# Study Abstract

A text document corpus, where each document is associated with multiple topics assigned by humans, is often used when evaluating an automatic topic identification algorithm. It is well known that humans assign topics differently when multiple people work on the same documents due to their background knowledge in the domain of the documents, educational level, language proficiency including vocabulary size, etc. For this reason, inter-indexer consistency values among multiple human indexers are computed. Looking into demographic surveys of the indexers might give us some clue about low/high consistency values.

To evaluate the effectiveness of an automatic topic identification algorithm against human indexers, we use cosine inter-indexer consistency. First, we develop a topic dictionary using the entire set of topics from both humans and our algorithm. Based on this dictionary, topic sets from each indexer (including our algorithm) are converted into topic vectors. Then, the cosine angles between each topic vector and the ‘centroid’ vector (i.e., average of all the vectors) are computed and then compared with each other. The bigger angle represents that the topic group is less consistent to the average topics.

The ‘golden standard’ topic-indexed document corpus, which is produced as a result of this study, will be used to improve our automatic topic identification algorithm.