Text-Based Ephemeral Clustering for Web Image Retrieval on Mobile Devices

Jose G. Moreno
Univesité de Caen Basse-Normandie
jose.moreno@unicaen.fr

Abstract

In this thesis, we present a study about visualization of clustered Web image results on mobile devices. The recent advances in the research areas of Information Retrieval (particularly in search results clustering, Web mobile interfaces, query intent mining) and Natural Language Processing (in collocation measures, high order similarity metrics) have enabled the findings that are laid out in this study.

Our specific contributions in this dissertation are: two algorithms, two datasets and an evaluation tool. The Dual C-means algorithm is the main product of these. Dual C-means can be seen as a generalization of our previous proposal the AGK-means. Both algorithms are based on word-word similarity metrics and on the classical K-means algorithm. A new dataset for a complete evaluation of search results clustering (SRC) algorithms is developed and presented. Similarly, a new Web image dataset is developed and used together with a new metric to measure the users' effort when a set of Web images is explored. Finally, we developed an evaluation tool for the SRC problem, in which we have implemented several classical and recent SRC metrics.

In order to validate our hypothesis, we performed a great deal of different experiments with the datasets mentioned above. Three valuable characteristics are evaluated in the proposed algorithms. First, clustering quality, in which classical and recent evaluation metrics are considered. Secondly, the labelling quality of each cluster is evaluated to make sure that all possible query intents are covered. Thirdly and finally, we evaluate the user's effort in exploring the clustered results presented as horizontal image strips. For these three, we use several datasets– some of which are built to evaluate individual or combinations of these characteristics.

We have made our conclusions on the numerous factors discussed in this dissertation. In essence, the proposed Dual C-means algorithm, is capable of obtaining proper clustering partitions and simultaneously ensuring high quality labels. When images are displayed on a mobile device's interface, our results indicate that the users' effort to explore the Web image results is reduced.

Supervisor: Gaël Dias

Available: https://hal.archives-ouvertes.fr/tel-01102604/

Datasets/tool: http://websrc401.greyc.fr