CROSS-DOMAIN SENTIMENT ENCODING THROUGH STOCHASTIC WORD EMBEDDING

Swapna Mohan. M[1], J. Umarani[2]

[1]Swapna Mohan. M, Research Scholar, Thanthai Hans Reover College (Autonomous), Elambalur, Perambalur
[2]J.Umarani, M.Sc, M. Phil, B.Ed, (Ph.D)., Assistant Professor, Dept. of Computer Applications, Thanthai Hans Reover College (Autonomous), Elambalur, Perambalur

---***---

Abstract - *Sentiment analysis is an important topic regarding identification of feelings, attitudes, emotions and opinions from text. A critical challenge for computerizing such analysis is the high manual annotation cost when conducting large-scale learning. However, the cross-domain methodology is a key solution for this. It operates commented reviews across domains and its success principally depend on the learning of a robust common characterization space across domains to support the sentiment classifier transfer. Over recent years, consequential effort has been pervaded to improve the cross-domain representation learning by designing increasingly more complex and elaborate model inputs and architectures. We support that it is not necessary to focus on design twist as this inevitably consumes more time for model training. Instead, we propose to traverse through a simple depicting the word contradiction and occurrence information and encode such information more accurately whilst aiming at lower machanical costs. The proposed methodology is unique and takes advantage of the stochastic embedding technique to tackle cross-domain sentiment alignment. Its effectiveness is benchmarked with over ten data tasks constructed from two review corpora, and is compared against ten classical and state-of-the-art methods.*

Key Words: *Security, Information, Outsourcing, Encryption, Protection etc…*

1. INTRODUCTION

Sentiment classification plays a significant role in many applications related to opinion mining and sentiment analysis, such as opinion extraction and summarization, review spam identification, user feeling analysis, contextual advertising, etc. The goal of sentiment classification is to automatically identify the sentiment polarity of a given text object, for instance, in terms of being positive, negative or neutral. Typical examples of such text objects include product reviews, which are generated by movie viewers, hotel customers, merchandise buyers, etc. The emotional tendency modeled through identifying sentiment polarity of the reviews can serve as a succinct yet informative indicator of the consumer attitude and opinion. This can potentially result in not only improved efficiency in the information sharing between the users, but also improved business solutions and services.

Focusing on reviews of a target product, a standard sentiment classifier can be built by training with a set of annotated example reviews of this product. Here, annotation refers to the process of assigning each review example a ground-truth sentiment polarity label. The sentiment polarities of new reviews for the same product can then be predicted by this trained classifier, [8]. Performance of such a system heavily relies on the availability and quality of the labelled example reviews. However, the process of manually annotating explosively growing online product reviews is very expensive and can be impractical. Therefore, there has been increasing interest on studying effective ways of reusing labeled reviews across different products. This is known as cross-domain sentiment classification, where a domain is referred to as a collection of reviews for a particular product.

2. EXISTING SYSTEM

Comparing the domain adaption strategies used by various state-of-the-art techniques, we can see that, in addition to the main task of sentiment classification, they usually enhance their learning through preparing extra tasks like detecting whether a pivot co-occurs with a domain-specific word, whether the different versions of the same pivot word in different domains possess similar enough representation vectors, whether a review contains a pivot, or whether the reviews from the source and target domains can be distinguished in the representation space, and so on; we refer to these as the auxiliary learning tasks. The learning algorithms are mostly built on spectral approaches which explore and preserve inherent data structure through matrix decompositions or neural networks which directly learn the review representations through structured processing of the content words based on different network architectures and exhaustive training.

We argue that instead of creating many auxiliary learning tasks and constructing complex models with elaborate design and input configurations, satisfactory domain adaptation can be achieved by preserving simple polarity and occurrence information of words in reviews. These are actually parts of the classical information utilized in early cross-domain sentiment classification works, that however failed to achieve good performance. We support that the unsatisfactory performance of past endeavors were potentially caused by the employed spectral approaches that were incapable of preserving accurately the desired information in their embedding spaces. Similar observations on poor neighbor preservation ability of spectral embeddings are also reported in the data visualization field. To tackle this issue, we propose a novel cross-domain sentiment representation learning model with its design inspired by the stochastic neighbor embedding method. It is built upon a simple mapping architecture to ease the computational cost, but we propose a more sophisticated approach for optimizing the mapping variables to achieve more accurate similarity structure preservation.

