Document Clustering Using Side Information for Mining Text Data

Kiran D. Gavali
Computer Engineering
Pune University
Pune, India

Saurabh S. Kamthe
Computer Engineering
Pune University
Pune, India

Ganesh B. Pingale
Computer Engineering
Pune University
Pune, India

Prasad U. Vedpathak
Computer Engineering
Pune University
Pune, India

***Abstract*—In many text mining applications, side-information is accessible alongside the text documents. Such side-information could be of various types, like document place of origin info, the links within the document, user-access behavior from internet logs, or other non-textual attributes that are embedded into the text document. Such attributes could contain an incredible quantity of information for cluster functions. However, the relative importance of this side-information is also troublesome to estimate, especially when a number of the knowledge is same. In such cases, it is often risky to include side-information into the text mining method, because it will either improve the standard of the illustration for the mining method, or will add noise. Therefore, we need a principled way to perform the mining method, therefore on maximize the benefits from exploitation this aspect info. In this paper, we design associate algorithmic rule which mixes classical partitioning algorithms with probabilistic models so as to make an efficient clustering approach. We tend to then show a way to extend the approach to the classification drawback. We tend to gift experimental results on a number of real knowledge sets so as the benefits of exploitation such an approach.**

Keywords*— Classification, Text Mining, Side Information, Data mining, Clustering

I. Introduction

The problem of text agglomeration arises within the context of the many application domains like the net, social networks, and alternative digital collections. The increasing amounts of text knowledge within the context of those massive on-line collections has led to Associate in Nursing interest in making ascendable and effective mining algorithms. An amazing quantity of labor has been wiped out recent years on the matter of agglomeration in text collections within the information and data retrieval communities. However, this work is primarily designed for the matter of pure text agglomeration, within the absence of different kinds of attributes. In several application domains, an amazing quantity of aspect info is additionally associated together with the documents. This is as a result of text documents usually occur within the context of a range of applications during which there could also be an oversized quantity of different kinds of information attributes or Meta info which can be helpful to the agglomeration method. Some samples of such side-information are as follows:

The application during which the system tracks user access behavior of internet documents. The access behavior of use could also be captured within the variety of internet logs. For each document, the meta-information corresponds to the browsing behavior of the various users. These logs are to enhance the standard of the mining method that is additional meaty to the user, and conjointly application sensitive. as a result of the logs will usually devour delicate correlations in content, that cannot be picked up by the raw text alone.

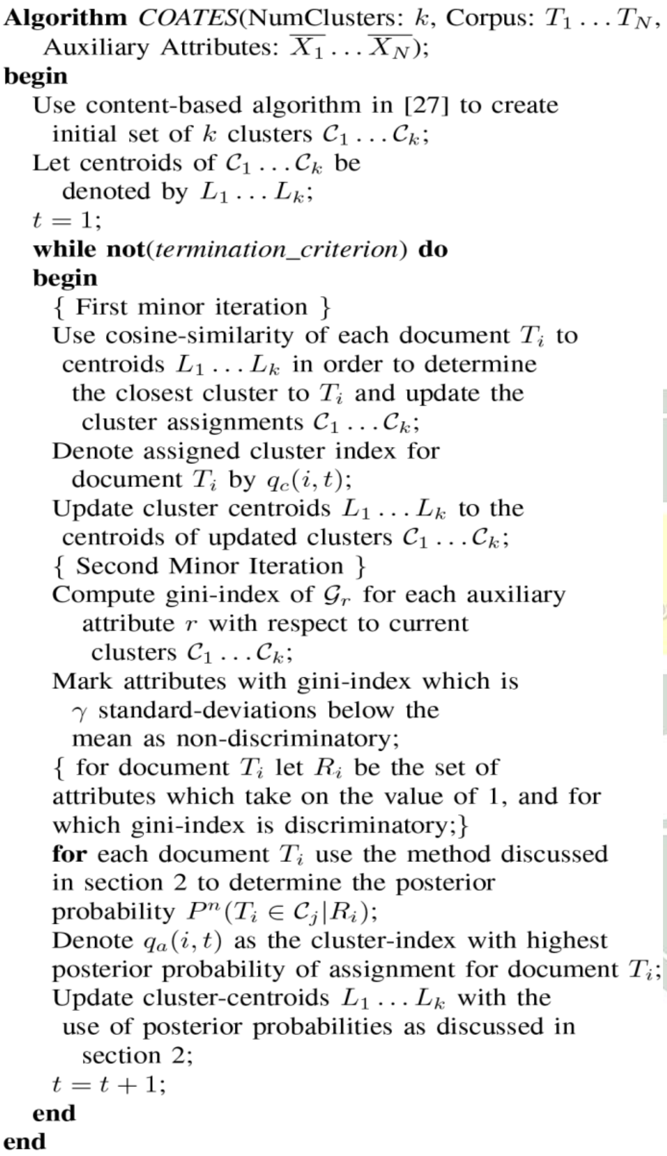
Several text documents contain links in between them, which might even be treated as attributes. These links contain plenty of helpful info for mining functions. Several internet documents have meta-data related to them that correspond to totally different sorts of attributes like the cradle or alternative info like possession, location, or maybe temporal info regarding the origin of the document. in an exceedingly variety of network and user-sharing applications, documents that square measure related to user-tags, may additionally be quite informative. Such side-information is helpful in raising the standard of the agglomeration method. The first goal of this paper is to review the agglomeration of data during which auxiliary information is on the market with text. Such situations square measure quite common in an exceedingly wide variety of knowledge domains.

