IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 16, Issue 5, Ver. I (Sep – Oct. 2014), PP 20-24
www.iosrjournals.org

An Efficient Method for Noisy Annotation Data Modeling

Sushama Shinde[1], Shyam Gupta[2]
[1](*Department of Computer Engg., Siddhant College of Engineering,Sudumbare,Pune/ Pune University, India)*
[2](*Department of Computer Engg., Siddhant College of Engineering, Sudumbare,Pune/Pune University , India)*

Abstract : *Probabilistic topic models are used for analyzing and extracting content-related annotations from noisy annotated discrete data like WebPages on WWW and these WebPages are stored using social bookmarking services with the help of social bookmarking services, reason behind this process most of time users can attach annotations freely, some annotations do not describe the semantics of the content, therefore they are noisy, simply they are not content related. The extraction of content-related annotations can be used as a prepossessing step in machine learning. Prepossessing step in machine learning is like text classification and image recognition, and can improve information retrieval performance. The proposed model is a generative model for content and annotations, where annotations are assumed to be originated either from topics that generated the content or from a general distribution unrelated to the content. We demonstrate the effectiveness of the proposed method with the help of synthetic data and real social annotation data for text and images.*
Keywords: *Book marking services, machine learning, social annotation , text classification .*

I. Introduction

Social annotations offer us a huge amount of user generated labeled data, see Yahoo! Delicious for examples. However, unlike expert-annotated data set, social annotations expose two liabilities: ambiguity and noise. Ambiguity rises when the users assign multiple tags to a single document. To the computer, every word in the document is related to all the tags. Noise is the nature of most user-generated content, and social annotation data is not an exception. Social annotation allows any string to be used as tags. In the meantime, the users are not professional annotators; they hold no responsibility to keep accuracy and consistency either. When we use social annotation as a labeled data set, reducing ambiguity and identifying noise tags are important. First, ambiguity reduction helps us to find out the real intention behind a tag, which leads to higher accuracy in related applications. On the other hand, the benefit of identifying and removing noise is straightforward. For example, if another user posts the news about web intelligent, we would not like to recommend "my favorite" to him, since he may be more interested in some other discipline. When dealing with Web-scale information, an automatic solution that can separate noise from good tags is more appropriate.

Content-unrelated annotations can often constitute disturbance if utilized for training samples in appliance learning jobs, such as self-acting text classification and likeness acknowledgement. Although the presentation of a classifier can usually be advanced by expanding the number of teaching samples, loud trials have a detrimental effect on the classifier. In this task advance classifier presentation if task can employ huge amounts of communal annotation facts and figures from which content unrelated annotations have been filtered out. Content unrelated annotations may furthermore constitute disturbance in data retrieval. There are numerous procedures offered for automatic annotation which has been proposed lately. But they implicitly suppose that all annotations are related to content, and to the best of our information, no attempt has been made to extract content-related annotations automatically.

These limitations are overwhelm by procedure offered in which probabilistic theme model for investigating and extracting content-related annotations from loud annotated facts and figures, which we call the loud annotation theme model (NATM) is presented. This procedure practically outperforms living procedures. although this procedure still needs expanding in number of directions. In this task we are giving the comprehensive NATM model. The NATM is based on theme forms. A theme form is a hierarchical probabilistic form, in which a article is modeled as a mixture of themes, and where a theme is modeled as a likelihood circulation over words. theme forms are effectively utilized for a broad variety of submissions encompassing data retrieval collaborative filtering, and visualization as well as for modeling annotated facts and figures. The NATM is an extension of correspondence latent Dirichlet allocation (Corr-LDA), which is a generative theme model for content and annotation. Since the Corr-LDA supposes that all annotations are associated to the content, it will not be used for dividing content-related annotations from content unrelated annotations.

The extraction of content-related annotations can be considered a binary classification difficulty. However, as considers genuine social annotation facts and figures, the annotations are not explicitly marked as content related/unrelated. thus, we will not use supervised binary classifiers such as the support vector machine

(SVM). The NATM is an unsupervised model, which can extract content-related annotations without content relevance labels.

