International Journal of Engineering Research and General Science Volume 3, Issue 1, January-February, 2015
ISSN 2091-2730

DEDUCTION TECHNIQUES FOR UNSUPERVISED CLUSTERING WITH SIMILARITY HIDDEN CLUSTERING MODEL

N.Senthilkumaran., MCA. M.Phil., Assistant Professor & Head, Department of Computer Applications, Email: n.senthilkumaran@hotmail.com.
S.Priya, Research Scholar, Email: priya91devi@gmail.com, Ph:8883038541. Vellalar College for Women (Autonomous), Erode

Abstract— Document clustering is automatically group related document into cluster. In this clustering frame work focus on correlations between the documents in the local patches are maximized while the correlations between the documents outside these patches are minimized simultaneously. The proposed systems are adopts both supervised and unsupervised constraints to demonstrate the effectiveness of the proposed algorithm in this framework. The novel proposed classical K-Mean clustering algorithm applied for data preprocessing in stop word removal, stemming and synonym word replacement to apply semantic similarity between words in the documents. In addition, content can be retrieved from text files, HTML pages as well as XML pages. Tags are eliminated from HTML files. Attribute name and values are taken as normal paragraph words in XML files and then preprocessing (stop word removal, stemming and synonym word replacement) is applied. In addition, TEXT, HTML and XML documents are cluster using cosine similarity model implement these research works.
Keywords: clustering,supervised constraints, un supervised constraints,co-clustering,k-means clustering

INTRODUCTION

Clustering is a data mining technique of grouping set of data objects into multiple groups or clusters so that objects within the cluster have high similarity, but are very dissimilar to objects in the other clusters. Dissimilarities and similarities are assessed based on the attribute values describing the objects. Clustering algorithms are used to organize data, categorize data, for data compression and model construction, for detection of outliers etc. Common approach for all clustering techniques is to find clusters centre that will represent each cluster. Cluster centre will represent with input vector can tell which cluster this vector belong to by measuring a similarity metric between input vector and all cluster centre and determining which cluster is nearest or most similar one .

Cluster analysis can be used as a standalone data mining tool to gain insight into the data distribution, or as a preprocessing step for other data mining algorithms operating on the detected clusters. Many clustering algorithms have been developed and are categorized from several aspects such as partitioning methods, hierarchical methods, density-based methods, and grid-based methods. Further data set can be numeric or categorical. Inherent geometric properties of numeric data can be exploited to naturally define distance function between data points. Whereas categorical data can be derived from either quantitative or qualitative data where observations are directly observed from counts.

When clustering textual data, one of the most important distance measures is document similarity. Since document similarity is often determined by word similarity, the semantic relationships between words may affect document clustering results. For example, sharing common named entities (NE) among documents can be a cue for clustering these documents together. Moreover, the relationships among vocabularies such as synonyms, antonyms, hypernyms, and hyponyms, may also affect the computation of document similarity. Consequently, introducing additional knowledge on documents and words may facilitate docu-ment clustering. To incorporate word and document constraints, we propose an approach called constrained information-theoretic coclustering (CITCC). It integrates constraints into the information theoretic coclustering (ITCC) framework [4], where KL-divergence is adopted to better model textual data. The constraints are modeled with two-sided hidden Markov random field (HMRF) regular-izations. We develop an alternating expectation maximiza-tion (EM) algorithm to optimize the model. As a result, CITCC can simultaneously cluster two sets of discrete random variables such as words and documents under the constraints extracted from both sides.

2. RELATED WORK

2.1 CITCC METHOD

Clustering is a popular technique for automatically organizing or summarizing a large collection of text there have been many approaches to clustering. Unlike traditional clustering methods that focus on 1D clustering, co clustering examines both document and word relationship at the same time. In addition to co clustering approaches, researchers have also developed constrained clustering methods to enhance document clustering.The purely un-supervised document clustering is often difficult, most constrained clustering approaches are semi-supervised and requires the use of manually labeled constraints.

The Constrained information theoretic co-clustering (CITCC) method describes the constrained co clustering problem as a two-sided HMRF regularized ITCC (HMRF2 -ITCC) model is formulated. Then we present how to use an alternating EM algorithm to optimize the model.

