Insights

International Journal for Multidisciplinary Research

Volume 1, Issue: 1

Ramesh Kumar, B and M. A. Prasidh, 2016, 1(1): 58-64.

QUERY WEIGHTING WITH TW K MEANS CLUSTERING AND MIND MAP PROCESS FOR WEB DATA

Ramesh Kumar, B and M. A. Prasidh
Department of Computer Science
Sree Narayana Guru College,
Coimbatore – 641 105,
Tamil Nadu, India.

ABSTRACT

In the literature the author's studies about clustering methods, but it's not enough work has done for clustering. In this proposed work single weighted queries are not discussed, but the optimal variable weighting is studied. With the help of Mindmap techniques the K Means Clustering is providing more accurate results than existing approaches. Here we compared TW-k-means with five clustering algorithms on three real-life data sets and the results have shown that the TW-k-means algorithm significantly outperformed the other five clustering algorithms in four evaluation indices. The proposed work we have done modification work with the two types of weights, compact views and important variables can be identified and the effect of low-quality views and noise variables can be reduced. So that TW-k-means can obtain better clustering results than individual variable weighting clustering algorithms from multi view data. Here we discussed the difference of the weights between TW-k-means and EW-k means algorithms. The experiments also revealed the convergence property of the view weights in TW-k-means.

Keywords - Clustering, Web databases, Mind map, TW K Means.

INTRODUCTION

To the best of our knowledge, very few existing studies could be universally applied to unstructured data, semi-structured data, structured data and graph data. Therefore, providing both effective and efficient search abilities over such heterogeneous collections within a single search engine remains a big challenge. As it is, the structure of the data, such as the potentially hierarchical embedding in Extended Mark up Languages documents, is not fully exploited for answering key word queries. It is also not seeking into account for result ranking in most search engines. Current implementations focus on either IR-style search to meaningfully rank the results, but ignore the rich

structural information, or Data Base style search to discover the answers by identifying structural relationships but employ a very straightforward ranking mechanism.

The clustering analysis is a process that partitions a set of the objects into the groups, or clusters in such a way that objects from the same cluster are similar and objects from different clusters are dissimilar. The clustering approaches that are studied assume that the data objects to be clustered are independent and identical, and are often modeled by a fixed length vector of feature/attribute values. In the recent surge of data mining research, this classical problem was reexamined in the context of large databases. However, homogeneity of the data objects to be clustered seems still the basic assumption.

Multi-view algorithms train two independent hypotheses which bootstrap by providing each other with labels for the unlabeled data. The training algorithms tend to maximize the agreement between the two independent hypotheses. Dasgupta have shown that the disagreement between two independent hypotheses is an upper bound on the error rate of one hypothesis; this observation explains at least some of the often remarkable success of multi view learning. It also provides rise to the question whether the multi-view approach can be used to improve clustering algorithms. Partitioning methods such as k-Means, k-Medoids, and EM and hierarchical, agglomerative methods are among the clustering approaches most frequently used in data mining. We study multi-view versions of these families of algorithms for document clustering.

MATERIALS AND METHODS

DATA MODEL

Clustering algorithms can be divided into two categories: generative (or model-based) approaches and discriminative approaches.

Model-based approaches attempt to learn generative models from the documents, with each model representing one cluster. Usually generative clustering approaches are based on the Expectation Maximization algorithm. The EM algorithm is an iterative statistical technique for maximum likelihood estimation in settings with incomplete data. Given a model of data generation, and data with some missing values, EM will locally maximize the likelihood of the model parameters and give estimates for the missing values. Similarity-based clustering approaches optimize an objective function that involve the pair wise document similarities, aiming at maximizing the average similarities within clusters and minimize the average similarities between clusters. Most of the similarity based clustering algorithms follow the hierarchical agglomerative approach, where a dendrogram is build up by iteratively merging closest examples/ clusters.

