User-interest-based document filtering via semi-supervised clustering

Na Tang and V. Rao Vemuri

Computer Science Dept.
University of California, Davis
Davis, CA 95616, USA
{natang, rvemuri}@ucdavis.edu

Abstract. This paper studies the task of user-interest-based document filtering, where users target to find some documents of a specific topic among a large document collection. This is usually done by a text categorization process, which divides all the documents into two categorizes: one containing all the desired documents (called positive documents) and the other containing all the other documents (called negative documents). However, in many cases, some documents among the negative documents are close enough to the positive documents, prompting a re-consideration (called *deviating negative* documents). Simply treating them as negative documents would deteriorate the categorization accuracy. We modify and extend a semi-supervised clustering method to conduct the categorization. Compared to the original method, our approach incorporates more informative initialization and constraints and in a result leads to better clustering results. The experiments show that our approach retrieves better (sometimes significantly improved) categorization accuracy than the original method in the presence of the *deviating negative* documents.

1 Introduction

A document filtering process is required in many information systems and applications (e.g. [11, 8, 2]), where users target to find some documents of a specific topic that they are interested in from a large document collection. This task of user-interest-based document filtering usually divides the document collection into two categories: positive documents and negative documents, where the former are those that users are interested in and the latter are those that users are not interested in. With increasingly maturing information retrieval and text mining techniques, this task of document categorization can be automated to a certain degree. The task can be done by representing the document collection in a vector space and then applying some learning algorithm to the vector space. The learning algorithms used for document categorization are usually divided into supervised learning, unsupervised learning and semi-supervised learning. Document categorization based on supervised learning is usually called document classification, in which case the model constructed from a set of labeled documents categorizes new unlabeled documents. Document categorization based

on unsupervised learning is usually called document clustering, in which case no labeled documents are available and the model categorizes all the unlabeled documents based on some clustering technique. Usually document clustering does not give as good categorization accuracy as document classification but it saves the effort of manual labeling. Semi-supervised document categorization, a case in which only limited labeled documents are available, provides a compromised solution; it requires some small effort in labeling, but still obtains good categorization results.

In realistic situations, quite often people show medium interest in some documents. These documents are hard to be categorized as positive or negative. Strictly speaking, they belong to the negative documents because they are not exactly what users look for. But they stand closer to the positive documents than to the other negative documents. For example, if users are looking for some documents talking about how to play tennis, then the documents that are not related to tennis definitely belong to negative documents. We call this kind of documents *pure negative* documents. During the searching process they might also get the documents that explain the history of tennis. These documents are also negative documents but they are biased toward the positive documents in some degree. We call this kind of documents *deviating negative* documents. It would deteriorate the categorization accuracy if they are simply considered as negative documents. In this paper, we propose a semi-supervised document clustering method to deal with this issue of borderline documents, namely *deviating negative* documents. Our approach performs a user-interest-based document retrieval task in the presence of *deviating negative* documents by modifying the semi-supervised clustering approach in [3].

The semi-supervised method in [3] made improvements to the standard K-means clustering by incorporating user supervision to the initialization process and the distance measure based on a probabilistic framework. Like [3], the proposed approach is also based on a probabilistic framework while taking advantage of more informative labeled data. Basically the documents are still divided into two classes, positive and negative, but in this paper we recognize the *deviating negative* documents among the negative documents while labeling documents. The initial cluster centroids are estimated from the labeled data. The cluster centroid of the negative documents is estimated from both the pure negative documents and *deviating negative* documents with a bias toward the pure negative documents. While assigning the instances to the clusters at each iteration, it is sensitive to the constraints provided by the more informative labeled data. In addition, it applies adaptive distance learning to be aware of the constraints and at the same time incorporate data variance. The experiments show that our approach is able to deal with the case when *deviating negative* documents are present. Compared to the semi-supervised clustering which does not recognize *deviating negative* documents, the proposed approach increases the categorization accuracy.

The rest of the paper is organized as follows: Section 2 describes the semi-supervised algorithm for the user-interest-based document filtering task. The

experimental results are shown at section 3. The related work is described section 4 and we conclude and discuss the future work at section 5.

2 Algorithm/Framework

This section explains how the standard K-means algorithm partitions documents into two clusters: the one containing positive documents and the one containing negative documents. And it then introduces how we incorporate a limited amount of labeled data into the clustering procedure while being aware of the *deviating negative* documents.