2.1 Survey

A. EMOTION RECOGNITION AND BRAIN MAPPING FOR SENTIMENT ANALYSIS: A REVIEW

The meteoric growth of the Internet has caused the increase in the amount of textual information available, such as in blogs, discussion forums and review sites on the web, where the texts surely have the emotion content. Emotion is one appearence of people behaviour and it is an important performance in human computer interaction (HCI). Human convey the emotion in the form of facial expression, speech and writing text. Recently, researchers in computational linguistic (CL) areas are fascinated in the attention of emotion for Sentiment Analysis (SA). SA naturally perceives the emotion conveyed by a text, and at the same time, distinguishing positive and negative valence. The wide areas of CL research, literally appreciable for scrutinizing the emotion dimension detection and searching the approaches and techniques in the term of emotion recognition (ER). There are two significant drifts of research in the area, the emotion recognition based on state affective computing and the real time using brain signal machines. The two areas have the same goal for getting the upgraded result in sentiment analysis with the mapping of emotion recognition provided. The revelaive work on emotion detection is comparatively rare and lacks empirical evaluation research. This paper demonstrates the overview of past and recent research on emotion recognition as well as some approaches and techniques used and shows the linked between both SA and ER.

B. FEATURE BASED OPINION SUMMARIZATION OF ONLINE PRODUCT REVIEWS

With the strikable development of Web 2.0, an amazing growth of the social-media and e-commerce is being witnessed. This expansion of e-commerce has been characterized by the availability of vast number of products online. The websites selling these products allow its customers to express their views freely about a purchased product and these reviews may reach up to hundreds. Hence, it becomes very difficult for the customers to read each and every review and keep track of all the pros and cons of a particular product with respect to all the features that it possesses. The present study targets this problem. Subsequently, a feature based opinion summary of the product reviews is proposed as a solution. The proposed algorithm based on frequent item-set mining gives a set of candidate features for a product. Numerous feature filtering methods like subset filtering, superset filtering and distance based filtering are utilized to get rid of the redundant and very basic features. Finally, sentiment lexicon is used to determine the popularity of the opinion word associated with the extracted feature(s) and a final summary is developed.

Achieve our design goals of both system security and usability, we divide each CG into three index vectors based on the structure of the CG, the type of concept and the value of concept. We apply order preserving symmetric encryption (OPSE) to our scheme to enhance security. Experimental results exhibit the efficiency of our proposed scheme.

C.TASC: TOPIC-ADAPTIVE SENTIMENT CLASSIFICATION ON DYNAMIC TWEETS

Sentiment classification is a topic-delicative task, i.e., a classifier trained from one topic will perform worse on another. This is particularly a problem for the tweets sentiment analysis. Since the topics in Twitter are very manifold, it is impossible to train a universal classifier for all topics. Moreover, compared to product scrutiny, Twitter lacks data labeling and a rating mechanism to acquire sentiment labels. The extremely sparse text of tweets also attain the performance of a sentiment classifier. In this paper, we propose a semi-supervised topic-adaptive sentiment classification (TASC) model, which starts with a classifier built on common features and mixed labeled data from various topics. It depreciate the hinge loss to adapt to unlabeled data and features including topic-related sentiment words, authors' sentiments and sentiment connections derived from“@” mentions of tweets, named as topic-adaptive features. Text and non-text features are extracted and naturally split into two views for co-training. The TASC learning algorithm updates topic-adaptive features based on the collaborative selection of unlabeled data, which in turn helps to select more reliable tweets to boost the performance. We also design the adapting model along a timeline (TASC-t) for dynamic tweets. A demonstration on 6 topics from published tweet corpuses demonstrates that TASC outperforms other well-known supervised and ensemble classifiers. It also beats those semi-supervised learning methods without feature adaption. Meanwhile, TASC-t can also achieve impressive accuracy and F-score. Finally, with timeline envision of “river” graph, people can intuitively grasp the ups and downs of sentiments' evolvement, and the intensity by color gradation.

