Automatic Distribution of Documents into Different Categories using Active Learning

M. Jayaprakash[1], D. John Aravindhar[2], E. R. Naganathan[3]

[1]School of Computing Sciences, Hindustan University

jp.cse.prakash@gmail.com
jaravindhar@hindustanuniv.ac.in

Abstract: ***Data mining extracts novel and useful knowledge from large repositories of data and has become an effective analysis and decision means in corporation in many information processing tasks, labels are usually expensive and the unlabeled data points are abundant. To reduce the cost on collecting labels, it is crucial to predict which unlabeled examples are the most informative, i.e., improve the classifier the most if they were labeled. Many active learning techniques have been proposed for text categorization, such as SVM Active and Transductive Experimental Design. However, most of previous approaches try to discover the discriminant structure of the data space, whereas the geometrical structure is not well respected. By minimizing the expected error with respect to the optimal classifier, they can select the most representative and discriminative data points for labeling. Experimental results on text categorization have demonstrated the effectiveness of proposed approach.***

Keywords: Text categorization, Novel Active learning, Manifold learning, labeled data

1. Introduction

The active learning techniques have been used for text categorization. Normally, in information processing task, main task is collecting the labels, but collecting labels are very expensive at the same time many unlabeled data points are there. To reduce the cost of collecting the labels unlabeled data points are more informative [7]. So we have to give the more concentration in unlabeled data points [4]. Most of previous approaches try to discover the structure of the data space, whereas the geometrical structure is not well respected. To propose a novel active learning algorithm, this is performed in the data manifold adaptive kernel space. The manifold adaptive kernel space reflects the underlying geometry of the data [9].

2. Literature Survey

2.1 Manifold Regularization

We propose a family of learning algorithms based on a new form of regularization that allows us to exploit the geometry of the marginal distribution. We focus on a semi-supervised framework that incorporates labeled and unlabeled data in a general-purpose learner. Some transductive graph learning algorithms and standard methods including support vector machines and regularized least squares can be obtained as special cases. We use properties of reproducing kernel Hilbert spaces to prove new Representer theorems that provide theoretical basis for the algorithms. As a result (in contrast to purely graph-based approaches) we obtain a natural out-of-sample extension to novel examples and so are able to handle both transductive and truly semi supervised settings. We present experimental evidence suggesting that our semi-supervised algorithms are able to use unlabelled data effectively. Finally we have a brief discussion of unsupervised and fully supervised learning within our general framework.

2.2 Spectral Regression

It considers the problem of document indexing and representation. Recently, Locality Preserving Indexing (LPI) was proposed for learning a compact document subspace. Different from Latent Semantic Indexing (LSI) which is optimal in the sense of global Euclidean structure, LPI is optimal in the sense of local manifold structure. However, LPI is not efficient in time and memory which makes it difficult to be applied to very large data set. Specifically, the computation of LPI involves eigen-decompositions of two dense matrices which is expensive. In this paper, we propose a new algorithm called *Regularized Locality Preserving Indexing* (RLPI). Benefit from recent progresses on spectral graph analysis, we cast the original LPI algorithm into a regression framework which enable us to avoid eigen-decomposition of dense matrices. Also, with the regression based framework, different kinds of regularizers can be naturally incorporated into our algorithm which makes it more flexible. Extensive experimental results show that RLPI obtains similar or better results comparing to LPI and it is significantly faster, which makes it an efficient and effective data preprocessing method for large scale text clustering, classification and retrieval.

2.3 Modeling Hidden Topics on Document Manifold

Topic modeling has been a key problem for document analysis. One of the canonical approaches for topic modeling is Probabilistic Latent Semantic Indexing, which maximizes the joint probability of documents and terms in the corpus. The major disadvantage of PLSI is that it estimates the probability distribution of each document on the hidden topics independently and the number of parameters in the model grows linearly with the size of the corpus, which leads to serious problems with over fitting. Latent Dirichlet Allocation (LDA) is proposed to overcome this problem by treating the probability distribution of each document over topics as a

hidden random variable. Both of these two methods discover the hidden topics in the Euclidean space. The document space is a manifold, either linear or nonlinear. In this paper, we consider the problem of topic modeling on intrinsic document manifold. Specifically, we propose a novel algorithm called Laplacian Probabilistic Latent Semantic Indexing (LapPLSI) for topic modeling. LapPLSI models the document space as a sub manifold embedded in the ambient space and directly performs the topic modeling on this document manifold in question.