2.1 Standard K-means for Document Partition

K-means clustering can best be described as a partitioning method, which partitions N data points into K mutually exclusive clusters that minimize the total distance between the data points and their cluster centroids. For the user-interest-based document filtering task, each document is treated as a data point; only two clusters are expected to be generated among the documents: C_1 (the cluster with positive documents) and C_{-1} (the cluster with negative documents). These 2-class clustering procedure via K-means can be described as: *1)* Randomly partition the documents into two clusters C_1 and C_{-1}; Estimate the centroids. *2)* For each document xi, calculate the distance (named D_{ij}) from x_i to each cluster centroid of C_j. If the x_i is closest to its own cluster, do nothing; otherwise, move it into the closest cluster. *3)* Re-estimate both the cluster centroids. *4)* Repeat *2)* and *3)* until no documents move from one cluster to another.

2.2 Semi-supervised Clustering with *Deviating Negative* Documents

The proposed approach incorporates labeled data into the K-means clustering framework to improve the clustering results. In this paper, the labeled data consists of the positive documents and the negative documents, where the latter includes the pure negative documents as well as the *deviating negative* documents. Based on these labeled documents, the clustering process is improved through three aspects: 1) initialization; 2) constraint-sensitive distance measure and 3) adaptive distance learning. These three improvements are explained in the following subsections. In the rest of the paper, the following notations are used: documents $\{x_i\}_{i\in P}$ are the labeled positive documents, documents $\{x_i\}_{i\in DN}$ are the *deviating negative* documents, documents $\{x_i\}_{i\in PN}$ are the pure negative documents and documents $\{x_i\}_{i\in U}$ are unlabeled documents. Here P, DN, PN, U are four disjoint subsets of $\{1, ..., N\}$ and $P+DN+PN+U = \{1, ..., N\}$. The function $l(x_i)$ stands for the label of document x_i, where

$$l(x_i) = \begin{cases} 1, & i \in P \\ 0, & i \in DN \\ -1, & i \in PN \\ unknown, & i \in U \end{cases}$$

Initialization The existence of the labeled data can provide prior information about the cluster distribution at the initial time and often results in good clustering. Therefore, instead of randomly initializing cluster centroids (section 2.1 step 1), the proposed approach estimates the cluster centroids from the limited labeled data. The cluster centroid of the positive documents is initialized with the mean of $\{x_i\}_{i \in P}$: $\frac{1}{|P|} \sum_{i \in P} x_i$. Because the topic of the *deviating negative* documents is close to that of the positive documents to some extent, if the cluster centroid is set to be the mean of all the negative documents, then the cluster centroid would be dragged toward that of the positive documents. Therefore, the proposed method initializes the cluster centroid of the negative documents with a weighted mean of the pure negative documents and the *deviating negative* documents: $\frac{1}{w_1 \cdot |DN| + w_2 \cdot |PN|} (w_1 \sum_{i \in DN} x_i + w_2 \sum_{i \in PN} x_i)$ where $w_1 < w_2$, which makes the cluster centroid of the negative documents biased toward the pure negative documents.

Constraint-sensitive Distance Measure The proposed approach enforces constraints that are induced by the labeled documents into the clustering procedure. As it is explained in section 2.1, the K-means algorithm for our document partition task aims to find the document clusters that minimize the overall distance of the documents from the cluster centroids. It modifies the distance measure so that the incorrect assignment of any labeled document x_i to cluster C_j ($j = \pm 1$) results in a certain degree of penalty, i.e., some increase in the distance of x_i from the centroid of C_j. By considering the similarity between the *deviating negative* documents and the positive documents, the proposed method weights the constraints so that the incorrect assignment of the *deviating negative* documents (i.e. assigning the *deviating negative* documents to the cluster of the positive documents) result in lighter penalty than the incorrect assignment of the pure negative documents. As it is mentioned in the original K-means algorithm, D_{ij} is the distance of a document x_i from the cluster centroid of C_j. The K-means assigns document x_i to cluster C_j with the minimum D_{ij} for any C_j. In the proposed method, instead of the pure distance D_{ij}, each document x_i is assigned to C_j to minimize the distortion NEW_D_{ij}, which is defined as:

$$NEW\_D_{ij} = D_{ij} + D_{ij} \cdot penalty(x_i, C_j),$$

where the penalty function is:

$$penalty(x_i, C_j) = \begin{cases} 0, & \text{if } l(x_i) = unknown \text{ || if } j = l(x_i) \text{ || } (j = -1 \text{ \&\& } l(x_i) = 0) \\ p_1, & \text{if } j = 1 \text{ \&\& } l(x_i) = 0 \\ p_2, & otherwise \end{cases}$$