Therefore, the paper extends the agglomeration approach to the matter of classification, which provides superior results attributable to the incorporation of facet info. The matter of classification has been wide studied within the data processing, info retrieval and info communities with applications in an exceedingly variety of various domains, like target promoting, identification in medical field, filtering of reports teams, and document organization. Text classification finds applications in an exceedingly wide selection of domains in text mining like filtering and organization of reports, document organization and retrieval, sentimental analysis, Email classification and spam filtering etc. Goal of this paper is to point out that the benefits of victimization side-information extend on the far side a pure agglomeration task, and might offer competitive benefits for a wider style of downside situations.

II. Clustering Technique For Side Information

The focus of this paper is to indicate the benefits of mistreatment side-information for mining text information extend beyond a pure bunch task that provides competitive advantages for a wider style of downside eventualities. The formula used for bunch of aspect information is COATES formula that corresponds to the fact that it's Content and Auxiliary attribute based mostly Text bunch formula.

The below fig shows Coates algorithm:



```
Algorithm COATES(NumClusters: $k$, Corpus: $T_1 \ldots T_N$,
  Auxiliary Attributes: $\overline{X_1} \ldots \overline{X_N}$);
begin
  Use content-based algorithm in [27] to create
    initial set of $k$ clusters $\mathcal{C}_1 \ldots \mathcal{C}_k$;
  Let centroids of $\mathcal{C}_1 \ldots \mathcal{C}_k$ be
    denoted by $L_1 \ldots L_k$;
  $t = 1$;
  while not(termination_criterion) do
  begin
    { First minor iteration }
    Use cosine-similarity of each document $T_i$ to
      centroids $L_1 \ldots L_k$ in order to determine
      the closest cluster to $T_i$ and update the
      cluster assignments $\mathcal{C}_1 \ldots \mathcal{C}_k$;
    Denote assigned cluster index for
      document $T_i$ by $q_c(i,t)$;
    Update cluster centroids $L_1 \ldots L_k$ to the
      centroids of updated clusters $\mathcal{C}_1 \ldots \mathcal{C}_k$;
    { Second Minor Iteration }
    Compute gini-index of $\mathcal{G}_r$ for each auxiliary
      attribute $r$ with respect to current
      clusters $\mathcal{C}_1 \ldots \mathcal{C}_k$;
    Mark attributes with gini-index which is
      $\gamma$ standard-deviations below the
      mean as non-discriminatory;
    { for document $T_i$ let $R_i$ be the set of
    attributes which take on the value of 1, and for
    which gini-index is discriminatory;}
    for each document $T_i$ use the method discussed
    in section 2 to determine the posterior
    probability $P^n(T_i \in \mathcal{C}_j | R_i)$;
    Denote $q_a(i,t)$ as the cluster-index with highest
    posterior probability of assignment for document $T_i$;
    Update cluster-centroids $L_1 \ldots L_k$ with the
      use of posterior probabilities as discussed in
      section 2;
    $t = t + 1$;
  end
end
```

The formula needs 2 phases:

