Fast Mining of Temporal Data Clustering

Dr.D.Suresh Babu, K.Navya
Department of Computer Science and Engineering
Kakatiya Institute of Technology and Science

Abstract: Temporal data clustering provides underpinning techniques for discovering the intrinsic structure and condensing information over temporal data. In this paper, we present a temporal data clustering framework via a weighted clustering produced by initial clustering analysis on different temporal data representations. In the existing system a novel weighted function guided by clustering validation criteria to reconcile initial partitions to candidate consensus partitions from different perspectives, and then, introduce an agreement function to further reconcile those candidate consensus partitions to a final partition.with the rapid growth of text documents, document clustering has become one of the main techniques for organizing large amount of documents into a small number of meaningful clusters. However, there still exist several challenges for document clustering, such as high dimensionality, scalability, accuracy, meaningful cluster labels, overlapping clusters, and extracting semantics from texts. In order to improve the quality of document clustering results, we propose an effective fast mining of temporal data clustering (fmtdc) approach that integrates association mining with an existing wordnet to alleviate these problems finally, each document is dispatched into more than one target cluster by referring to these candidate clusters, and then the highly similar target clusters are merged. The experimental results proved that our approach outperforms the influential document clustering methods with higher accuracy.

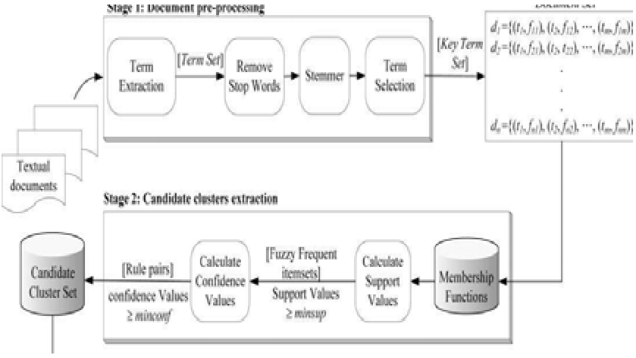
Keywords:temporal document clustering, kb datasets, association rule minining,fuzzy sets,hypernyms,weighted conses function

INTRODUCTION:

Temporal data are ubiquitous in the real world and there are. Unlike static data, there is a high amount of dependency among temporal data and many application areas ranging from multimedia information processing to temporal data mining the proper treatment of data dependency or correlation becomes critical in temporal data processing. Temporal clustering analysis provides an effective way to discover the intrinsic structure and condense information over temporal data by exploring dynamic regularities underlying temporal data in an unsupervised learning way. Its ultimate objective is to partition an unlabeled temporal data set into clusters so that sequences grouped in the same cluster are coherent. In general, there are two core problems in clustering analysis, i.e., model selection and grouping. The former seeks a solution that uncovers the number of intrinsic clusters underlying a temporal data set, while the latter demands a proper grouping rule that groups coherent sequences together to form a cluster matching an underlying distribution.

In our approach, the key terms will be extracted from the document set, and the initial representation of all documents is further enriched by using hypernyms of WordNet in order to exploit the semantic relations between terms. Then, an association mining clustering algorithm for texts is employed to discover a set of highly-related item sets, which contain key terms to be regarded as the labels of the candidate clusters. Finally, each document is dispatched into more than one target cluster by referring to these candidate clusters, and then the highly similar target clusters are merged. We conducted experiments to evaluate the performance based on Classic Web KB datasets. The experimental results proved that our approach outperforms the influential document clustering methods with higher accuracy.

FRAME WORK:


Stage 1: Document pre-processing
Term Extraction
[Term Set]
Remove Stop Words
Stemmer
Term Selection
[Key Term Set]
$d_1=\{(t_{1},f_{11}),(t_{2},f_{12}),\cdots,(t_{m},f_{1m})\}$
$d_2=\{(t_{1},f_{21}),(t_{2},f_{22}),\cdots,(t_{m},f_{2m})\}$
$d_n=\{(t_{1},f_{n1}),(t_{2},f_{n2}),\cdots,(t_{m},f_{nm})\}$
Textual documents
Stage 2: Candidate clusters extraction
Candidate Cluster Set
[Rule pairs]
confidence Values ≥ minconf
Calculate Confidence Values
[Fuzzy Frequent itemsets]
Support Values ≥ minsup
Calculate Support Values
Membership Functions

