Reda Bouadjenek, Scott Sanner, Yihao Du
Department of Mechanical and Industrial Engineering
University of Toronto, Toronto, Canada
29 April 2019

To: Editor, Journal of Information Systems

Dear Editor,

Please find enclosed our manuscript "Relevance- and Interface-driven Clustering for Visual Information Retrieval", which we submit for consideration for publication in the Journal of Information Systems.

Our paper addresses the problem of search results of spatio-temporal data which are often displayed on a map or other visual interface. However, given the massive volume of available information in many applications (e.g., thousands of geolocated tweets matching a query), displaying all relevant results would often result in a saturated and unreadable display.

In many settings, it is natural to assume that search results cluster into spatially, temporally, and topically related content that can be aggregated and presented as a single unit rather than individual results. Such approaches leverage the cluster hypothesis of information retrieval, which posits that documents in the same cluster should behave similarly with respect to information needs. However, we argue in the paper that most existing work on aggregation for visual search that has sought to exploit the cluster hypothesis has focused on K-means and related unsupervised clustering methods that do not necessarily guarantee that clusters of matching search results are highly relevant. Moreover, the use of clustering algorithms such as K-means requires the design of a complex distance metric; for example, consider that space is often measured by Euclidean distance while keyword content is often measured by cosine distance and both of these distances need to be combined into a single distance metric for K-means. Such ad-hoc metric specifications do not necessarily guarantee the coherence of clusters from a visual, temporal and keyword content perspective.

To address the problems described above, we propose in this paper a clustering algorithm that given the relevance probability of each tweet to the search query, could automatically generate highly relevant clusters covering a large fraction of relevant content while explicitly optimizing for interface-driven desiderata of spatial, temporal, and keyword coherence without requiring any ad-hoc specification of a complex distance metric over these dimensions. Specifically, we present a novel relevance-driven clustering objective that extends standard information retrieval metrics to clusters.

We believe that this work provides the basis for new perspective of developing relevance-driven clustering algorithms formulated as an optimization problem. We note that all the material used and developed in this paper is publicly accessible and therefore will be available to the broader research community for future studies building on our approach.

We look forward to hearing from you.

Sincerely,

Reda, Scott, and Yihao