A new algorithm for two-mode text clustering

Livia Celardo[1], Domenica Fioredistella Iezzi[2], Maurizio Vichi[1]

[1] "La Sapienza" University - Rome
[2] "Tor Vergata" University - Rome

Keywords: co-clustering, text clustering, unstructured data

Today are available an enormous quantity of textual information in all fields, especially on the Web, that has become a strategic tool in finding and collecting data. According to IDC (2007), over 95% of the digital universe is unstructured data, and 80% of all stored data of organizations is unstructured. This massive amount of documents has made it difficult for users to obtain relevant information from them (Aliguliyev, 2009a); so, to manage information overload, it became necessary partitioning data into more compact forms, with the aim of describe, systematize and retrieve information (Balbi, 2012); it has been proposed various techniques, one of which is text clustering (Aliguliyev, 2009b). Text Clustering is an unsupervised one-way method that allows to classify large sets of documents in groups based on their attributes (Iezzi, 2012) with the aim of reproducing the internal structure of the data (Iezzi, 2010). While one-mode partition is widely utilized for information retrieval, very useful but still underused is the co-clustering approach (or two way clustering) that concerns simultaneous partitioning of rows and columns. In text mining contest, co-clustering is a very useful methodology (Balbi, 2012); the strength of this approach lies in finding clusters of documents characterized by groups of terms (Balbi, Miele and Scepi, 2010) with a high dimensionality reduction (Tjhi and Chen, 2006). The aim is to implement a new co-clustering procedure to classify not only the terms, but also the documents on the basis of the distinctive contents within every text; this process is planned to obtain the best clustering of words/documents in terms of the higher level of results interpretability.

References

Aliguliyev, R. M. (2009a). Performance evaluation of density-based clustering methods. *Information Sciences*, *179*(20), 3583-3602.

Aliguliyev, R. M. (2009b). Clustering of document collection–A weighting approach. *Expert Systems with Applications*, *36*(4), 7904-7916.

Balbi, S., Miele, R., and Scepi, G. (2010). Clustering of documents from a two-way viewpoint. In *10th Int. Conf. on Statistical Analysis of Textual Data*.

Balbi, S. (2012). Beyond the curse of multidimensionality: high dimensional clustering in text mining. Bolasco S. and Iezzi D.F. (by) *Advances in Textual Data Analysis and Text Mining - Special Issue Statistica Applicata - Italian Journal of Applied Statistics*.

J. F. Gantz, D. Reinsel, C. Chute, W. Schlichting, J. McArthur, S. Minton, I. Xheneti, A. Toncheva, and A. Manfrediz. (2007). *The expanding digital universe: A forecast of worldwide information growth through 2010*. IDC White paper, pages 1–20, March 2007.

Iezzi, D. F. (2010). Topic connections and clustering in text mining: an analysis of the JADT network. *Statistical Analysis of Textual Data, Rome, Italy*, *2*(29), 719-730.

Iezzi, D. F. (2012). Centrality measures for text clustering. *Communications in Statistics-Theory and Methods*, *41*(16-17), 3179-3197.

Tjhi, W. C., and Chen, L. (2006). A partitioning based algorithm to fuzzy co-cluster documents and words. *Pattern Recognition Letters*, *27*(3), 151-159.