D. AN EFFECTIVE HYBRID CUCKOO SEARCH WITH HARMONY SEARCH FOR REVIEW SPAM DETECTION

In the recent years, online comments are one of most important source of customer opinion. Nowadays consumer can attain knowledge about the products and service from online review resources, using which they can make decisions. This may predominate Opinion Spam, where spammers may manipulate and fake reviews to promote artificially or devalue the products and other services. Opinion spam detection is done by extricating meaningful features from the text, and identifying the spam reviews using machine learning techniques. This demonstration results in a very high dimensional feature space. These features are inapplicable, redundant, and noisy which may affect the performance of the classifier. Therefore, a good accent selection method is required in order to speed up the processing rate, predictive accuracy. Evolutionary algorithms for accent selection can be used to handle these high-dimensional feature spaces which eliminate the noisy and irrelevant features. In this work, an constructive hybrid feature selection method using Cuckoo Search with Harmony search is proposed and Naive Bayes is used for classifying the review into spam and ham.

E. POLARITY SHIFT DETECTION APPROACHES IN SENTIMENT ANALYSIS: A SURVEY

Sentiment analysis is a process of recognizing and categorizing opinions expressed in a piece of text. It categorizes the text into positive, negative or neutral. Lexicon-based and Supervised Machine Learning-based are the two main techniques in sentiment analysis. Bag-of-words model is used to represent the text as a course of independent words and machine learning algorithms are used for classification.

3. PROPOSED SCHEME

In this paper, we solve the problem of how to enable a searchable encryption system with the support of semantic extension. Our work is one of only a few to study ranked search over encrypted data represented by CGs in Cloud Computing. We choose CG among various modes of knowledge representation as our semantic representation tool. In our scheme, we apply a state-of-the-art technique i.e., text summarization and Tregex, a tool for simplifying sentences to summarize the document. We introduce the existing scheme of constructing CGs to help generate CGs. To attain our design goals for both system security and usability, we divide each CG into three index vectors in response to the structure, the type of concept and the value of the concept of CG. Ranked search greatly improves system usability by returning the matching files in ranked order with regard to certain relevance criteria. Thus, to stipulate how the document convince the query and ranks the returned file, we use the – ”text summarization score” (TSS), which can gauge the extent to which the documents match their summarizations according to the relevance score. Moreover, if the score is higher, the document is more coincident with the original document. To protect the privacy of the TSS, we then integrate order preserving symmetric encryption. At the same time by using OPM (One to Many Preserving Mapping) technique more number of search results are stored in an graphical representation which reduces the storage space. The top results are ranked which is a benefit for the user i.e.: the search results are ranked according to the accuracy level and based on that top k ranked results, the search outputs are provided to the users in an efficient manner.

Advantages

•Increases more efficient search results.

•The search results increases the efficiency of the results.

•As the search results uses top k rank search results it is more prior.

•Reduces the cloud storage as it stores the results in a graphical representation and provides the searchable content to the user in a more specific manner.

•Exact search results are maintained in a distributed database which enhances the user preferences in an enormous way to attain the search contents.

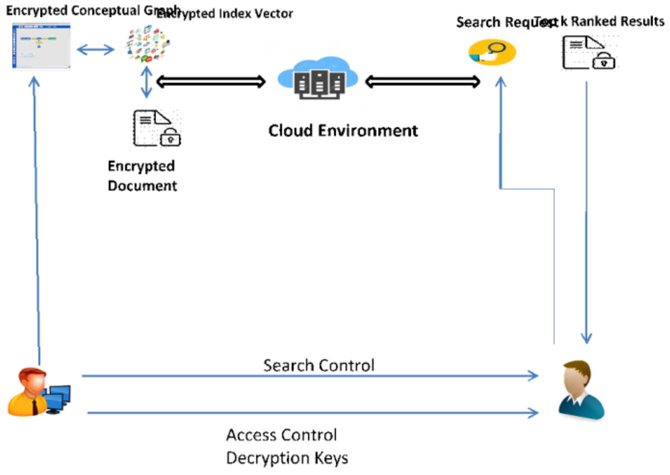

Fig -2: Architecture Diagram

Module Description

Cloud Owner Login

At the initial the cloud owner has to provide their login details to login to their preferred cloud login account. After login process the admin can concentrate on the user activities such as searching process and so on. So that the entire process has been under the control of the admin such as user accounts, search results and so on.

User Registration

To access the account details the user have to register their preferred details to gain the account details from the cloud owner. After registration each user has been provided with unique ID and password to access their account info. So with the help of the registered queries the admin gathers the basic details of the user.

User Login

After registration the user can login to their preferred account to access or search the preferred details. The details of each user and their data has been stored with privacy and security as they have separate username and login to access the data and as well as for search.