2.1.1 PROBLEM FORMULATION

The document set and word set as D= {d_1; d_2; ...; d_M} and V={ U_1, U_2... ; U_V}. Then the joint probability of p(d_m; U_i) can be computed based on the co- occurrence count of d_m and U_i

$$q(d_m, v_i) = p(\hat{d}_{k_d}, \hat{v}_{k_v}) p(d_m|\hat{d}_{k_d}) p(v_i|\hat{v}_{k_v}),$$

where $\hat{d}_{k_d}$ and $\hat{v}_{k_v}$ are cluster indicators, k_d and k_v are the cluster indices, is used to approximate p=(d_m, U_i) by minimizing the Kullback-Leibler (KL) divergence where $\hat{\mathcal{D}}$ and $\hat{\mathcal{V}}$ are the cluster sets $p(\mathcal{V}|d_{k_d})$ denotes a multinomial distribution based on the probabilities $(p(v_1|\hat{d}_{k_d}), \ldots, p(v_V|\hat{d}_{k_d}))^T$, $p(v_i|\hat{d}_{k_d})) = p(v_i|\hat{v}_{k_v}) p(\hat{v}_{k_v}|\hat{d}_{k_d})$ and $p(v_i|\hat{v}_{k_v}) = p(v_i)/p(l_{v_i} = \hat{v}_{k_v})$. Symmetrically, we can define the probability for words: $p(\mathcal{D}|\hat{v}_{k_v})$ denotes a multinomial distribution based on the probabilities

$(p(d_1|\hat{v}_{k_v}), \ldots, p(d_V|\hat{v}_{k_v}))^T$, $p(d_i|\hat{v}_{k_v})) = p(d_i|\hat{d}_{k_d}) p(\hat{d}_{k_d}|\hat{v}_{k_v})$ and $p(d_i|\hat{d}_{k_d}) = p(d_i)/p(l_{d_i} = \hat{d}_{k_d})$.

$$\begin{aligned} & D_{KL}(p(\mathcal{D}, \mathcal{V}) \| q(\mathcal{D}, \mathcal{V})) \\ &= D_{KL}(p(\mathcal{D}, \mathcal{V}, \hat{\mathcal{D}}, \hat{\mathcal{V}}) \| q(\mathcal{D}, \mathcal{V}, \hat{\mathcal{D}}, \hat{\mathcal{V}})) \\ &= \sum_{k_d}^{K_d} \sum_{d_m : l_{d_m} = k_d} p(d_m) D_{KL}(p(\mathcal{V}|d_m) \| p(\mathcal{V}|\hat{d}_{k_d})) \\ &= \sum_{k_v}^{K_v} \sum_{v_i : l_{v_i} = k_v} p(v_i) D_{KL}(p(\mathcal{D}|v_i) \| p(\mathcal{D}|\hat{v}_{k_v})), \end{aligned}$$

2.2 UNSUPERVISED CONSTRAINTS

Unsupervised Constraints show how to generate additional semantic constraints for clustering. Specifically, introduce named-entity-based document constraints and un reliable word -based constraints using the following approaches.

Document Constraints

In practice, document constraints constructed based on the human annotations are difficult to obtain. To scope with this problem, in this work, we propose new methods to derive "good but imperfect" constraints using information retrieval automatically extracted from either the content of a document (e.g., NE constraints) or existing knowledge sources (e.g., Wordnet constraints). Similarly, if two documents share the same organization names such as "AIG," "Lehman Brothers," and "Merrill Lynch," then both of them may be belong to the same document cluster about the financial markets. Consequently, the document must-link constraints can be constructed from the correlated named entities such as person, location, and organization. Specifically, if there are overlapping NEs in two documents and the number of overlapping NEs is larger than a predefined threshold, add a must-link to these documents.

Word Constraints

Besides named-entity-based document constraints, it is possible to incorporate additional lexical constraints derived from existing knowledge sources to further improve clustering results. In our experiment result the information in WordNet, an online lexical database, to construct word constraints. Specifically, the semantic distance of two words can be computed based on their relationships in WordNet.