A. Data format

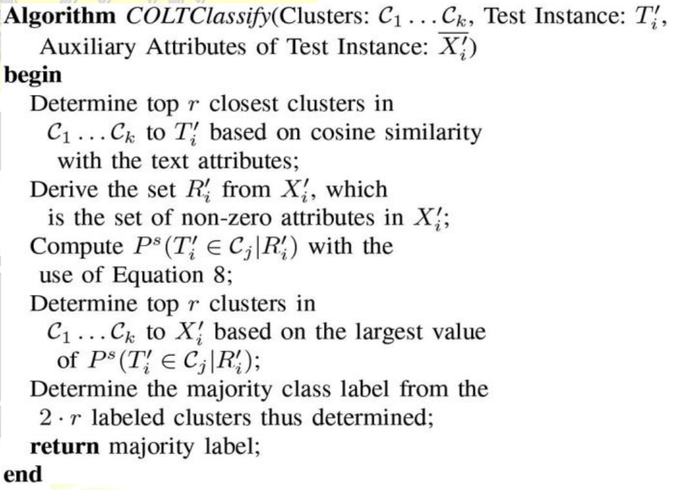
It's a light-weight data format introduce that a standard text bunch approach is employed with none side-information. For this purpose, it uses the k-means clustering formula. The rationale that this formula is employed, it is a straightforward formula which may quickly and expeditiously provide an inexpensive initial place to begin. The partitioning and the centroids created by the clusters fashioned within the 1st phase offer associate degree initial place to begin for the second part. The first part relies on text info solely, not the auxiliary info.

B. Main Phase

This part starts off with these initial teams, and iteratively reconstructs these clusters with the utilization of each the text content and also the auxiliary info. Alternating iterations that use the text content and auxiliary attribute information so as to enhance the standard of the bunch are performed during this step. These iterations are content iterations and auxiliary iterations severally. The combination of the content iteration and auxiliary iteration is stated as a significant iteration. Every major iteration contains 2 minor iterations that like the auxiliary and text-based ways severally.

III. Classification Based On Clusiering Of Side Information

In this section, discussion of a way to extend the approach to classification is completed. This paper can extend the sooner bunch approach so as to include management, and it creates a model that summarizes the category distribution within the knowledge in terms of the clusters. Then, it'll show a way to use the summarized model for effective classification. For extension of classification to the matter of bunch, COLT algorithmic rule is employed for classification of facet info that refers to the very fact that it's a Content and auxiliary attribute-based Text classification algorithmic rule. This algorithmic rule uses a supervised bunch approach so as to partition the info into completely different clusters. This partitioning is then used for the needs of classification. The algorithmic rule works in three steps.



```
Algorithm COLTClassify(Clusters: $\mathcal{C}_1 \ldots \mathcal{C}_k$, Test Instance: $T_i'$,
  Auxiliary Attributes of Test Instance: $\overline{X_i'}$)
begin
  Determine top $r$ closest clusters in
    $\mathcal{C}_1 \ldots \mathcal{C}_k$ to $T_i'$ based on cosine similarity
      with the text attributes;
  Derive the set $R_i'$ from $X_i'$, which
    is the set of non-zero attributes in $X_i'$;
  Compute $P^s(T_i' \in \mathcal{C}_j | R_i')$ with the
    use of Equation 8;
  Determine top $r$ clusters in
    $\mathcal{C}_1 \ldots \mathcal{C}_k$ to $X_i'$ based on the largest value
      of $P^s(T_i' \in \mathcal{C}_j | R_i')$;
  Determine the majority class label from the
    $2 \cdot r$ labeled clusters thus determined;
  return majority label;
end
```

