents to their respective shards. Additionally, in order to limit the number of shards that are much larger or much smaller than the average shard, large shards are re-partitioned using the same clustering algorithm, and small shards are merged together. The SB^2 K-means algorithm has been shown to result in relatively balanced shard maps (in terms of size) that enable effective selective search.

Dai et al. [1] noticed that the topics of content-based partitions do not necessarily align with the intent or information needs of a user. To resolve this mismatch, they use query logs to bias the similarity function towards terms that are more likely to be used by users. They also applied the same bias to the centroids used to initialize the K-means algorithm, by clustering the word embeddings of frequently occurring query terms and using those clusters to create an initial set of topics. The authors show that both the new similarity function (`QKLD`) and the new centroid seed selection (`QInit`) improve the performance of a selective search system. Additionally, they published the resulting shard maps for Gov2 and ClueWeb09B, so other researchers could easily use these for their own research on shard maps or selective search.[3]

While the algorithms and shard maps were made publicly available, there is not yet a public implementation of either clustering algorithm. This makes it difficult to replicate their results on alternative datasets, or apply one of the algorithms in a practical setting. In order to make the algorithms usable by the general public, and make it easier for researchers or search engine developers to implement and experiment with selective search systems, we release an open source implementation of SB^2 K-means, including the extensions `QKLD` and `QInit`. We use `scikit-learn`'s `CountVectorizer` to tokenize and vectorize the input documents.[4] We use GloVe embeddings [3] and `scikit-learn`'s agglomerative clustering implementation to run `QInit`. The K-means algorithm is implemented in Cython,[5] in order to make it more efficient and usable on large collections. Our implementation will be published as a Python package on PyPI.[6]

Our ongoing work focuses on replicating the shard maps by Dai et al., and ensuring our implementation produces shard maps of similar quality. We will also generate and publish shard maps for other collections, to verify how well selective search performs for alternate settings and datasets.

ACKNOWLEDGMENTS

This work has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101070014 (OpenWebSearch.EU, `https://doi.org/10.3030/101070014`).

REFERENCES

[1] Zhuyun Dai, Chenyan Xiong, and Jamie Callan. Query-Biased Partitioning for Selective Search. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, CIKM '16, pages 1119–1128, New York, NY, USA, October 2016. Association for Computing Machinery.

[2] Anagha Kulkarni. *Efficient and Effective Large-scale Search*. PhD thesis, Carnegie Mellon University, April 2013.

[3] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global Vectors for Word Representation. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics.

* {gijs.hendriksen, djoerd.hiemstra, arjen.devries}@ru.nl

[1] `http://ir.dcs.gla.ac.uk/test_collections/access_to_data.html`

[2] `https://lemurproject.org/clueweb09/`

[3] `https://boston.lti.cs.cmu.edu/appendices/CIKM2016-Dai/`

[4] `https://scikit-learn.org/`

[5] `https://cython.org/`

[6] `https://pypi.org/`