Word document is construct word must-links based on semantic distances, for example, we can add a word must-link if the distance between two words is less than a threshold, additional lexical information can be seamlessly incorporated into the clustering algorithm to derive better word clusters. Moreover, since word knowledge can be transferred to the document side during coclustering, with additional word constraints, it is possible to further improve document clustering.

2.3 KL DIVERGENCE

In probability theory and information theory, the Kullback–Leibler divergence (also information divergence, information gain, relative entropy or KLIC, here abbreviated as KL divergence) is a non-symmetric measure of the difference between two probability distributions *P* and *Q*. Specifically, the Kullback–Leibler divergence of *Q* from *P*, denoted $D_{KL}(P||Q)$, is a measure of the information lost when *Q* is used to approximate *P*:[4] The KL divergence measures the expected number of extra bits required to code

samples from *P* when using a code based on *Q*, rather than using a code based on *P*. Typically *P* represents the "true" distribution of data, observations, or a precisely calculated theoretical distribution. The measure *Q* typically represents a theory, model, description, or approximation of *P*.Although it is often intuited as a metric or distance, the KL divergence is not a true metric for example, it is not symmetric: the KL divergence from *P* to *Q* is generally not the same as that from *Q* to *P*. However, its infinitesimal form, specifically its Hessian, is a metric tensor: it is the Fisher information metric.

2.4 K MEANS CLUSTERING

K-means clustering is a data mining the machine learning algorithm used to cluster observations into groups of related observations without any prior knowledge of those relationships. The k-means algorithm is one of the simplest clustering techniques and it is commonly used in medical imaging, biometrics and related fields.

The k-means algorithm is an evolutionary algorithm that gains its name from its method of operation. The algorithm clusters observations into k groups, where k is provided as an input parameter. It then assigns each observation to clusters based upon the observation's proximity to the mean of the cluster. The cluster's mean is then recomputed and the process begins again. Here's how the algorithm works:

Step 1: The algorithm arbitrarily selects k points as the initial cluster centers ("means").

Step 2: Each point in the dataset is assigned to the closed cluster, based upon the Euclidean distance between each point and each cluster center.

Step 3: Each cluster center is recomputed as the average of the points in that cluster.

step 4: Steps 2 and 3 repeat until the clusters converge. Convergence may be defined differently depending upon the implementation, but it normally means that either no observations change clusters when steps 2 and 3 are repeated or that the changes do not make a material difference in the definition of the clusters.

2.5. PROPOSED CLUSTERING METHODS

2.5.1 K-MEANS

K-means is the most important flat clustering algorithm. The objective function of K- means is to minimize the average squared distance of objects from their cluster centers, where a cluster center is defined as the mean or centroid μ of the objects in a cluster C.

The ideal cluster in K-means is a sphere with the centroid as its center of gravity. Ideally, the clusters should not overlap. A measure of how well the centroids represent the members of their clusters is the Residual Sum of Squares (RSS), the squared distance of each vector from its centroid summed over all vectors.

K-means can start with selecting as initial clusters centers K randomly chosen objects, namely the seeds. It then moves the cluster centers around in space in order to minimize RSS. This is done iteratively by repeating two steps until a stopping criterion is met

- Reassigning objects to the cluster with closest centroid
- Recomputing each centroid based on the current members of its cluster.

The following termination conditions as stopping criterion for using termination process

- A fixed number of iterations I has been completed.
- Centroids μ i do not change between iterations.
- Terminate when RSS falls below a pre-established threshold.

2.5.2 CLASSICAL K-MEANS ALGORITHMS (Both TEXT, HTML, and XML Documents)

1. procedure KMEANS(X,K)
2. {s1, s2, · · · , sk} SelectRandomSeeds(K,X)
3. for i ← 1,K do
4. μ(Ci) ← si
5. end for
6. repeat
7. min k~x n −~μ(C k)k C k = C k [{~x n }
8. for all C k do
9. μ(C k) = 1
10. end for
11. until stopping criterion is met
12. end procedure

The proposed algorithm fall within a subcategory of the flat clustering algorithms, called Model-based clustering. The model-based clustering assumes that data were generated by a model and then tries to recover the original model from the data. This model then defines clusters and the cluster membership of data. The proposed algorithm is a generalization of K-Means algorithm in which

the set of K centroids as the model that generate the data. It alternates between an expectation step, corresponding to reassignment, and a maximization step, corresponding to re computation of the parameters of the model.