Weighted Consensus Function

In this module basic weighted consensus function is the use of the pairwise similarity between objects in a partition for evident accumulation, where a pairwise similarity matrix is derived from weighted partitions and weights are determined by measuring the clustering quality with different clustering validation criteria. In the Partition Weighting Scheme X is a data set of N objects and there are M partitions, where the cluster number in M partitions could be different, obtained from the initial clustering analysis. Our partition weighting scheme assigns a weight to each Pm in terms of a clustering validation criterion, and weights for all partitions based on the criterion collectively form a weight vector. After looking into all existing of clustering validation criteria, we select three criteria of complementary nature for generating weights from different perspectives as elucidated Modified Huber's index (MHI), Dunn's Validity Index (DVI). They will be used to weight the similarity matrix, respectively.

Sentence Importance Calculation (K_i) = $\sum_{i=1} T_i \, doc_j$

Also calculate the weight-age of doc_j in sample S_i,

$$I = \sum_{k=1}^{j} \left\{ \sum_{i=1}^{n} T_i \, doc_j \right\}$$

Frequency value of doc_j in sample S_i,
F =doc1/ Number of documents in sample S_i

DOCUMENT ANALYSIS:

Definition 4: The key term set of a document set D={d1,d2,d3,…di,…dn} denoted KD= {t1,t2,t3,..tj,..tm},is a subset of the term set TD, including only meaningful key terms, which do not appear in a well defined stop word list, and satisfy the pre defined minimum tf-idf threshold α, the minimum tf-df threshold β and the minimum tf2 threshold γ. Based on the above definitions the representation of a document can be derived by algorithm 1.

Algorithm 1. Document preprocessing algorithm

```
Input:
1. A document set D={d1,d2,d3….di,..dn}
2. A well defines stop word list.
3. The minimum tf-idf threshold α.
4. The minimum tf-df threshold β.
5. The minimum tf2 threshold γ.
Output: The key term set of D, KD
Method:
Step 1: Extract the term set TD={t1,t2,t3,…tj,..ts}
Step 2. Remove all stop words from TD.
Step 3. Apply word stemming for TD
Step 4. For each di Є D do
For each tj Є TD do
1. Evaluate its tfidfij, tfdfij anf tf2 weights.
2. Retain the term if tfidfij >= α, tfdfij >= β and tf2
ij >= γ.
Step 5. Obtain the key term set KD based on the previous
steps.
Step 6. For each di Є D do
For each tj Є KD do
1. Count its frequency in di to obtain
di={(t1,fi1),(t2,fi2),….(tj,fij),..(tm,fim)}.
```

let us consider one example a document set D={d1,d2,d3,….d10} containing 10 documents. By algorithm 1, we might obtain the derived representation of D and its key term set KD (stock, record, profit,

TERM CONSTRUCTION

The objective of the second module is based on the usage of WordNet for generating a richer document representation of the given document set. As the relationships of relevant terms have been predefined in WordNet, in this module, we intend to use the hypernyms provided by WordNet as useful features for document clustering. After key terms are extracted from the document set, they can be organized based on the hierarchical relationship of WordNet to construct term trees. A term tree is constructed by matching a key term in WordNet and then navigating upwards for five levels of hypernyms. Eventually, all term trees can be regarded as a term forest for the document set D. Using hypernyms can help our approach magnify hidden similarities to identify related topics, which potentially leads to better clustering quality. For example, a document talking about 'sale' may not be associated with a document about 'trade' by the clustering algorithm, if there are only 'sale' and 'trade' in the key term set. But, if a more general term 'commerce' is added to both documents, their semantic relationship can be revealed. Hence, we enriched the representation of each document with hypernyms based on WordNet to find semantically-related documents. Based on the key terms appeared in a document, the representation of this document is enriched by associating them with the term trees accordingly.

