Joint Latent Dirichlet Allocation for Social Tags

INTRODUCTION

- The explosion of online media, many websites such as Flickr, Delicious and Cite ULike provide the tagging services to annotate media objects (e.g., image, video, music and web page) in the forms of short text. These data, termed as social tags, present us new opportunities as well as challenges to better understand the multimedia objects or user behavior patterns on the internet, and have been widely used in many applications including multimedia retrieval , image analysis [6–8], recommender system, and user modeling. However, although general text analysis methods such as LDA or its variants have shown promise for social tag analysis, these models fail to consider the following two unique characteristics of social tags.

Abstract

- Social tags, serving as a textual source of simple but useful semantic metadata to reflect the user preference or describe the web objects, has been widely used in many applications. However, social tags have several unique characteristics, i.e., sparseness and data coupling (i.e. non-IIDness), which makes existing text analysis methods such as LDA not directly applicable. In this article, we propose a new generative algorithm for social tag analysis named Joint Latent Dirichlet Allocation, which models the generation of tags based on both the users and the objects, and thus accounts for the coupling relationships among social tags. The model introduces two latent factors that jointly influence tag generation: the user's latent interest factor and the object's latent topic factor, formulated as user-topic distribution matrix and object-topic distribution matrix, respectively. A Gibbs sampling approach is adopted to simultaneously infer the above two matrices as well as a topic-word distribution matrix. Experimental results on four social tagging data sets have shown that our model is able to capture more reasonable topics and achieves better performance than five state-of-the-art topic models in terms of the widely used point-wise mutual information (PMI) metric. In addition, we analyze the learnt topics showing that our model recovers more themes from social tags while LDA may lead the topic vanishing problems, and demonstrate its advantages in the social recommendation by evaluating the retrieval results with Mean Reciprocal Rank (MRR) metric. Finally, we explore the joint procedure of our model in depth to show the non-IID characteristic of social tagging process.

INTRODUCTION

- The explosion of online media, many websites such as Flickr, Delicious and CiteULike provide the tagging services to annotate media objects (e.g., image, video, music and web page) in the forms of short text. These data, termed as social tags, present us new opportunities as well as challenges to better understand the multimedia objects or user behavior patterns on the internet, and have been widely used in many applications including multimedia retrieval , image analysis , recommender system, and user modeling .

EXISTING SYSTEM

- semantic metadata to reflect the user preference or describe the web objects, has been widely used in many applications. However, social tags have several unique characteristics, i.e., sparseness and data coupling (i.e. non-IIDness),which makes existing text analysis methods such as LDA not directly applicable.
- These data, termed as social tags, present us new opportunities as well as challenges to better understand the multimedia objects or user behavior patterns on the internet, and have been widely used in many applications including multimedia retrieval.

DISADVANTAGES

- A Gibbs sampling approach is adopted to simultaneously infer the above two matrices as well as a topic-word distribution matrix. Experimental results on four social tagging data sets have shown that our model is able to capture more reasonable topics and achieves better performance than five state-of-the-art topic models in terms of the widely used point-wise mutual information (PMI) metric.

PROPOSED SYSTEM

- we propose a new generative algorithm for social tag analysis named Joint Latent Dirichlet Allocation, which models the generation of tags based on both the users and the objects, and thus accounts for the coupling relationships among social tags. The model introduces two latent factors that jointly influence tag generation: the user's latent interest factor and the object's latent topic factor, formulated as user-topic distribution matrix and object-topic distribution matrix, respectively.
- A Gibbs sampling approach is adopted to simultaneously infer the above two matrices as well as a topic-word distribution matrix. Experimental results on four social tagging data sets have shown that our model is able to capture more reasonable topics and achieves better performance than five state-of-the-art topic models in terms of the widely used point-wise mutual information (PMI) metric.

ADVANTAGES

- the user's latent interest factor and the object's latent topic factor, formulated as user-topic distribution matrix and object-topic distribution matrix, respectively. A Gibbs sampling approach is adopted to simultaneously infer the above two matrices as well as a topic-word distribution matrix.
- Experimental results on four social tagging data sets have shown that our model is able to capture more reasonable topics and achieves better performance than five state-of-the-art topic models in terms of the widely used point-wise mutual information (PMI) metric.
- In addition, we analyze the learnt topics showing that our model recovers more themes from social tags while LDA may lead the topic vanishing problems, and demonstrate its advantages in the social recommendation by evaluating the retrieval results wit Mean Reciprocal Rank (MRR) metric.

HARDWARE REQUIREMENTS

- Processor - Pentium –III
- Speed - 1.1 Ghz
- RAM - 256 MB(min)
- Hard Disk - 20 GB
- Floppy Drive - 1.44 MB
- Key Board - Standard Windows Keyboard
- Mouse - Two or Three Button Mouse
- Monitor - SVGA

SOFTWARE REQUIREMENTS

- Operating System : Windows 8
- Front End : Java /DOTNET
- Database : Mysql/HEIDISQL

CONCLUSION

- In this paper, we proposed a Joint Latent Dirichlet Allocation method for modeling social tags. By considering the non-IID characteristic of social tags, our model introduces two factors, the user interest factor and the object latent topic factor, to jointly affect the generative procedure of tags. As our model utilizes the collaborative information among the users and the objects to extract more explicit information from tags, experiments on four publicly available data sets have demonstrated the advantages of our models compared with other five topic models in terms of PMI scores. Besides, we analyze the learnt topics in Delicious data set and show that our model recovers more thematically topics from social tags while LDA may lead to vanishing topics by considering these social data in only one perspective. Following that, we further show the advantages of JLDA in the social recommendation application. Finally, we explore the joint procedure of social tags at the topic level and explain that our model is more interpretable than LDA in the generative procedure of social tags.

REFERENCES

[1] X. Li, C. G. Snoek, and M. Worring, "Learning social tag relevance by neighbor voting," Multimedia, IEEE Transactions on, vol. 11, no. 7, pp. 1310–1322, 2009.

[2] M. Levy and M. Sandler, "Music information retrieval using social tags and audio," Multimedia, IEEE Transaction on, vol. 11, no. 3, pp. 383–395, 2009.

[3] L. Chen, D. Xu, I. W. Tsang, and J. Luo, "Tag-based image retrieval improved by augmented features and group-based refinement," Multimedia, IEEE Transactions on, vol. 14, no. 4, pp. 1057–1067, 2012.

[4] Y. Mao, H. Shen, and C. Sun, "A probabilistic topic model with social tags for query reformulation in informational search," in Advanced Data Mining and Applications, 2011.

[5] L. Wu, R. Jin, and A. Jain, "Tag completion for image retrieval," PAMI Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2013.