USING HUMAN ANNOTATIONS DISCOVERS EXTERNAL DATA FEATURES INTEGRATION USING SHARING

P.R.N. Prasad [1*], CH. Raja Jacob [2*]

1. M.Tech (CSE) Student, Dept of CSE, Nova College of Engg & Tech, Jangareddygudam, Dist: W.Godavari, AP, India
2. Assoc.Professor, Dept of CSE, Nova College of Engg & Tech, Jangareddygudam, Dist: W.Godavari, AP, India

Keywords:
CASTLE, streaming data, privacy preserving.

Abstract: *In this paper, we present Continuously Anonymizing Streaming data via Adaptive Clustering (CASTLE), a cluster-based scheme that anonymizes data streams on-the-fly and, at the same time, ensures the freshness of the anonymized data by satisfying specified delay constraints. We further show how CASTLE can be easily extended to handle diversity. Our extensive performance study shows that CASTLE is efficient and effective the quality of the output data. Most of existing privacy preserving techniques, such as k-anonymity methods, are designed for static data sets. As such, they cannot be applied to streaming data which are continuous, transient and usually unbounded. Moreover, in streaming applications, there is a need to offer strong guarantees on the maximum allowed delay between an incoming data and its anonymized output.*

1. INTRODUCTION

The World Wide Web has influenced many aspects of our lives, changing the way we communicate, conduct business, shop, entertain, and so on. However, a large portion of the Web data is not organized in systematic and well structured forms, a situation which causes great challenges to those seeking for information on the Web. Consequently, a lot of tasks enabling users to search, navigate and organize web pages in a more effective

way have been posed in the last decade, such as searching, page rank, web clustering, textclassification, etc. To this end, there have been a lot of successful stories like Google, Yahoo, Open Directory Project (Dmoz), Clusty, just to name but a few. Inspired by this trend, the aim of this thesis is to develop efficient systems which are able to overcome the difficulties of dealing with sparse data. The main motivation is that while being overwhelmed by a huge amount of online data, we sometimes lack data to search or learn effectively. Let take web search clustering as an example. In order to meet the real-time condition, that is the response time must be short enough, most of online clustering systems only work with small pieces of text returned from search engines. Unfortunately those pieces are not long and rich enough to build a good clustering system. A similar situation occurs in the case of searching images only based on captions. Because image captions are only very short and sparse chunks of text, most of the current image retrieval systems still fail to achieve high accuracy. As a result, much effort has been made recently to take advantage of external resources like learning with knowledge-base support, semi-supervised learning, etc. in order to improve the accuracy. These approaches, however, have some difficulties: (1) constructing a knowledge base is very time-consuming & labor-intensive, and (2) the results of semi-supervised learning in one application cannot be reused in another one even in the same domain. In the thesis, we introduce two general frameworks for learning with hidden topics discovered from large-scale data collections: one for clustering and another for classification. Unlike semi-supervised learning, we

[*] ***P.R.N. Prasad***
M.Tech (CSE) Student, Dept of CSE, Nova College of Engg & Tech, Jangareddygudam, Dist: W.Godavari, AP, India

approach this issue from the point of view of text/web datasome analysis that is based on recently successful topic analysis models, such as Latent Semantic Analysis, Probabilistic-Latent Semantic Analysis, or Latent Dirichlet Allocation. The underlying idea of the frameworks is that for a domain we collect a very large external data collection called “universal dataset”, and then build the learner on both the original data (like snippets or image captions) and a rich set of hidden topics discovered from the universal data collection. The general frameworks are flexible and general enough to apply for a wide range of domains and languages. Once we analyze a universal dataset, the resulting hidden topics can be used for several learning tasks in the same domain. This is also particularly useful for sparse data mining. Sparse data like snippets returned from a search engine can be expanded and enriched with hidden topics. Thus, a better performance can be achieved. Moreover, because the method can learn with smaller data (the meaningful hidden topics rather than all unlabeled data), it requires less computational resources than semi-supervised learning.