Step 1: Extract key terms from the document set
Step 2: Apply WordNet hierarchical relationship of WordNet
Step 3: Construct Term trees by matching a key term in WordNet
Step 4: Navigate upwards for five levels of hypernyms
Step 5: Construct term forest for the document set D

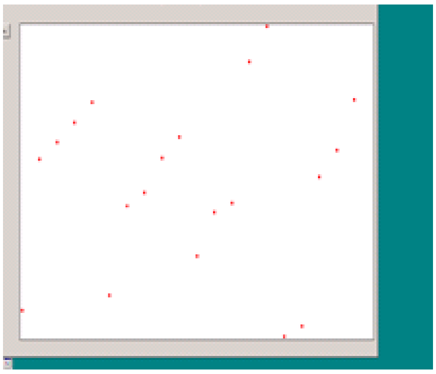
CANDIDATE CLUSTERS EXTRACTION

After the above processes, documents are converted into structured term vectors. Then, the data mining algorithm is executed to generate itemsets and output a candidate clusters set. In the following, we define the membership functions and present our association mining algorithm for texts. The membership functions each pair (tj, fij) of a document di can be transformed into a fuzzy set with its frequency being represented by three fuzzy regions, namely Low, Mid, and High, to depict its grade of membership within di. Each fuzzy value wij has a corresponding membership function, to convert the key tem frequency fij into a value of the range, where

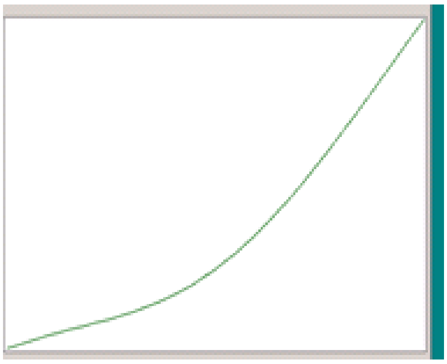
The fuzzy frequent itemsets are generated based on predefined membership functions and the minimum support value θ, from a large textual document set, and obtains a candidate cluster set according to the minimum confidence value λ. Since each discovered fuzzy frequent itemset has an associated fuzzy count value, it can be regarded as the degree of importance that the itemset contributes to the document set.

EXPERIMENT RESULTS:

Clusters of documents without association



After the association rule mining and cluster extraction of documents



CONCLUSION:

Although numerous document clustering methods havebeen extensively studied for many years, the highComputation complexity and space need still make the clustering methods inefficient. Hence, reducing theheavy computational load and increasing the precisionof the unsupervised clustering of documents are important issu hierarchical document clustering approach, based on thefuzzy association rule mining, for alleviating theseproblems satisfactorily.In our approach, we start with the document pre-processing stage; then employ by using the fuzzy association data mining method insecond stage; which generate a candidate cluster set,and merge the high similar clusters. Our experiments show that the accuracy of our algorithm important candidate clusters for document clustering to increase the accuracy quality of documentclustering. Therefore, it is worthy extending in reality for concentrating on huge text documents management.

Our future work focuses on the following two aspects:

1. For improving the performance of the document clustering algorithm, the soft computing approach i.e rough set approach can be applied. Further the algorithm can be improved for higher accuracy by using domain knowledge like WordNet
2. An efficient incremental clustering algorithm can be applied for assigning new document to the most similar existing cluster be proposed as the future direction.

REFERENCES:

[1] Beil, F., Ester, M., & Xu, X. "*Frequent term-based text clustering". In* Proceedings of the 8th ACM SIGKDD int'l conf. on knowledge discovery and data mining(pp. 436–442), 2002.

[2] Chen, C. L., Tseng, Frank S. C., & Liang, T. "*Hierarchical document clustering using fuzzy association rule mining".* In Proceedings of the 3rd international Conference of innovative computing information and control (ICICIC2008) (pp. 326–330), 2008.

[3] Fung, B. C. M., Wang, K., & Ester, M. "*Hierarchical document clustering using frequent itemsets*". Master thesis, Simon Fraser University, 2002.

[4] Fung, B. C. M., Wang, K., & Ester, M. "*Hierarchical document clustering using frequent itemsets".* In Proceedings of the 3th SIAM int'l conf. on data mining(pp. 59–70), 2003.

[5] Shahnaz, F., Berry, M. W., Pauca, V. P., & Plemmons, R. J. "*Document clustering using nonnegative matrix factorization*". Information Processing and Management, 42(2), 373–386, 2006.

[6] Shihab, K. "*Improving clustering performance by using feature selection and extraction techniques".* Journal of Intelligent Systems, 13(3), 135–161, 2004.

[7] Steinbach, M., Karypis, G., & Kumar, V. "*A comparison of document clustering techniques*". In Proceedings of the KDD workshop on text mining, 2000.

[8] Hong, T. P., Lin, K. Y., & Wang, S. L. "*Fuzzy data mining for interesting generalized association rules*". Fuzzy Sets and Systems, 138(2), 255–269,2003.