2.4 Text Categorization with Support Vector Machines

This paper explores the use of Support Vector Machines (SVMs) for learning text classifiers from examples. It analyzes the particular properties of learning with text data and identities why SVMs are appropriate for this task. Empirical results support the theoretical findings. SVMs achieve substantial improvements over the currently best performing methods and behave robustly over a variety of different learning tasks. Furthermore, they are fully automatic, eliminating the need for manual parameter tuning.

2.5 Query by Committee Made Real

Training a learning algorithm is a costly task. A major goal of active learning is to reduce this cost. In this paper we introduce a new algorithm, KQBC, which is capable of actively learning large scale problems by using selective sampling. The algorithm overcomes the costly sampling step of the well known *Query By Committee* (QBC) algorithm by projecting onto a low dimensional space. KQBC also enables the use of kernels, providing a simple way of extending QBC to the non-linear scenario. Sampling the low dimension space is done using the *hit and run* random walk. We demonstrate the success of this novel algorithm by applying it to both artificial and a real world problems.

3. Existing System

In existing, the collecting labels used to group many documents. SVMActive technique has used in existing, to collects data points. This method is used to reduce the size of the version space as much as possible. It is difficult to measure the version space. In existing, three approximations are used to collects data points. One of them which select the points closest to the current decision boundary is called Simple Margin.

4. Data Categorization

The text Categorization for collecting documents which is related to user's query. We have to find the data space efficiently. Here data space is used to overcome some of the problem encountered in data integration system. The aim is to reduce the effort required to setup a data integration system by relying on existing matching and mapping generation techniques. The proposed algorithm has shown good performance for text categorization on 20Newsgroup, Reuters-21578, and RCV1, especially when only a small number of examples can be labeled. There are several problems that need to be investigated in the future. They have to reduce the high dimensional into low dimensional. Dimension represents the data elements that categories each item in a data set into non-overlapping regions.

5. Manifold Experimental Design

5.1 Over all Manifold Architecture

The following architecture are used in developing the system for collecting documents which is related to user's query. They have to find the data space efficiently.

- User request
- Query Analysis
- Creating labels
- Text classification
- Clustering and retrieving

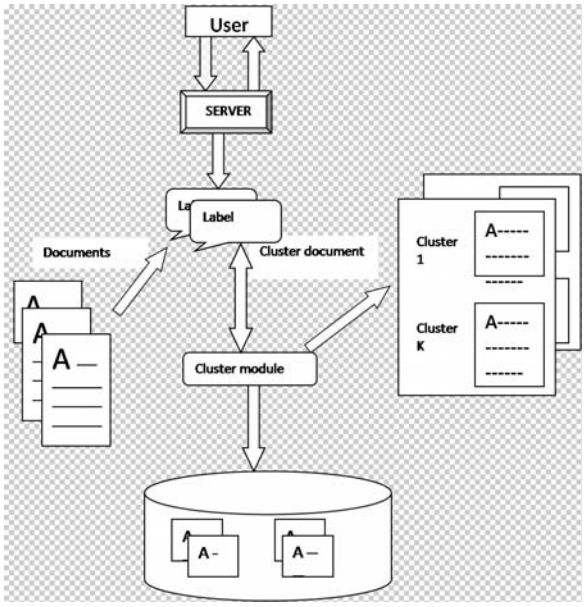

Figure 1: Describes the System Architecture process

5.2 User request

The Input is send to User's query and the output is Validate from server

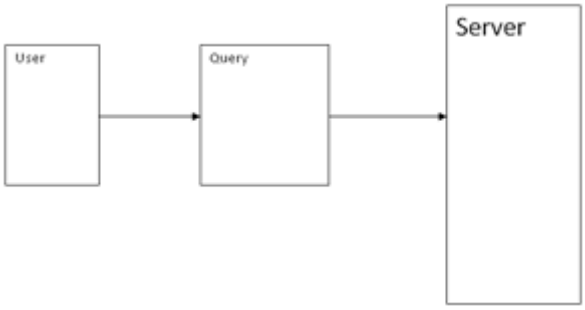

Figure 2: User Request from web content

First module is user request. This module is used to get the query from the user. This module will come after successfully giving the correct username and password by