Encrypted Data

The data and the info stored by the user has been in an encrypted format for privacy preference. The encrypted has been stored and can be accessed by the separate user by their desired access info. The cloud database stores the user data in an encrypted format with the help of the AES algorithm.

Search Request

The user used to search their desired search request by using the proposed techniques and the results for the searched data has been provided to the user with the higher efficiency. The results for the searched data are top k ranked results that has been provided by the cloud server. The search process of the USSCG scheme is an algorithm named CGM (conceptual graphs match).

Access Control

The scheme is designed to prevent the cloud server from learning additional information regarding the document collection and query. The scheme is designed to prevent the cloud server from learning additional information regarding the document collection and query. There are still multiple index vectors in the query that cannot be distinguished well, we will rank the result according to the encrypted relevance scores by the modified OPSE We denote OPM as our one-to-many order-preserving mapping function with parameter:

Search Control

Text Summarization (TS) always tries to determine the ”meaning” of documents. Essentially, TS techniques are categorized as Extractive and Abstractive. In this paper, we focus on extractive summarization. Extractive summaries (ES) produce some of the most significant sentences extracted from the document instead of rebuilding. To solve the problem of semantic search based on conceptual graphs over encrypted outsourced data effectively in the above system model, our system has the following design goals.

Ranked search: The proposed scheme is designed to provide not only semantic query based on CG but also accurate result ranking.

Search Efficiency: The scheme aims to achieve low communication and computation overhead.

Algorithm Used

```
Algorithm 1 CGM
Input: F, D1, D2, D3, Q1, Q2, Q3, TSS
Output: F(Q1)
for each document Fi in F containing D1, D2, D3 do
if RScore(D1;Q1) > Score then
Insert D2, D3 into F(Q1);
RScore(D2;Q2);
RScore(D3;Q3);
Insert a new element (RScore(D2;Q2),
RScore(D3;Q3), FID) into RList;
else
return;
end if
end for
for each Fj , Fk in F(Q1) do
if RScore(Dj2;Qj2) == RScore(Dj2;Qk2) then
if RScore(Dj3;Qj3) == RScore(Dj3;Qk3) then
Compare the TSS of Fj , Fk to sort the elements of
F(Q1);
else
According to the RScore(D3;Q3) to sort the elements
of F(Q1);
end ifelse
According to the RScore(D2;Q2) to sorttheelements
of F(Q1);
end if
end for
```

Algorithm 2 One-to-many Order-preserving Mapping- OPM

```
Input: D, R, m, id(F)
Output: c
while jDj! = 1 do
fD;Rg   Binary Search(K;D;R;m);
end while
coin R  TapeGen(K; (D;R; 1 jj m; id(F)));
c coin   R;
Return c;.
```

4. CONCLUSION

In this paper, we define the problem of semantic search based on conceptual graphs over encrypted outsourced data for the first time. We choose CGs among various methods of knowledge representation to represent the documents. To generate the CGs, we apply a state-of-the-art technique, ie.,text summarization and Tregex a tool for simplifying sentences in our method. And to achieve our design goals of both system security and usability, we divide each CG into three index vectors based on the structure of the CG, the type of concept and the value of concept. We apply order preserving symmetric encryption (OPSE) to our scheme to enhance security. Experimental results exhibit the efficiency of our proposed scheme.

REFERENCES

[1] E. Cambria, "Affective computing and sentiment analysis," IEEE Intelligent Systems, vol. 31, no. 2, pp. 102–107, 2016.

[2] L.-W. Ku, Y.-T. Liang, H.-H. Chen et al., "Opinion extraction, summarization and tracking in news and blog corpora." in AAAI spring symposium: Computational approaches to analyzing weblogs, vol. 100107, 2006.

[3] S. Liu, X. Cheng, F. Li, and F. Li, "Tasc: topic-adaptive sentiment classification on dynamic tweets," IEEE Transactions on Knowledge and Data Engineering, vol. 27, no. 6, pp. 1696–1709, 2015.

[4] F. Li, M. Huang, Y. Yang, and X. Zhu, "Learning to identify review spam," in IJCAI Proceedings-International Joint Conference on Artificial Intelligence, vol. 22, no. 3, 2011, p. 2488.

[5] P. H. Calais Guerra, A. Veloso, W. Meira Jr, and V. Almeida, "From bias to opinion: a transfer-learning approach to real-time sentiment analysis," in Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, 2011, pp.150–158.