2. THE PROBLEM OF MODELING TEXT CORPORA AND HIDDEN TOPIC ANALYSIS

2.1. Latent Semantic Analysis

The main challenge of machine learning systems is to determine the distinction between the lexical level of “what actually has been said or written” and the semantic level of “what is intended” or “what was referred to” in the text or utterance. This problem lies in twofold: (i) polysemy, i.e., a word has multiple meaning and multiple types of usage in different context, and (ii), synonymy and semantically related words, i.e, different words mat have similar sense. They at least in certain context specify the same concept or the same topic in a weaker sense. Latent semantic analysis (LSA - Deerwester et al, 1990) [13][24][26] is the well-known technique which partially addresses this problem. The key idea is to map from the document vectors in word space to a lower dimensional representation in the so-called concept space or latent semantic space. Mathematically, LSA relies on singular value decomposition (SVD), a well-known factorization method in linear algebra.

a. Latent Semantic Analysis by SVD In the first step, we present the text corpus as term-by-document matrix where elements (i, j) describes the occurrences of term i in document j. Let X be such a matrix, X .

b. Applications

The new concept space typically can be used to:

-Compare the documents in the latent semantic space. This is useful to some typical learning tasks such as data clustering or document classification.

- Find similar documents across languages, after analyzing a base set of translated documents.

-Find relations between terms (synonymy and polysemy). Synonymy and polysemy are fundamental problems in natural language processing:

- Synonymy is the phenomenon where different words describe the same

idea. Thus, a query in a search engine may fail to retrieve a relevant

document that does not contain the words which appeared in the query.

- Polysemy is the phenomenon where the same word has multiple meanings.

So a search may retrieve irrelevant documents containing the desired words in the wrong meaning. For example, a botanist and a computer scientist looking for the word "tree" probably desire different sets of documents.

-Given a query of terms, we could translate it into the concept space, and find matching documents (information retrieval).

c. Limitations

LSA has two drawbacks:

-The resulting dimensions might be difficult to interpret. For instance, in

{(car), (truck), (flower)} --> {(1.3452 * car + 0.2828 * truck), (flower)}

the (1.3452 * car + 0.2828 * truck) component could be interpreted as "vehicle".

However, it is very likely that cases close to

{(car), (bottle), (flower)} --> {(1.3452 * car + 0.2828 * bottle), (flower)} will occur. This leads to results which can be justified on the mathematical level,

but have no interpretable meaning in natural language.

The probabilistic model of LSA does not match observed data: LSA assumes that

words and documents form a joint Gaussian model (ergodic hypothesis), while a

Poisson distribution has been observed. Thus, a newer alternative is probabilistic

latent semantic analysis, based on a multinomial model, which is reported to give

better results than standard LSA.

3. PROBABILISTIC LATENT SEMANTIC SPACE

Let us consider topic-conditional multinomial distribution over vocabulary as points on the)|(. Zp 1 − M dimensional simplex of all possible multinomial. Via convex hull, the K points define a 1 −≤ K L dimensional sub-simplex. The modeling assumption expressedby (1.1) is that conditional distributions for all documents are approximated by a multinomial representable as a convex combination of in which the mixture component uniquely define a point on the spanned sub-simplex which can identified with a concept space. A simple illustration of this idea is shown in)|(dwP)|(zwP)|(dzP) Figure 1.

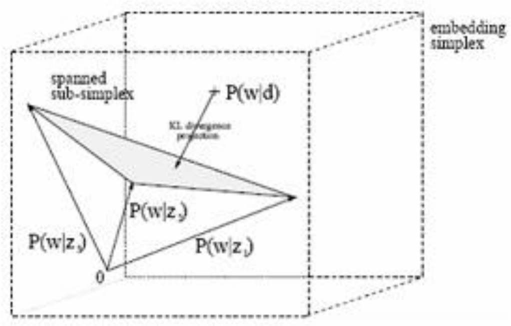

Figure 1.. Sketch of the probability sub-simplex spanned by the aspect mode

3.1. LIMITATIONS

In the aspect model, notice that is a dummy index into the list of documents in the training set. Consequently, d is a multinomial random variable with as many possible values as there are training documents and the model learns the topic mixtures only for those documents on which it is trained. For this reason, pLSI is not a well-defined generative model of documents; there is no natural way to assign probability to a previously unseen document. d)|(dzpA further difficulty with pLSA, which also originate from the use of a distribution indexed by training documents, is that the numbers of parameters grows linearly with the number of training documents. The parameters for a K-topic pLSI model are K multinomial distributions of size V and M mixtures over the K hidden topics. This gives KV + KM parameters and therefore linear growth in M. The linear growth in parameters suggests that the model is prone to overfitting and, empirically, overfitting is indeed a serious problem. In practice, a tempering heuristic is used to smooth the parameters of the model for acceptable predictive performance. It has been shown, however, that overfitting can occur even when tempering is used (Popescul et al., 2001, [41]). Latent Dirichlet Allocation (LDA - which is described in section 1.3. overcomes both of these problems by treating the topic mixture weights as a K-parameter hidden random variable rather than a large set of individual parameters which are explicitly linked to the training set.