Here the constants satisfy the condition $p_1 < p_2$. The iterated conditional modes (ICM), applied in [3] to find the optimal assignment based on the distance measure, is not used in this paper because the exact label of the documents under supervision are known while only pairwise constraints (must-link and cannot-link) are provided in [3]. Because of the same reason, the constraints are enforced

in the clustering procedure in a simpler way than [3, 12]. Furthermore, the weight function is sensitive to the distance of the point from the cluster centroid and it also provides lighter penalty (p_1) for the incorrect assignment of the *deviating negative* documents, in which case $j = 1$ and $l(x_i) = 0$.

In general, the constraint-sensitive distance measure discourages constraint violations while being aware of the real distance between points. In addition, the penalty of violations by *deviating negative* documents is differentiated from the penalty of violations by pure negative documents by taking into account the topic closeness between *deviating negative* documents and positive documents.

Adaptive Distance Learning The pure distance D_{ij} from document x_i to cluster C_j can be estimated from any distance measure such as Euclidean distance, Cosine distance, I-divergence and so on. However, instead of using the static distance, which may fails to capture the real notion of distance in a clustering procedure, parameterized distance measures are used to incorporate the user-specified constraints and data variance.

One of the commonly used distance measure - Euclidean distance - is parameterized in this paper. Suppose the centroid of cluster C_j is c_j, the pure Euclidean distance is defined as:

$$D_{ij} = \sqrt{(x_i - c_j)^T (x_i - c_j)}$$

Then the parameterized Euclidean distance is defined as follows:

$$D_{ij}^A = \sqrt{(x_i - c_j)^T \cdot A \cdot (x_i - c_j)},$$

where A is a positive diagonal matrix. Therefore, the final distortion NEW_D_{ij}, which the clustering process tries to minimize for each document x_i, is parameterized as:

$$\text{NEW\_}D_{ij}^A = \sqrt{(x_i - c_j)^T \cdot A \cdot (x_i - c_j)},$$

The parameter matrix A is first initialized with an identity matrix and then updated at each iteration after the cluster centroids are re-estimated. The updating rule is:

$$\begin{aligned} a_k &= a_k + \frac{\partial New\_D}{\partial a_k} \\ &= a_k + (\sum_{i=1}^{N} \frac{\partial D_{i,assigned\_l(x_i)}}{\partial a_k} + penalty(x_i, assigned\_l(x_i)) \cdot \sum_{i=1}^{N} \frac{\partial D_{i,assigned\_l(x_i)}}{\partial a_k}), \end{aligned}$$

where $assigned\_l(x_i)$ stands for the assigned label for document x_i at the current iteration and

$$\frac{\partial D_{i,assigned\_l(x_i)}}{\partial a_k} = \frac{x_{ik} c_{assigned\_l(x_i),k}}{2\sqrt{(x_i - c_j)^T \cdot A \cdot (x_i - c_j)}}.$$

In essence, the adaptive distance learning brings similar documents closer and pushes dissimilar documents further apart. In this way more cohesive clusters are generated, which facilitate the partitioning process.

As a whole, combined with these three improvements, the proposed algorithm is summarized in the following chart:

Semi-supervised Clustering with Deviating Negative Documents

Input: Set of documents $\{x_i\}_{i=1}^{N}$, index of labeled positive documents, deviating negative documents and pure negative documents respectively: *P*, *DN*, *PN*

Output: Disjoint 2-partitioning of $\{x_i\}_{i=1}^{N}$

1. Initialize centroids of clusters C_1 and C_{-1} with $\frac{1}{|P|}\sum_{i\in P} x_i$ and $\frac{1}{w_1 \cdot |DN| + w_2 \cdot |PN|}(w_1 \sum_{i\in DN} x_i + w_2 \sum_{i\in PN} x_i)$ respectively.
2. For each $i \in \{1,...N\}$, calculate the parameterized distance from document *i* to cluster C_j, i.e., $New\_D_{ij}^{A}$. If document *i* is closest to its own cluster, do nothing; otherwise, move it into the closest cluster.
3. Re-estimate the cluster centroids with $c_j = \frac{1}{|C_j|}\sum_{x_i \in C_j} x_i$; Update parameter matrix *A*.
4. Repeat 2 & 3 until no documents moving from one cluster to another.

