Empirical Studies in Multi-Channel Clustering

Hila Weisman
hila.weisman@nice.com

Peter Izsak
peter.izsak@nice.com

Victor Shafran
victor.shafran@nice.com

NICE Systems
Zarhin 13, Ra'anana

Abstract

A major challenge in customer-oriented enterprises is conducting Multi-Channel Interaction Analytics: gaining insight from collaborated data sources. Today, end users interact with companies using websites, mobile applications, text messages, social media platforms or simply by approaching the nearest company store. When customers interact using these channels, they express themselves differently with a large degree of lexical variation, rendering Text Analytics (TA) solutions which treat each interaction channel equally as a non-effective approach to the task.

A prevalent TA solution for single-channel analytics is to automatically find the main topics in a set of interactions. Specifically, our in-house solution utilizes a state-of-the-art clustering algorithm to produce a division of the interactions to topics (clusters) with their corresponding labels. Our goal is to conduct thorough empirical studies on the characteristics of each channel with respect to different clustering algorithms, levels of supervision and evaluation metrics. We compare the performance of our in-house clustering solution within the different channels. In addition, we demonstrate that regardless of the chosen clustering algorithm, its performance will have great variance between the different channels.

Clustering algorithms can be classified as data-centric or description-centric [1]. Suffix Tree Clustering [2] represents a description-centric approach to

clustering and aims to create the best groupings that can be clearly labelled. Kmeans clustering [3] represents a data-centric approach to clustering and aims to directly represent the data at the cost of unclear labeling. Our in-house solution consists of a data-centric clustering algorithm which utilizes a rich set of linguistic and statistical features.

Our initial evaluation starts with comparing the performance of each of the three above-mentioned clustering algorithms over various channels by measuring the purity of the clusters. Preliminary results show that all clustering algorithms suffer from high performance variance between the different channels (almost 40% change in purity), indicating the need for further studies and adaptations. We continue with a per-channel evaluation by measuring the performance of each clustering algorithm. To that end, we use various performance measures such as purity, normalized Mutual Information, qualitative coherence of labeling etc.

We conclude with a discussion on the sensitivity of features per channel and describe future development directions we aim to pursue. For example, we would like to use insights from the Multi-Channel evaluations to tackle an even harder task, namely, Cross-Channel Analytics, where interactions from different channels are connected and viewed as a single interaction. For example, a customer makes a call to a contact center and then sends a feedback as an SMS message. We aim to utilize the models and insights from our studies to create a TA solution which treats Multi-Channel interactions as cross-channel.

References

[1] Carpineto, Claudio et al. *A survey of web clustering engines*. ACM Computing Surveys (CSUR). 2009.

[2] Zamir, Oren and Oren Etzioni *Web document clustering: A feasibility demonstration*. Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval. 1998.

[3] Lloyd., S. P. *Least squares quantization in PCM*. IEEE Transactions on Information Theory 28. 1982.