3.2. LATENT DIRICHLET ALLOCATION

Latent Dirichlet Allocation (LDA) [7][20] is a generative probabilistic model for collections of discrete data such as text corpora. It was developed by David Blei, Andrew Ng, and Michael Jordan in 2003. By nature, LDA is a three-level hierarchical Bayesianmodel in which each item of a collection is modeled as a finite mixture over an underlying set of topics. Each topic, in turn, modeled as an infinite mixture over an underlying set of topic probabilities. In the context of text modeling, the topic probabilities provide an explicit representation of a document. In the following sections, we will discuss more about generative model, parameter estimation as well as inference in LDA.of data storage implementation. It can contains noisy ratio also very less. We find out destination results with less PSNR and BER.

3.3.GENERATIVE MODEL IN LDA

Given a corpus of M documents denoted by { } M dddD 21 = , in which each document number m in the corpus consists of Nm words drawn from a vocabulary of terms , the goal of LDA is to find the latent structure of “topics” or “concepts” which captured the meaning of text that is imagined to be obscured by “word choice” noise. Though the terminology of “hidden topics” or “latent concepts” has been encountered in LSA and pLSA, LDA provides us a complete generative model that has shown better results than the earlier approaches. iw{ V tt ,..., 1 }Consider the graphical model representation of LDA as shown in Figure 1, the generative process can be interpreted as follows: LDA generates a stream of observable words , partitioned into documents nm w , m dr. For each of these documents, a topic proportion m ϑ r is drawn, and from this, topic-specific words are emitted. That is, for each word, a topic indicator is sampled according to the document – specific mixture proportion, and then the corresponding topic-specific term distribution nm z , nm z ,φ r used to draw a word. The topics k φ r are sampled once for the entire corpus. The complete (annotated) generative model is

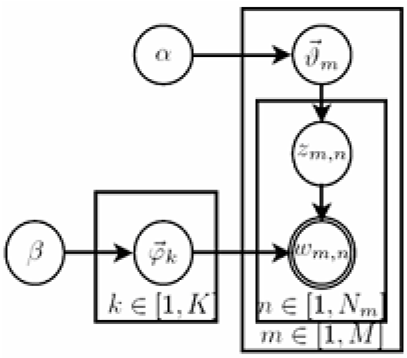

Figure 2. Graphical model representation of LDA - The boxes is “plates” representing replicates. The outer plate represents documents, while the inner plate represents the repeated choice of topics and words within a document

4. ALGORITHM

Algorithm : Gibbs sampling algorithm for Latent Dirichlet Allocation.

```
- Initialization

zero all count variables, () z m n m n ( ) t z n ,   z n
for all documents  do  [ ] Mm ,1 ∈
for all words   in document  do  [ m Nn ,1 ∈ ] m
sample topic index  ~Mult(1/K)  nm z ,
increment document-topic count: ( )1 + s m n
increment document-topic sum:  1 + m n
increment topic-term count: ( )1 + t sn
increment topic-term sum:  1 + z n
end for
end for
- Gibbs sampling over burn-in period and sampling period
while not finished do
for all documents  do  [] Mm ,1 ∈
for all words   in document  do  [ m Nn ,1 ∈ m
- for the current assignment of  to a term t for word  :  z nm w ,      decrement counts and sums: ( )
1 − zm n ; 1 − m n ;( )1 − tzn ; 1 − zn
- multinomial sampling acc. To Eq. 1.15 (decrements from previous step):    sample topic index  ( ) wzzpz iir r,|~ ~−
- use the new assignment of  to the term t for word  to:  z nm w ,
 increment counts and sums: ( )1 + zm nr; ; 1 + tznr 1 + znr
   end for
 end for
-  check convergence and read out parameters
 if converged and L sampling iterations since last read out then
  - the different parameters read outs are averaged
    read out parameter set Φacc. to Eq. 1.16
    read out parameter set Θacc. to Eq. 1.17
  end if
end while.
```

5. PREPROCESSING AND TRANSFORMATION

Data preprocessing and Transformation are necessary steps for any data mining process in general and for hidden topics mining in particular. After these steps, data is clean, complete, reduced, partially free of noises, and ready to be mined. The main steps for our preprocessing and transformation are described in the subsequent sections and shown in the following chart:

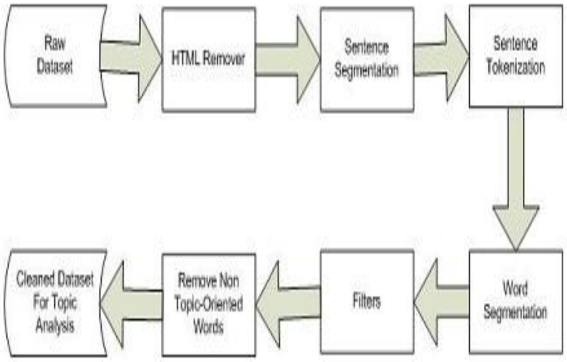

Figure 3. Pipeline of Data Preprocessing and Transformation

5 CLASSIFICATION WITH HIDDEN TOPICS

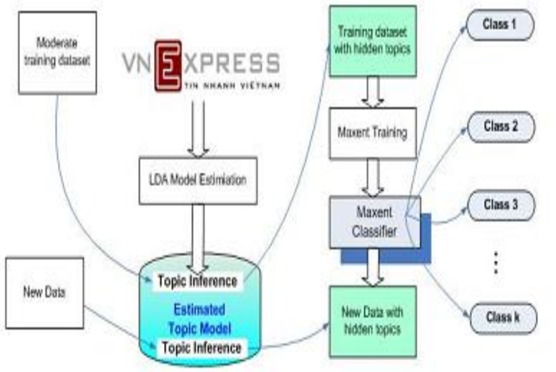

Figure 4. Classification with VnExpress topics

The objective of classification is to automatically categorize new coming documents into one of k classes. Given a moderate training dataset, an estimated topic model and k classes, we would like to build a classifier based on the framework in Figure 4.1. Here, we use the model estimated from VnExpress dataset with LDA . In the following subsections, we will discuss more about important issues of this deployment.

a. Data Description

For training and testing data, we first submit queries to Google and get results through Google API [19]. The number of query phrases and snippets in each train and test dataset are shown in Table 4.1 Google search results as training and testing dataset. The search phrases for training and test data are designed to be exclusive. Note that, the training and testing data here are designed to be as exclusive as possible.

b. Combining Data with Hidden Topics

The outputs of topic inference for train/new data are topic distributions, each of which corresponds to one snippet. We now have to combine each snippet with its hidden topics.

This can be done by a simple procedure in which the occurrence frequency of a topic in the combination depends on its probability. For example: a topic with probability greater than 0.03 and less than 0.05 have 2 occurrences, while a topic with probability less than 0.01 is not included in the combination.

5. PERFORMANCE EVALUATION

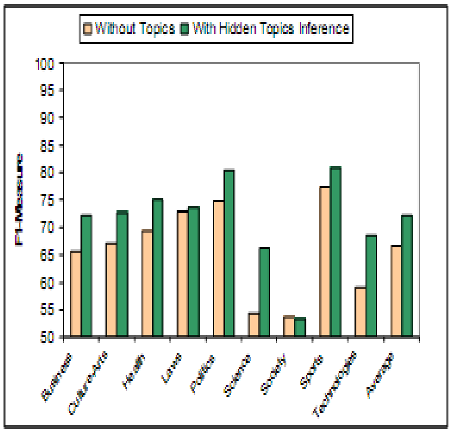

Figure 5. F1-Measure for classes and average (over all classes) in learning with 60 topics

shows the results of learning with different settings (without topics, with 60, 80, 100, 120, 140 topics) among which learning with 60 topics got the highest F-1 measure (72.91% in comparison with 66.57% in baseline). When the number of topics increase, the F-1 measures vary around 70-71% (learning with 100, 120, 140 topics). This shows that learning with hidden topics does improve the performance of classifier no matter how many numbers of topics is chosen. depicts the results of learning with 60 topics and different number of training examples. Because the testing dataset and training dataset are relatively exclusive, the performance is not always improved when the training size increases. In any cases, the results for learning with topics are always better than learning without topics. Even with little training dataset (1300 examples), the F-1 measure of learning with topics is quite good (70.68%). Also, the variation of F-1 measure in experiments with topics (2% - from 70 to 72%) is smaller than one without topics (8% - from 62 to 66%). From these observations, we see that our method does take effects even with little learning data.