3 Experiments

Experiments are conducted on the Syskill and Webert Web Page Ratings (SW) [9], the 20 Newsgroup data set (20NG) [1] and the heart-disease webpage set (HD) from [11]. SW contains four data sets of HTML pages relating to four different topics. A user rated each page in a 3-point scale (hot, median and cold) which indicates his interest in that page. We select 3 sets from SW (*bands*, *biomedical*, *goats*) and treat the "hot" documents as positive documents, "median" documents as *deviating negative* documents, "cold" documents as pure negative documents. The 20NG data set contains about 20,000 documents on different subjects from 20 UseNet discussion groups. We select 2 subsets from 20NG: *ibm_gra_mac* and *ibm_x_mac*. The subset *ibm_gra_mac* contains 600 documents, 200 randomly selected from the group *comp.sys.ibm.hardware* as positive documents, 200 randomly selected from the group *comp.sys.mac.hardware* as *deviating negative* documents and 200 randomly selected from the group *comp.graphics* as pure negative documents. The subset *ibm_x_mac* also contains 600 documents. The 200 positive and 200 *deviating negative* documents are randomly selected from the same groups as *ibm_gra_mac* while the 200 pure negative documents are randomly selected from the group *comp.windows.x*. HD contains 288 HTML pages that are divided into positive documents, *deviating negative* documents and *pure negative* documents based on a user's interest. For each data set, 10%

of documents are chosen as the labeled data, and the remaining as the unlabeled data.

The preprocessing includes document representation and feature selection. We represent each document as a vector via a TF-IDF model (Term Frequency - Inverse Document Frequency). With the TF-IDF vector representation, each jth item of the vector i (representing document i) is determined by the number of times that it appears in document i (TF) as well as the number of documents that this word appears (IDF). The dimension of the vectors is decided by the vocabulary size, which tends to be large even with a small set of documents. Instead of using all the words, a smaller number of best words can be selected for further clustering. This can lead to significant savings of computer resources and processing time. It is called feature selection because each word is considered as a feature for clustering. In the proposed method, the following equation (see [7]) is used for feature selection. It evaluates the quality of a word w:

$$q(w) = \sum_{i=1}^{N} f_i^2 - \frac{1}{N}\left[\sum_{i=1}^{N} f_i\right]^2 .$$

Here f_i is the frequency of word w in document d_i and N is the total number of documents. In our experiments, the dimension of the vectors is set to be 128.

The pair of weights for initialization (w_1,w_2) prevents the centroid of negative documents from biased toward positive documents, while the pair of penalties for constraint-sensitive distance (p_1, p_2) provides more sensitive constraints with the presence of *deviating negative* documents. The selection of the initialization parameters, namely the weights and penalties, is a problem that is yet to be addressed. Like the learning parameter in many machine learning methods, these values are likely to be problem dependent. Some local search is probably involved. For expediency, both (w_1,w_2) and (p_1,p_2) are set at (0.5,1) in this study, which satisfies $w_1 < w_2$ and $p_1 < p_2$.

The categorization accuracies with different document sets and different methods are shown at Table 1. The method "K-means_3C" is the standard K-means algorithm by treating *deviating negative* documents as a separated class (totally 3 classes) while the method "K-means_2C" is the standard K-means algorithm with only 2 classes: the positive class and the negative class. Similarly, the method "semi_K-means_3C" and "semi_K-means_2C" is the semi-supervised K-means approach presented in [3] with 3 classes and 2 classes respectively. The method "semi_K-means_DND" is the proposed approach, i.e., the semi-supervised K-means by assuming 2 classes while considering *deviating negative* documents (DND) during clustering procedure.

The experimental results show that both "K-means" and "semi_K-means" with 2 classes offer better performance than those with 3 classes (by treating *deviating negative* documents as a separated class). The reason is that the topic of *deviating negative* documents are not totally separated from either *pure negative* documents or *positive* documents, which confuses the 3-class clustering. The results also show that, among all the approaches with 2 classes, our approach retrieves the best categorization accuracies over all the document sets. It indicates

Table 1. A Comparison of categorization accuracies with different methods and document set.

	bands	biomedical	goats	ibm_gra_mac	ibm_x_mac	HD
K-means_3C	0.557	0.588	0.500	0.453	0.521	0.535
semi_K-means_3C	0.574	0.611	0.557	0.588	0.640	0.552
K-means_2C	0.656	0.595	0.500	0.652	0.548	0.563
semi_K-means_2C	0.672	0.687	0.529	0.731	0.598	0.689
semi_K-means_DND	**0.754**	**0.702**	**0.571**	**0.740**	**0.688**	**0.693**

that the 2-class clustering and the additional semi-supervision on the *deviating negative* documents, which are the documents similar to positive documents but not exactly what users want, is able to give more informative constraints and in a result lead to better clustering accuracy for filtering purpose.