In next section II we are presenting the literature survey over different methods presented. In section III, the proposed approach and its system block diagram is depicted. In section IV we are presenting the current state of implementation and results achieved. Finally conclusion and future work is predicted in section V.

II. Literature survey

In the literature survey, number of models has been proposed for automatic annotation especially for images. However, because they do not model the relevance between the content and annotation, they cannot be used for extracting content-related annotations. These automatic annotation methods are based on supervised classifiers, in which all annotations of training samples are considered even if they are unrelated to the content. This means they cannot be employed for the automatic generation of content-related annotations.

In S. Golder and B.A. Huberman [2], they present the Collaborative tagging describes the method by which numerous users add metadata in the form of keywords to distributed content. Collaborative tagging has developed in popularity on the world wide world wide web, on sites that permit users to tag bookmarks, photographs and other content. In this paper they investigated the structure of collaborative tagging schemes as well as their dynamic aspects. Specifically, discovered regularities in client undertaking, tag frequencies, kinds of tags utilized, bursts of attractiveness in bookmarking and amazing steadiness in the relative percentages of tags inside a given URL. They also present a dynamic form of collaborative tagging that forecasts these stable patterns and concerns them to imitation and distributed knowledge.

In K. Barnard, P. Duygulu, [3], present a new approach for modeling multi-modal data sets, focusing on the specific case of segmented images with associated text. Learning the joint distribution of image regions and words has many applications. By the considering in detail predicting words associated with whole images (auto-annotation) and corresponding to particular image regions (region naming). Auto-annotation might help organize and access large collections of images. Region naming is a model of object recognition as a process of translating image regions to words, much as one might translate from one language to another. Learning the relationships between image, regions and semantic correlates (words) is an interesting example of multi-modal data mining, particularly because it is typically hard to apply data mining techniques to collections of images. In this they developed a number of models for the joint distribution of image regions and words, including several which explicitly learn the correspondence between regions and words. In this studied multi-modal and correspondence extensions to Hofmann‟s hierarchical clustering/aspect model, a translation model adapted from statistical machine translation and a multi-modal extension to mixture of latent Dirichlet allocation (MoM-LDA). All models are assessed using a large collection of annotated images of real.

In D.M. Blei and M.I. Jordan [4], advised the difficulty of modeling annotated facts and figures and facts and figures with multiple kinds where the example of one kind (such as a caption) serves as a description of the other type (such as an image). Describe three hierarchical probabilistic blend models that are aimed at such data, culminating in the Corr-LD a form, a latent variable that is productive of modeling the joint circulation of both types and the conditional circulation of the annotation granted the prime type.

In S. Feng, R. Manmatha, and V. Lavrenko [5], suggested that retrieving images in response to textual queries requires some information of the semantics of the image. How can both automatic likeness annotation and retrieval (using one phrase queries) from images and videos using a multiple Bernoulli relevance model. The form supposes that a teaching set of images or videos along with keyword annotations is supplied. Multiple keywords are supplied for a likeness and the exact correspondence between a keyword and a likeness is not provided. Each likeness is partitioned into a set of rectangular regions and a real-valued characteristic vector is computed over these regions. The relevance model is a junction likelihood circulation of the word annotations and the likeness characteristic vectors and is computed using the training set. The word probabilities are estimated utilizing a multiple Bernoulli form and the likeness feature probabilities using a non-parametric kernel density estimate.

III. System Architecture.

In this section , the proposed approach and its system block diagram is depicted.