2.5.3 COSINE SIMILARITY

In this approach, two documents are selected. Then the vector values for two documents are found out. The cosine similarity measure is applied. Then the correlation between two documents is found out using the following formula.

$$Corr(u,v) = \frac{u^T v}{\sqrt{u^T u}\sqrt{v^T v}} = \left\langle \frac{u}{\|u\|}, \frac{v}{\|v\|} \right\rangle$$

Correlation Formula

For example, the string "I have to go to school" is present in one document. the string "I have to go to temple" is present in other document.

Then the data is prepared such that

[i , have , to , go, school , temple] = [1,1,2,1,1,0]

[i , have , to , go , school , temple] = [1,1,2,1,0,1]

[i , have , to , go , school , temple] = [1,1,2,1,0,1]

Formula: cos = 1*1 + 1*1 + 2*2 + 1*1 + 1*0 + 0*1 / sqrt((1^2 + 1^2 + 2^2 + 1^2 + 1^2 + 0^2) * 1^2 + 1^2 + 2^2 + 1^2 + 1^2 + 0^2).

3. EXPERIMENTAL RESULT

The following Table 1 describes experimental result for proposed system analysis. The table contains weight of text document, HTML documents and XML documents, weight of clustering text document, HTML documents, XML documents and average of text document, HTML documents, and XML documents clustering details are shown.

s.no	Weight of Document (TEXT, XML, HTML)	Weight of Clustering Document			Proposed system Clustering Document [%]		
		HTML	XML	TEXT	HTML	XML	TEXT
1	200	166	172	176	83	86	88
2	250	230	234	234	92	93.6	93.60
3	300	286	287	291	95.33	97	97
4	350	334	338	342	95.42	96.57	97.71
5	400	391	394	398	97.75	98.5	99.50
6	450	436	442	445	96.88	98.22	98.89
7	500	473	479	489	94.60	95.80	97.80
8	550	537	545	547	97.62	99.10	99.45
9	600	586	589	592	97.77	98.17	98.62
10	650	638	643	646	98.15	98.92	99.38

The following **Fig 1** describes experimental result for proposed system analysis. The table contains weight of text document, HTML documents and XML documents, weight of clustering text document, HTML documents, XML documents and average of text document, HTML documents, and XML documents clustering details are shown.