4 Related Work

Some other semi-supervised clustering algorithms [3, 4, 6, 12] and semi-supervised classification algorithms [8, 10, 5] are available and can be applied to the user-interest-based document filtering task. Basically semi-supervised clustering incorporate limited labeled data to guide the clustering process while semi-supervised classification uses unlabeled data to improve classification. However, none of them consider the issue of *deviating negative* documents. Ignorance of this kind of documents may deteriorate the categorization results. This paper deals with this issue under a limited amount of user supervision.

5 Conclusion and Discussion

This paper presented a semi-supervised clustering approach to user-interest-based document filtering. It modifies a semi-supervised clustering algorithm in [3] in order to be sensitive to user interest especially to the presence of the *deviating negative* documents. This approach was empirically tested with the Syskill and Webert Web Page Ratings, the 20 Newsgroup data set and a webpage set from [11]. The experiments show that our approach retrieves better categorization accuracy than the method in [3] with the presence of the *deviating negative* documents.

A number of interesting problems are left for future research:

1. Labeling data selection: Selecting appropriate documents for labeling would have an influence in the clustering results. We propose to incorporate an active learning algorithm to actively select samples for labeling in the future.

2. Feature selection: The labeled data may have a good insight about the feature selection. By combining the method used in this paper and the information gain technique, which is usually used for classification when adequate labeled data is available, we may get words of better quality for categorization.

3. Incremental documents clustering: The information environments tend to be dynamic and it is desirable to have an adaptive clustering method to deal with continuously growing document set.

4. Other applications: Besides document categorization and document filtering, some other applications involving 2-class classification may also take advantage of the proposed method to deal with the issue of *deviating negative* instances, which are the instances belonging to the negative class but close to the positive instances.

References

1. 20 newsgroup data set. http://people.csail.mit.edu/ jrennie/20Newsgroups, last visited Jan. 19th, 2005.
2. Baldi, P., Frasconi, P. and Smyth, P.: Modeling the Internet and theWeb: Probabilistic Methods and Algorithms. Chapter 4: Text Analysis. Wiley, (2003).
3. Basu, S., Bilenko, M. and Mooney, R.J.: A probabilistic framework for semi-supervised clustering. In Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2004). Seattle, WA, (2004).
4. Basu, S., Banerjee, A. and Mooney, R.J.: Semi-supervised Clustering by Seeding. In Proceedings of the 19th International Conference on Machine Learning (ICML-2002). Sydney, Australia, (2002).
5. Blum, A. and Mitchell, T.: Combining labeled and unlabeled data with co-training. In Proceedings of the 11th Annual Conference on Computational Learning Theory. (1998).
6. Cohn, D., Caruana, R. and McCallum, A.: Semi-supervised clustering with user feedback. Technical Report TR2003-1892, Cornell University, (2003).
7. Dhillon, I., Kogan, J. and Nicholas, C.: Feature Selection and Document Clustering, in Survey of Text Mining. Springer-Verlag, New York. (2004) Chapter 4.
8. Liu, B., Dai, Y., Li, X., Lee, W. S. and Yu, P. S.: Building Text Classifiers Using Positive and Unlabeled Examples. In Proceedings of the Third IEEE International Conference on Data Mining. Melbourne, Florida, (2003).
9. Pazzani, M.: Syskill and Webert Web Page Ratings. http://ncdm171.lac.uic.edu:16080/kdd/databases/SyskillWebert/SyskillWebert.task.html, last visited Jan. 19th, 2005.
10. Nigam, K., McCallum, A. K., Thrun, S. and Mitchell, T.: Text Classification from Labeled and Unlabeled Documents using EM. Machine Learning. Vol. 39. (2000).
11. Tang, N. and Vemuri, V. R.: Web-based Knowledge Acqusition to Impute Missing Values for Classification. In Proceedings of the 2004 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI/IAT-2004). Beijing, China, (2004).
12. Wagstaff, K., Cardie, C., Rogers, S. and Schroedl, S.: Constrained K-Means Clustering with Background Knowledge. In Proceedings of 18th International Conference on Machine Learning (ICML-2001), (2001).