3.1 Architecture

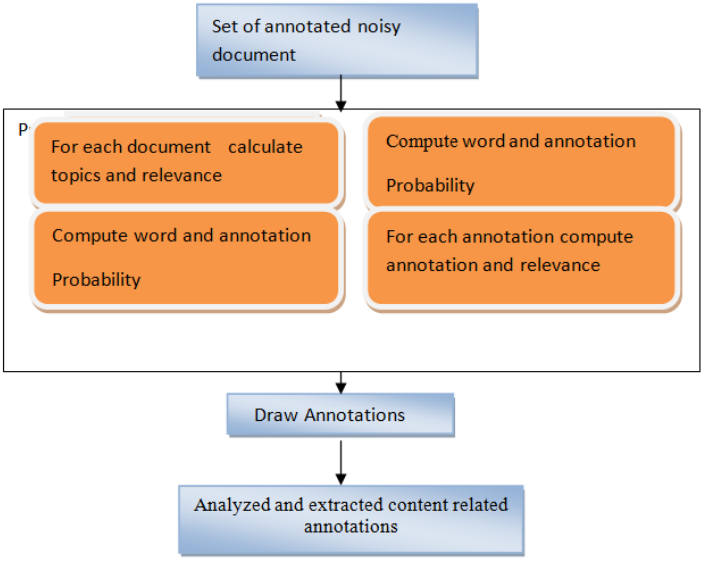

Fig.3.1 System Architecture

3.2 Proposed Work :

In this paper the aim is to design the extended approach for probabilistic topic model for analyzing and extracting content-related annotations from noisy annotated discrete data such as WebPages stored using social bookmarking services. This proposed approach is designed using the existing efficient model called NATM (noisy annotation topic model). We are presenting the extended version of NATM by adding the functionalities in it. This extended method is called XNATM (extended NATM). In this model we are adding functionality which is used to incorporate user information into the model for modeling social annotation data. This approach was not included in NATM. This proposed model can be applied in both implicit and partially explicit relevance settings, and it can also be used as the preprocessing for different classifiers as well as for modeling noisy annotated data. The proposed model is a generative model for content and annotations, in which the annotations are assumed to originate either from topics that generated the content or from a general distribution unrelated to the content.

3.3 Advantages of Proposed Algorithm

- Even if annotations have the same name, the XNATM is able to consider some of them to be related to the content and others not by taking account of the relevance of each annotation to the content.
- XNATM does not require any thresholds because the classification is explicitly achieved by inferring a latent variable that represents the relevance of each annotation to the content.
- XNATM simultaneously models content and annotations with their relevance in one probabilistic framework.
- As compared to NATM, in XNATM the user information is also incorporated in model which will helpful for efficiently modeling social annotation data.

3.4 Algorithm

1. Draw relevance probability
2. Draw content-unrelated annotation probability
3. For each topic k=1,…..,k;
    a. Draw word probability
        b. Draw annotation probability
4. For each document d=1,…..,D;
    a. Draw topic for word
    b. For each word n=1,….,Nd;

i. Draw topic for word
ii. Draw word

c. For each annotation m=1,….,Md;
i. Draw topic for annotation
ii. Draw relevance
iii. Draw annotation

3.5 Mathematical Model

Input Sets: Block catalog dataset.

Process:

The joint distribution on words, annotations, topics for words, topics for annotations, and relevance is described as follows:

$P(W,T,Z,C,R\backslash\alpha,\beta,\gamma,\eta) = P(Z\backslash\alpha)P(W\backslash Z,\beta)P(T\backslash C,R,\gamma)P(R\backslash\eta)P(C\backslash Z)$, ……..(1)

Where, $\boldsymbol{W}=\{^{\omega}{}_{d}\}^{D}{}_{d=1,}$ $\boldsymbol{T}=\{t_d\}^{D}{}_{d=1,}$ $\boldsymbol{Z}=\{z_d\}^{D}{}_{d=1,}$
$C=\{c_d\}^{D}{}_{d=1,}\boldsymbol{c_d}=\{c_{dm}\}^{Md}{}_{m=1,}\boldsymbol{R}=\{r_d\}^{D}{}_{d=1}$, And $r_{d=}\{r_{dm}\}^{md}{}_{m=1.}$

We can integrate out multinomial distribution parameters, $\{\Theta_d\}^{D}{}_{d=1}$, $\{\Phi_k\}^{K}{}_{k=1}$,and $\{\psi_\kappa\}^{K}{}_{\kappa=0}$