IV. Experimental Results

In this section, we tend to compare our bunch and classification methods against variety of baseline techniques on real and artificial knowledge sets. We tend to consult with our bunch approach as Content and Auxiliary attribute based mostly Text clustering (COATES). Because the baseline, we tend to used 2 completely different methods: (1) Associate degree economical projection based mostly bunch approach that adapts the k-means approach to text. This approach is wide far-famed to supply glorious bunch results in a awfully economical method. We tend to consult with these algorithms as SchutzeSilverstein [text only] all told figure legends in the experimental section. (2) We tend to adapt the k-means approach with the utilization of each text and aspect info directly. We tend to consult with this baseline as K-Means [text+side]. For the case of the classification downside, we tend to test the COLT ways against the subsequent baseline methods: (1) We tested against a Naive mathematician Classifier that uses solely text. (2) We tend to tested against associate degree SVM classifier that uses solely text. (3) We tend to tested against a supervised bunch technique which uses each text and aspect info. Thus, we tend to compare our algorithms with baselines that are chosen in such the way that we are able to value the advantage of our approach over each a pure text-mining technique and a natural different that uses each text and side-information. In order to adapt the k-means approach to the case wherever each text and side-information is employed, the auxiliary attributes were merely used as text-content within the type of “pseudo-words” within the assortment. This makes it comparatively simple to switch the k-means formula thereto case. We will show that our approach has important benefits for both the bunch and classification issues.

A. *Formatting*

During this step, it uses a supervised k means that approach so as to perform the formatting, with the employment of strictly text content. The category memberships of the records in every cluster are pure for the case of supervised formatting. Thus, the k-means bunch algorithmic rule is changed, in order that every cluster solely contains records of a specific category.

B. *Cluster-Training Model Construction*

During this section, a mixture of the text and side-information is employed for the needs of making a cluster primarily based model. As within the case of formatting, the purity of the clusters in maintained throughout this section.

Conclusion

In this paper, we tend to bestowed ways for mining text knowledge with the utilization of side-information. Several varieties of text database contain an outsized quantity of side-information or meta-information, which can be utilized in order to boost the agglomeration method. So as to style the agglomeration method, we tend to combined associate degree repetitive partitioning technique with a likelihood estimation method that computes the importance of various forms of side-information. This general approach is employed so as to style each agglomeration and classification algorithms. We tend to gift results on real data sets illustrating the effectiveness of our approach. The results show that the utilization of side-information will greatly enhance the standard of text agglomeration and classification, while maintaining a high level of potency.

Acknowledgment

We are glad to express our sentiments of gratitude to all who rendered their valuable guidance to us. We would like to express our appreciation and thanks to Dr. D. M. Yadav, Principal, JSPM’s Rajashri Shahu School of Engg. & Research, Pune. We are also thankful to Prof. R. H. Kulkarni, Head of Department, Computer Engg., JSPM’s Rajashri Shahu School of Engg. & Research, Pune. We would like to thank our teachers for their assistance and support. Further we would like to acknowledge all of those friends who helped us gathering dataset. We thank the anonymous reviewers for their comments.

References

[1] C. C. Aggarwal and H. Wang, Managing and Mining Graph Data. New York, NY, USA: Springer, 2010.

[2] C. C. Aggarwal, Social Network Data Analytics. New York, NY, USA: Springer, 2011.

[3] C. C. Aggarwal and C.-X. Zhai, Mining Text Data. New York, NY, USA: Springer, 2012.

[4] C. C. Aggarwal and C.-X. Zhai, “A survey of text classification algorithms,” in Mining Text Data. New York, NY, USA: Springer, 2012.

[5] C. C. Aggarwal and P. S. Yu, “A framework for clustering massive text and categorical data streams,” in Proc. SIAM Conf. Data Mining, 2006, pp. 477–481.

[6] C. C. Aggarwal, S. C. Gates, and P. S. Yu, “On using partial supervision for text categorization,” IEEE Trans. Knowl. Data Eng., vol. 16, no. 2, pp. 245–255, Feb. 2004.

[7] C. C. Aggarwal and P. S. Yu, “On text clustering with side information,” in Proc. IEEE ICDE Conf., Washington, DC, USA, 2012.

[8] R. Angelova and S. Siersdorfer, “A neighborhood-based approach for clustering of linked document collections,” in Proc. CIKM Conf., New York, NY, USA, 2006, pp. 778–779.

[9] A. Banerjee and S. Basu, “Topic models over text streams: A study of batch and online unsupervised learning,” in Proc. SDM Conf. 2007, pp. 437–442.

[10] J. Chang and D. Blei, “Relational topic models for document networks,” in Proc. AISTASIS, Clearwater, FL, USA, 2009, pp. 81–88.

[11] D. Cutting, D. Karger, J. Pedersen, and J. Tukey, “Scatter/Gather: A cluster-based approach to browsing large document collections,” in Proc. ACM SIGIR Conf., New York, NY, USA, 1992, pp. 318–329.