6.CONCLUSION

In this paper we have presented We have presented a general framework to build classification and matching/ranking models for short and sparse text/Web data by taking advantage of hidden topics from large-scale external data collections. The framework mainly focuses on several major problems we might have when processing such kind of data: data sparseness and synonym/homonym problems. Our approach provides a way to makesparse documents more related and topic-focused by performing topic inference for them with a rich source of global information about words/terms and concepts/topics coming from universal datasets. The integration of hidden topics helps uncover and highlight underlying themes of the short and sparse documents, helping us overcome difficulties like synonyms, hyponyms, vocabulary mismatch, noisy words for better classification, clustering, matching, and ranking. In addition to sparseness and ambiguity reduction, a classifier or matcher built on top of this framework can handle future data better as it inherits a lot of unknown words from the universal dataset. Also, the framework is general and flexible to be applied to different languages and application domains.We have carried out two careful experiments for two evaluation tasks and they have empirically shown how our framework can overcome data sparseness and ambiguity in order to enhance classification, matching, and ranking performance.

REFERENCES

1. L. Baker and A. McCallum. Distributional clustering of words for text classification. In ACM SIGIR, 1998.
2. P. Baldi, P. Frasconi, and P. Smyth. Modeling the Internet & the Web: probabilistic methods & algorithms. Wiley, 2003.
3. S. Banerjee, K. Ramanathan, and A. Gupta. Clustering short texts using Wikipedia. In ACM SIGIR, 2007.
4. [4] A. Berger, A. Pietra, and J. Pietra. A maximum entropy approach to natural language processing. Comp. Ling., vol.22, no.1, pp.39–71, 1996.
5. R. Bekkerman, R. El-Yaniv, N. Tishby, and Y. Winter. Distributional word clusters vs. words for text cate. JMLR, vol.3, pp.1183–1208, 2003.
6. D. Blei, A. Ng, and M. Jordan. Latent Dirichlet Allocation. JMLR, vol.3, pp.993–1022, 2003.
7. D. Blei and J. Lafferty. A correlated topic model of Science. The Annalsof Applied Statistics, vol.1, no.1, pp.17-35, 2007.
8. D. Bollegala, Y. Matsuo, and M. Ishizuka. Measuring semantic similarity between words using Web search engines. In WWW, 2007.
9. [9] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co–training. In COLT, 1998.
10. Broder, M. Fontoura, V. Josifovski, and L. Riedel. A semantic approach to contextual advertising. In ACM SIGIR, 2007.
11. L. Cai and T. Hofmann. Text categorization by boosting automatically extracted concepts. In ACM SIGIR, 2003.
12. J. Cai, W. Lee, and Y. Teh. Improving WSD using topic features. In EMNLP-CoNLL, 2007.
13. P. Chatterjee, D. Hoffman, and T. Novak. Modeling the clickstream: Implications forWeb-based advertising efforts. Marketing Science, vol.22,no.4, pp.520–541, 2003.
14. K.Suresh,R.MadanaMohana and Dr.A.RamaMohan Reddy "Improved FCM algorithm for Clustering on Web Usage Mining", IJCSI International Journal of Computer Sciencec Issues, Vol.8 Issue 1, January 2011,ISSN(Online):1694-0814. http://www.ijcsi.org/papers/IJCSI-8-1-42-45.pdf.
15. M. Ciaramita, V. Murdock, and V. Plachouras. Semantic associations for contextual advertising. Journal of Electronic Commerce Research, vol.9,no.1, pp.1–15, 2008.

AUTHOR :

Mr.Peethani Raja Nagendra Prasad has received his Bachelor of Degree in COMPUTER SCIENCE & ENGINEERING from SASI Institute of Engineering and Technology, Tadepalligudem, West Godavari District, A.P., Affiliated to J.N.T.U., Hyderabad, and Pursuing M.Tech in COMPUTER SCIENCE & ENGINEERING from NOVA College of Engineering and Technology, Vegavaram, Jangareddygudam, West Godavari Dist, Affiliated to J.N.T.U., Kakinada, AP, India.

Mr.Ch.Raja Jacob, well known Author and Excellent Teacher Received M.C.A and M.Tech (CSE) from Acharya Nagarjuna University is working as Associate Professor and H.O.D, Department of MCA, M.Tech Computer Science Engineering, NOVA College of Engineering and Technology, He is an Active Member of ISTE. He has 7 years of Teaching Experience in various Engineering Colleges. To his credit couple of publications both National and International Conferences/Journals. His area of Interest includes Data Warehouse and Data Mining, Information Security, Flavors of Unix Operating Systems and other advances in Computer Applications.