Because we use Dirichlet distributions for their priors, which are conjugate to multinomial distributions. the first term on the right hand side of (1) is calculated by

$P(Z\backslash\alpha) = \Pi^{D}{}_{d=1} \int P(z_d\backslash\Theta_d)\, P(\Theta_d\backslash\alpha)d\Theta_d$ and we have the following equation by integrating out $\{\boldsymbol{\theta}_d\}^{D}{}_{d=1}$:

$$P(W|Z,\beta) = \left(\frac{\tau(\beta W)}{\tau\beta^{W}}\right)_{k} \prod_{k} \frac{\prod_{w}\tau(N_{kw}+\beta)}{\tau(N_k+\beta W)} \quad \text{…….. (2)}$$

Where N_{kw} is the number of times word ω has been assigned to topic k, and $N_k = \sum_{\omega} N_{k\omega}$.

Where $K \in \{0,\ldots,K\}$, And κ=0 indicates irrelevance to the content. $M_{\kappa t}$ is the number of times annotation t has been identified as content unrelated if κ =0, or as content-related topic κ if $\kappa \neq 0$,and M_{Kt} . the bornoulli parameter **λ** can also be integrated out because we use a beta distribution for the prior, which is the conjugate prior of a Bernoulli distribution, the fourth term is given as follow:

$$P(R|\eta) = \frac{\tau(2\eta)\tau(M_{0+\eta})\tau(M-M_{0+\eta})}{\tau(\eta)^{2}\tau(M+2\eta)} \quad \text{…….. (3)}$$

Where M is the number of annotations, and $M_{o\ is}$ the number of content unrelated annotations,

IV. Work Done

Output Sets: Content related annotations.

4.1. Input Dataset:

For implementation purpose BLOCK CATLOG data set were use.

4.2 Hardware and Software Used

Hardware Configuration

- processor : Pentium iv 2.6 ghz
- ram : 512 mb dd ram
- monitor : 15” color
- hard disk : 20 gb
- floppy drive : 1.44 mb
- CD drive : lg 52x
- keyboard : standard 102 keys
- mouse : 3 buttons

Software Configuration

Front End : JAVA
Tools Used : Eclipse
Operating System : Windows XP/7

4.3 Results of Practical Work

The results of common procedure explained for XNATM are the actual results . After following all the procedure given in implementation details ,the results for the proposed system are compared with the results obtained by existing system .And applying Navie Bayes algorithm to existing system is a new contribution .The following graph is obtained after this comparison .This graph shows that the results are more accurate .

The graph shows that the results obtained by proposed system are more accurate than existing system . After the discussion on various aspects of Social annotation and the extended NATM model for finding the relevance between word and annotated data, the following results were found .Those results are given in the implementation details. The results of common procedure explained for XNATM are the actual results .

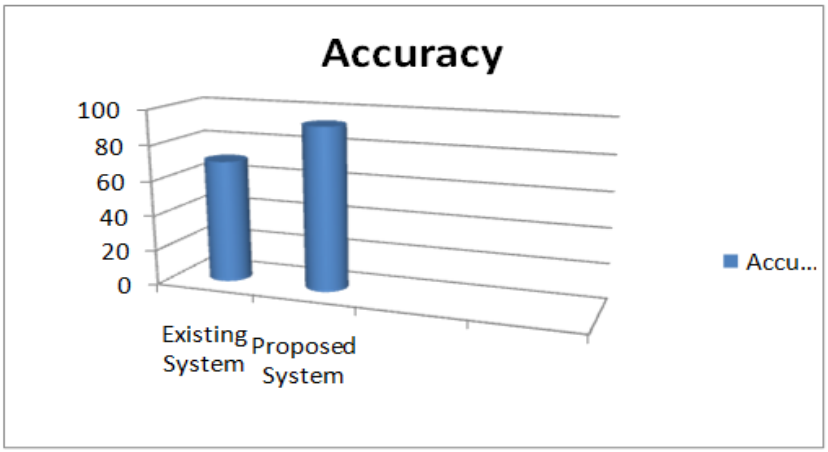

Fig. 4.1 Accuracy Graph

Fig 4.1 shows the accuracy graph for the system. It shows that the accuracy increased by applying XNATM strategy . This is the advantage of using the proposed system and comparing it to the existing system.