the user. In this module, we are creating user request window to create this will use Form action to that module. User wants to send a query to particular server, server will receive that particular request which coming from user now server wants to identify that particular from where is request coming? To identify only we are giving form action is user to identify a particular application. Server program is running in one machine and client program is running in one particular machine. Now client send a query this query will be received by server to which User giving queries to get a response which is collect from server's database. So server must have some amount of data then only user can get some response.

5.3 Query Analysis

The Input is Query received from user and the output is Query will be analysis by server

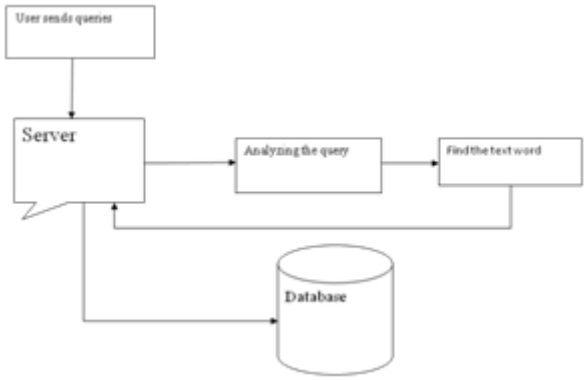

Figure 3: Query Analyzing

In this module query will processed by server. Server will receive query once user sent a query. Once received a query server should do one thing that is have to find the document and send it to user. This is the process of the server. Result will be generated according to the user's query. Server must do some steps once receive the query from server (i.e.) receive the query and search relevant document from the database and take documents and send it to user. Server will be running in one port no. Server will be run then only user can give query to serve. Server must run before the running of the user [4]. So we first run the server the server then run the client. Server split the query into words then searches the document relevant to user's query this is said to be as the semantic searching [5]. This serve must maintain the database then only server can able to store the document and retrieve the documents.

5.4 Creating labels & Text Analysis

The Input is Get the all documents and the output is Label will be generated in third module we are creating labels. Normally labels are used to find the set of data in a group. Each and every label having some amount of data information and some unlabeled data points are abundant. In the project we will create the label for efficient retrieval of data for user's query [4]. Many records will be stored in a group this is said to be label. After creating the label we have to analysis the text then store it into appropriate label. While doing this it will be retrieved by efficiently [9]. Text analysis is main thing in this project. Text will be analysis by labels.

5.5 Text classification

The Input is Split the text and the output is similar document from the classification.

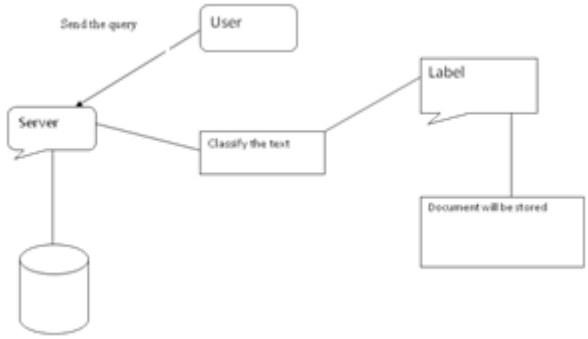

Figure 4: Classification of data

Text classification is next step of the text analysis. In our project, we will be separate the text documents into two sets. One is said to be as training set another one is testing set. We will be storing the records into the database. Each record contains a set of attributes; one of the attributes is the class then fined a model for class attribute as a function of the values of other attributes [5]. Normally we are creating the labels these labels having some amount of records these records should be relevant to our labels. User send the query to server this query will be analyzing and then take main word with the help of this we will easily classify the text.

5.6 Clustering and retrieving

The Input is selected documents from database and the output is send query from taken document to user.