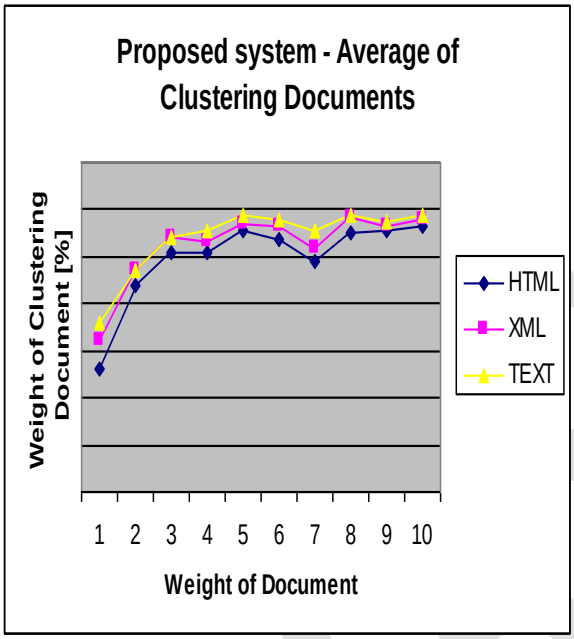

Figure 1 - Average of Clustering Documents

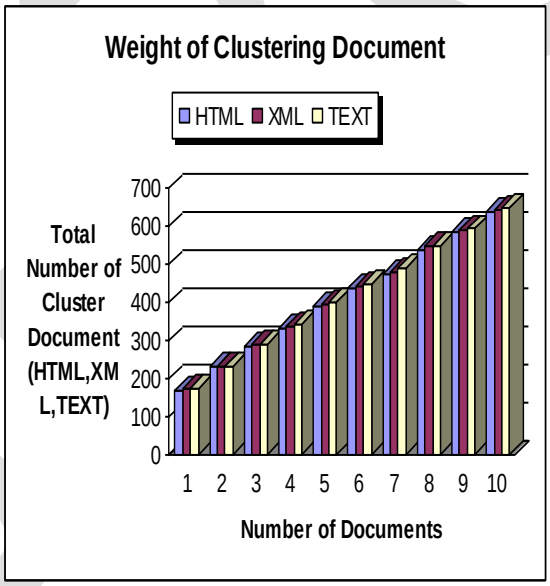

Figure 2- Weight of Clustering Documents

4. CONCLUSION

This proposed framework demonstrated how to construct various document and word constraints and apply them to the constrained coclustering process. A novel constrained coclustering approach is proposed that automatically incorporates various word and document constraints into information-theoretic coclustering. It demonstrates the effectiveness of the proposed method for clustering textual documents.

There are several directions for future research. The current investigation of unsupervised constraints is still preliminary. Furthermore, the algorithm consistently outperformed all the tested constrained clustering and coclustering methods under different conditions. The enhanced cosine similarity approach results in better clustering process.

5. FURTHER WORK

The future enhancements can be made for documents of different languages. Investigation for better text features that can be automatically derived by using natural language processing or information extraction tools can be made.

REFERENCES

[1] Yangqiu Song , Shimei Pan, Shixia Liu And Furu Wei ,"Constrained Text Coclustering With Supervised And Unsupervised Constraints"Vol.25,No.6,2013.
[2] F. Wang, T. Li, and C. Zhang, "Semi-Supervised Clustering via Matrix Factorization," Proc. SIAM Int'l Conf. Data. Mining (SDM), pp. 1-12, 2008
[3] R.G. Pensa and J.-F. Boulicaut, "Constrained Co-Clustering of Gene Expression Data," Proc. SIAM Int'l Conf. Data Mining (SDM), pp. 25-36, 2008.
[4] A.Banerjee, I. Dhillon, J. Ghosh, S. Merugu, and D.S. Modha, "A Generalized Maximum Entropy Approach to Bregman Co-Clustering andMatrix Approximation," J. Machine Learning Research, vol. 8, pp. 1919-1986, 2007.
[5] B.Long, X. Wu, Z. (Mark) Zhang, and P. S. Yu. Unsupervised Learning on K-partite Graphs, In Proc. of SIGKDD, 317-326, 2006.
[6] R. Pensa, C. Robardet, and J-F. Boulicaut. Towards constrained co-clustering in ordered 0/1 data sets. In Proceedings ISMIS 2006, volume 4203 of LNCS, pages 425–434, Bari, Italy, 2006. Springer.
[7] B. Kulis, S. Basu, I. Dhillon, R. J. Mooney. Semi- Supervised Graph Clustering: A Kernel Approach. In Proc. of ICML, 457-464, 2005.
[8] H. Cho, I.S. Dhillon, Y. Guan, and S. Sra, "Minimum Sum-Squared Residue Co-Clustering of Gene Expression Data," Proc. Fourth SIAM Int'l Conf. Data. Mining (SDM), 2004.
[9] I.S. Dhillon, S. Mallela, and D.S. Modha, "Information-Theoretic Co-Clustering," Proc. Ninth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), pp. 89-98, 2003.
[10] I.S. Dhillon, "Co-Clustering Documents and Words Using Bipartite Spectral Graph Partitioning," Proc. Seventh ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), pp. 269-274, 2001.
[11] K. Nigam, A.K. McCallum, S. Thrun, and T.M. Mitchell, "Text Classification from Labeled and Unlabeled Documents using EM," Machine Learning, vol. 39, no. 2/3, pp. 103-134, 2000.
[12] Jain, M. Murty, and P. Flynn, "Data Clustering: A Review," ACM Computing Surveys, vol. 31, no. 3, pp. 264-323, 1999.

WEB REFERENCES

1. http://msdn.microsoft.com
2. http://www.c-sharpcorner.com
3. http://www.codeproject.com
4. http://www.programmersheaven.com