Unfortunately, keyword search techniques used for locating information from collections of (Web) documents cannot be used on data stored in databases. In relational databases, information

needed to answer a keyword query is often split across the tables/tuples, due to normalization. This database contains paper titles, their authors and citations extracted from the DBLP repository. It depicts partial information paper title and authors about a particular paper. As we can see, the information is distributed across seven tuples related through foreign key references. In keyword based search, we need to identify tuples containing the keywords and ascertain their proximity through links. Answers to keyword queries on the Web are often only the starting point for further browsing to locate required information. Similar browsing facilities are needed in the context of searching for information from databases.

Multi-View EM Clustering

In this section we want to analyze whether we can extend EM based cluster algorithms, so that they incorporate the multi-view setting with independent views. Different EM applications differ in specific models. We focus on models that are suitable for document clustering. Gaussian models could be used for multi-view EM as well, but are not applicable for document clustering. We firstly describe the general EM algorithm extended for two views, then we describe two instances of this algorithm and present and analyze empirical results.

General Multi View EM Algorithm

In the field of semi-supervised learning, co-EM based methods Positive results on the co-EM algorithm for the problem of semi-supervised learning lead to the question whether co-EM can improve on EM for unsupervised learning setting as well. The co-EM algorithm. In each iteration i, each view v finds the model parameters £(v) i which maximize the likelihood given the expected values for the hidden variables of the other view. In turns M, E steps in view one and M, E steps in view two are executed. The single expectation and maximization steps are equivalent to the E and Msteps of the original EM algorithm. The algorithm is not guaranteed to converge. Our experiments show that the algorithm often does not converge. As displayed in Table 1, we do not run the algorithm until convergence but until a special stopping criterion is met.

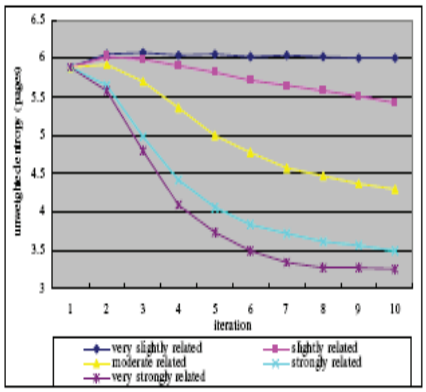
RESULTS AND FINDINGS

The results show that the degree of how tightly the objects are interrelated with each other has significant impact on the clustering accuracy. When objects become more strongly interrelated, our iterative the clustering results may improve in accuracy. But improvement will decrease, as we can see that when the relative density of relationships is larger than 80%, the clustering accuracy is very close to that of 100% relationships. More relationships will not help much compared to the cost of computational complexity.

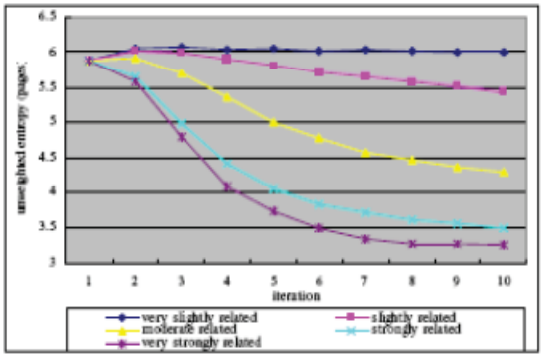
Table 1. Multi-View EM.

Input: Unlabeled data $D = \{(x_1^{(1)}, x_1^{(2)}), \ldots, (x_n^{(1)}, x_n^{(2)})\}$.

