{tag} {/tag}

IJCA Proceedings on International Conference in Computational Intelligence (ICCIA2012)

iccia - Number 6

Year of Publication: 2012

Authors:
Lalit A. Patil

S M. Kamalapur

Dhananjay Kanade

{bibtex}iccia1047.bib{/bibtex}

Abstract

Web mining techniques such as clustering help to organize the web content into appropriate subject based categories so that their efficient search and retrieval becomes manageable. Traditional WebPages clustering typically uses only the page content (usually the page text) in an appropriate feature vector representation such as Bags of words, termfrequency /inverse document frequency ,etc. and then applies standard clustering algorithms(e.g. K-means, Suffix tree, Query directed clustering). For example, Users can provide captions for images on the

internet, provide tags to WebPages and other media content they regularly browse on the internet, etc. Therefore such user – generated content can provide useful information in various form such as meta-data or in more explicit ways such as tags. Typically, WebPages clustering algorithms only use feature extracted from the page text. However, the advent also social –bookmaking websites, such as StumbleUpon and Delicious has led to a huge amount of usergenerated content such as the information that is associated with the WebPages. In multi-view learning, the feature can be split into two subset alone is sufficient for learning. Here as for, unsupervised learning algorithms, multiple views of the data can often help in extracting better features. Canonical Correlation Analysis (CCA) is an unsupervised feature extraction technique for finding dependencies between two (or more) views of the data by maximizing the correlations between the views in a shared subspace. But the drawbacks of the CCA is it gives The first approach is based on an annotation based probabilistic latent semantic analysis (LSA) over document-word and tagword co-occurrence matrices

References

- Anusua Trivedi, Piyush Rai, Scott L. DuVall “Exploiting Tag and Word Correlations for Improved Webpage Clustering “SMUC’10, October 30, 2010, Toronto, Ontario, Canada. Copyright 2010 ACM.
- S. Poomagal, Dr. T. Hamsapriya, “K-means for Search Results clustering using URL and Tag contents “978-1-61284- 764-1/11/$26.00 ©2011 IEEE.
- Lu, C., Chen, X., and Park, E. K. Exploit the tripartite network of social tagging for web clustering. In CIKM ’09 (2009), pp. 1545–1548.
- Ramage, D., Heymann, P., Manning, C. D., and Garcia- Molina, H. Clustering the tagged web. In WSDM ’09 (2009)
- Kakade, S. M., and Foster, D. P. Multi-view regression via canonical correlation analysis. In COLT’07 (2007)
- Ando, R. K., and Zhang, T. Two-view feature generation model for semi-supervised learning. In ICML ’07 (2007)
- Bach, F. R., and Jordan, M. I. Kernel independent component analysis. Journal of Machine Learning Research 3 (2003)
- Bao, S., Xue, G., Wu, X., Yu, Y., Fei, B., and Su, Z. Optimizing web search using social annotations. In WWW ’07 (2007)
- Bickel, S., and Scheffer, T. Multi-view clustering. In ICDM ’04 (Washington, DC, USA, 2004), IEEE Computer Society,
- Blaschko, M. B., and Lampert, C. H. Correlational spectral clustering. In CVPR (2008).
- http://www.stumbleupon.com
- http://www.delicious.com.

Index Terms

Computer Science Computational Intelligence

Keywords

Canonical Correlation Analysis probabilistic latent semantic analysis term-frequency Web page clustering