V. Conclusion

In this suggested scheme form we have a theme form for extracting content related to annotations from loud annotated data. Here it is possible to request suggested model in both implicit and partially explicit relevance backgrounds, and it is mostly utilized for pre-processing different classifiers and also for modelling noisy annotated facts and figures. We have verified experimentally that the suggested method can extract content-related annotations appropriately, and also be utilized for investigating social annotation facts and figures. Definitely our outcomes have been encouraging to designated day.

It is essential to extend our approach in a number of directions. Primarily we need to work out the number of topics automatically by expanding the suggested form to a nonparametric Bayesian model. Second, we want to incorporate user data into the form for forming communal annotation facts and figures. Third, a structure to deal with content-unrelated annotations can be used in models other than topic forms. Eventually, the suggested method is theoretically applicable to various types of annotation facts and figures; we will affirm this in added experiments.

In future work for this work can be used for voting mechanism for classification. CART is the ultimate classification tree that has revolutionized the entire field of advanced analytics and inaugurated the current era of data mining. CART, which is continually being improved, is one of the most important tools in modern data mining. Others have tried to copy CART but no one has succeeded as evidenced by unmatched accuracy, performance, feature set, built-in automation and ease of use. Designed for both non-technical and technical users, CART can quickly reveal important data relationships that could remain hidden using other analytical tools.

References

[1] S. Golder and B.A. Huberman, "Usage Patterns of Collaborative Tagging Systems," J. Information Science institute, vol. 32, no. 2, pp. 198-208, 2006.
[2] K. Barnard, P. Duygulu, D. Forsyth, N. de Freitas, D.M. Blei, and M.I. Jordan, "Matching Words and Pictures," J. Machine Learning Research, vol. 3, pp. 1107-1135, 2003.
[3] D.M. Blei and M.I. Jordan, "Modeling Annotated Data," Proc. 26th Ann. Int"l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR "03), pp. 127-134, 2003
[4] S. Feng, R. Manmatha, and V. Lavrenko, "Multiple Bernoulli Relevance Models for Image and Video Annotation," Proc. IEEE CS Conf. Computer Vision and Pattern Recognition (CVPR "04), vol. 2, pp. 1002-1009, 2004.
[5] J. Jeon, V. Lavrenko, and R. Manmatha, "Automatic Image Annotation" and "Retrieval Using Cross-Media Relevance Models," Proc. 26th Ann. Institutes" ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR "03), pp. 119-126, 2003.
[6] J. Jeon and R. Manmatha, "Using Maximum Entropy for Automatically Image Annotation," Proc. Image and Video Retrieval (CIVR "04), pp. 24-32, 2004.
[7] D.M. Blei, A.Y. Ng, and M.I. Jordan, "Latent Dirichlet Allocation," J. Machine Learning Research, vol. 3, pp. 993-1022, 2003.
[8] T. Hofmann, "Probabilistic Latent Semantic Analysis," Proc. 15th Conf. Uncertainty in Artificial Intelligence (UAI "99), pp. 289-296, 1999.
[9] T. Hofmann, "Collaborative Filtering via Gaussian Probabilistic Latent Semantic Analysis," Proc. 26th Ann. Int"l ACM SIGIR Conf. Research and Development in Information Inst., pp. 259-266, 2003.
[10] T. Iwata, S. Watanabe, T. Yamada, and N. Ueda, "Tracking Model for Analyzing Consumer Purchase Behaviour," Proc. 21st Int"l Joint Conf. Artificial Intelligence (IJCAI "09), pp. 1427-1432, 2009.

Proceedings Papers:

[11] Tomoharu Iwata, Takeshi Yamada, "Modelling Noisy Annotated Data with Application to Social Annotation", 2013 IEEE.