1. Initialize $\Theta_0^{(2)}, T, t = 0$.
2. E step view 2: compute expectation for hidden variables given the model parameters $\Theta_0^{(2)}$
3. Do until stopping criterion is met:
   (a) For $v = 1 \ldots 2$:
      i. $t = t + 1$
      ii. M step view v: Find model parameters $\Theta_t^{(v)}$ that maximize the likelihood for the data given the expected values for the hidden variables of view $\bar{v}$ of iteration $t - 1$
      iii. E step view v: compute expectation for hidden variables given the model parameters $\Theta_t^{(v)}$
   (b) End For v.
4. return combined $\hat{\Theta} = \Theta_{t-1}^{(1)} \cup \Theta_t^{(2)}$



(a)



(b)

Figure-1. Clustering accuracy when dataset evolves from slightly interrelated to strongly interrelated

The density of relationships between the two types of objects has significant impact on the clustering accuracy. In Figure 1 we empirically analyze how clustering accuracy evolves when the

two types of objects are interrelated from slightly to strongly. In this experiment, we randomly select 20%, 40%, 60%, 80% and 100% of the user/web-page relationships to represent different degree of how objects are interrelated. The parameter □is fixed to 0.4 here.

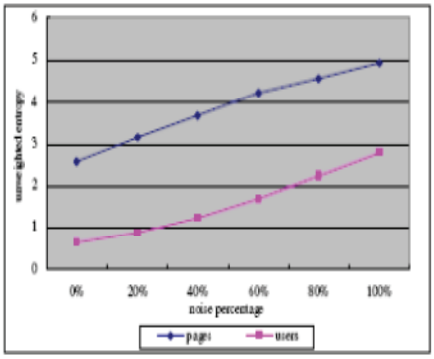


Figure 2. Clustering accuracy along with noise ratio

The next two comparative studies will show the effectiveness of weighted clustering and our link analysis algorithm under noise. Figure 3 shows that a clustering algorithm which incorporates the weights of objects will obtain better final clustering results, especially, in large noise ratios. By considering their importance value, important users or web-pages will have larger impact in forming the correct clusters. This may decrease the influence of the noise relationships.

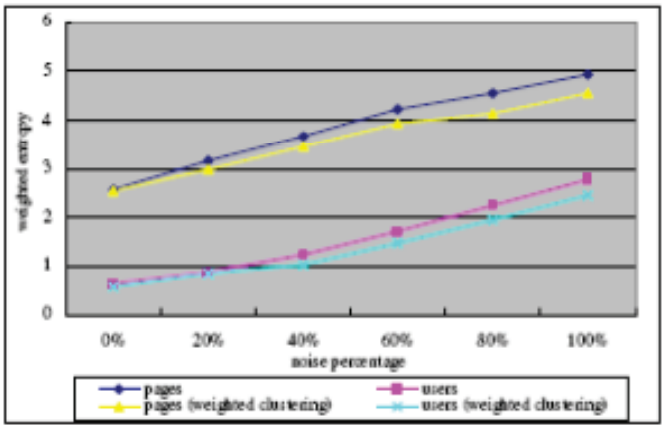


Figure 3. Comparison between importance weighted clustering and un-weighted clustering

From these experiments we can draw the conclusion that the weighted iterative clustering in our framework behaves well compared to traditional approaches. The introduction of relationship in defining object features and reinforcement clustering by relationship proves effective for the interrelated multi-type data objects.

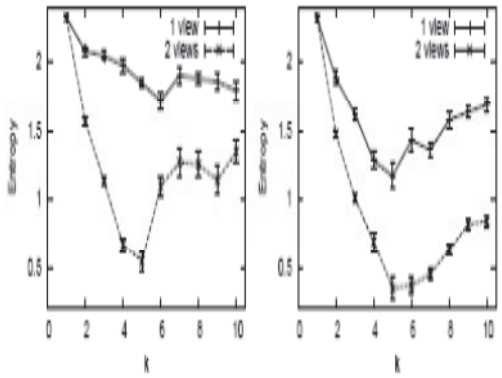


Figure 4. Single and multi-view mixture-ofmultinomials EM (left) and spherical k-Means (right) for the artificial data set and different k.