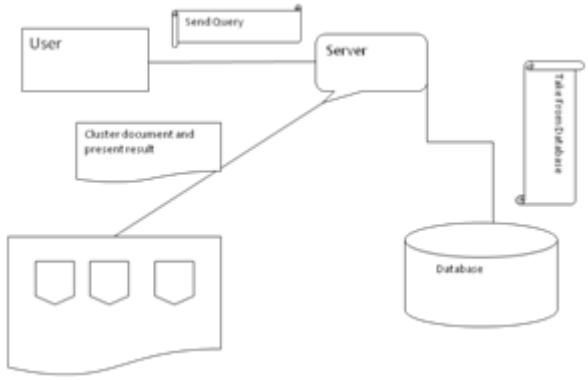

Figure 5: Retrieving the different data

In this module, we are creating the group of documents and store it into database. Here clustering is used to group the number of documents in a single group [5]. Document will be retrieved from the database depending upon the user's query. Once query has received from the user it will be split and then search the documents from the database [9]. While clustering we need to integrate the data. Data integration is combining data residing in

different sources. For example two similar companies need to merge their database.

6. Methodologies

In order to incorporate the manifold structure into the active learning process. To perform active learning tasks in manifold adaptive kernel space. Manifold adaptive experimental design algorithm is used. Data and experimental settings of text categorization to be implemented.
The Frequency and Documents for each associated data be categorized using the below formula.

$$tf - idf = (1 + \log tf) * \log N/df$$

The Term frequency (tf) and Inverse document frequency (idf) are to be search and cluster the documents.

- The manifold structure into the reproducing kernel space.
- In which they leads to manifold adaptive kernel space.
- Kernel trick is usually applied in the hope of discovering the nonlinear structure in the data by mapping the original nonlinear observations into a higher dimensional linear space.

7. Conclusion

In this project the active learning techniques for text categorization. Unlike most of previous active learning approaches which explore either euclidean or data-independent nonlinear structure of the data space, our proposed approach explicitly takes into account the intrinsic manifold structure. The Future enhancement is to use the Document clustering to merge the document to produce the results. The methods are Text (and hypertext) categorization, Image classification. This is used to download documents in separate manner. The Input is Retrieved the document and the output is Download the particular document.

References

[1] R. Angelova and G. Weikum(2006),"Graph-based Text Classification: Learning from your neighbors," Proc. 29th Int'l Conf. Research and Development in Information Retrieval.
[2] A. C. Atkinson and A, N. Donev(2007), Optimum Experimental Designs, with SAS. Oxford Univ. Press
[3] M. Belkin and P. Niyogi (2001), "Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering," Advances in Neural Information Processing Systems, vol. 14, pp.585-591.
[4] M. Belkin, P. Niyogi, and V. Sindhwani (2006), "Manifold Regularization: A Geometric Framework for Learning from Examples," J. Machine Learning Research, vol. 7, pp. 2399-2434.
[5] D. Cai, X. He, W.V. Zhang, and J. Han (2007), "Regularized Locality Preserving Indexing via Spectral Regression," Proc. 16th ACMConf. Information and Knowledge Management (CIKM '07), pp. 741-750.
[6] D. Cai(2008), X. He, X. Wu, and J. Han, "Non-Negative Matrix Factorization on Manifold", Proc. Int'l Conf. Data Mining (ICDM '08)
[7] H. Seung, M. Opper, and H. Sompolinsky (1992), "Query by Committee," Proc. Fifth Ann. Workshop Computational Learning Theory (COLT '92), pp. 287-294.
[8] T. Joachims (1998), "Text Categorization with Support Vector Machines: Learning with Many Relevant Features," Proc. European Conf. Machine Learning (ICML), pp. 137-142.
[9] D. Cai, Q. Mei, J. Han, and C. Zhai(2008), "Modeling Hidden Topics on Document Manifold," Proc. 17th ACM Conf. Information and Knowledge Management (CIKM '08), pp. 911-920

Author Profile

Mr M. Jayaprakash received bachelor's degree (Adhiparasakthi Engineering College) in Computer Science and Engineering from Anna University, Chennai, India in 2008 and doing master's degree in Computer Science and Engineering in Hindustan University, Chennai.