The comparison of multi-view mixture-of-multinomial EM and spherical k-Means with their single-view counterparts for the WebKB data set is shown in Figure 1. The number of clusters is set to k = 6. Figure 2 displays the same setting, but with different number of desired clusters k. We notice a tremendous improvement of cluster quality with the multi-view algorithms. Figure 4 displays results for the artificial data set built from the 20 newsgroup data set, as described. On this data set the multi-view algorithms improve the entropy even more than for the WebKB data set. The total independence property of the artificial data set seems to support the success of multi-view EM. Figure 4 shows the results for the six document data sets without natural multi-view property, where we randomly split the available attribute sets into two subsets. In ten of twelve cases the multi-view outperform the single-view algorithms significantly.

CONCLUSION AND FUTURE WORK

TW k- means can compute weights for views and individual variables simultaneously in the process of clustering. With the two types of weights, compact views and important variables can be identified and effect of low-quality views and noise variables can be reduced. TW-k-means can obtain better clustering results than individual variable weighting clustering algorithms from multi view data. We used two real-life data sets to investigate the properties of two types of weights in TW-k-means. We discussed the difference of the weights between TW-k-means and EW- k means algorithms As such, it is a new variable weighting method for clustering of multi view data.

The work we will combine in the future is two-level variable weighting method with other techniques such as fuzzy techniques, subspace clustering techniques, semi-supervised techniques etc. We will investigate the approaches that can automatically group variables in the clustering process.

We are exploring support for external links, such as HTML HREFs, to aid in browsing. Such support is particularly useful when integrating data's from multiple web databases. Other planned

system features include authorization mechanisms to selectively expose data to different users. Query evaluation with keywords matching metadata can be relatively slow, since a large number of tuples may be defined to be relevant to the keyword. This problem also arises with non-metadata keywords that match large number of nodes. We are working on techniques to fast up such queries by not performing backward search from large numbers of nodes, and instead searching forwards from probable information of nodes corresponding to more selective keywords.

REFERENCES

1. J. Mui and K. Fu, "Automated Classification of Nucleated Blood Cells Using a Binary Tree Classifier," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 2, no. 5, pp. 429-443, May 1980.
2. J. Wang, H. Zeng, Z. Chen, H. Lu, L. Tao, and W. Ma, "ReCoM: Reinforcement Clustering of multiType Interrelated Data Objects," Proc. 26th Ann. Int'l ACM SIGIR Conf. Research and Development in Informaion Retrieval, pp. 274-281, 2003.
3. S. Bickel and T. Scheffer, "Multi-view Clustering," Proc. IEEE Fourth Int'l Conf. Data Mining, pp. 19-26, 2004.
4. K. Kailing, H. Kriegel, A. Pryakhin, and M. Schubert, "Clustering Multi-Represented Objects with Noise," Proc. Eighth Pacific-Asia Conf. Knowledge Discovery and Data Mining, H. Dai, R. Srikant, and C. Zhang, eds., vol. 3056, pp. 394-403, Springer Berlin/Heidelberg, 2004.
5. V.R. de Sa, "Spectral Clustering with Two Views," Proc. IEEE 22nd Int'l Workshop Learning with Multiple Views (ICML), pp. 20-27, 2005.
6. D. Zhou and C. Burges, "Spectral Clustering and Transductive Learning with Multiple Views," Proc. 24th Int'l Conf. Machine Learning, pp. 1159-1166, 2007.
7. M.B. Blaschko and C.H. Lampert, "Correlational Spectral Clustering," Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR '08), pp. 1-8, 2008.
8. K. Chaudhuri, S. Kakade, K. Livescu, and K. Sridharan, "Multiview Clustering via Canonical Correlation Analysis," Proc. 26th Ann. Int'l Conf. Machine Learning, pp. 129-136, 2009.
9. G. Tzortzis and C. Likas, "Multiple View Clustering Using a Weighted Combination of Exemplar-Based Mixture Models," IEEE Trans. Neural Networks, vol. 21, no. 12, pp. 1925-1938, Dec. 2010.
10. B. Long, P. Yu, and Z. Zhang, "A General Model for Multiple View Unsupervised Learning," Proc. Eighth SIAM Int'l Conf. Data Mining (SDM '08), 2008.