Tripartite hidden topic models for personalised tag suggestion

Morgan Harvey[1], Mark Baillie[1], Ian Ruthven[1] and Mark Carman[2]
[1] Strathclyde University, Computer and Information Sciences Department
[2] University of Lugano, Faculty of Informatics

Introduction

Social tagging systems provide methods for users to categorise resources using their own choice of keywords (or "tags") without being bound to a restrictive set of predefined terms. Unfortunately this freedom of word choice comes at a significant cost in terms of both data sparsity and ambiguity of tag meanings.

Tagging systems typically provide simple tag recommendations to increase the number of tags assigned to resources.

In this work we extend the latent Dirichlet allocation (Blei 2003) topic model to include user data and use the estimated probability distributions in order to provide personalised (and therefore hopefully more relevant) tag suggestions to users.

Modelling topics

Topic models attempt to probabilistically uncover the underlying semantic structure of a collection of resources modelled over a number of hidden topics which are assumed to be present in the collection. Each word is assumed to be generated based on both the probability that word given the topic $P(w \mid z)$ and the probability of the topic given the word's parent document $P(z \mid d)$. In our new model, the Tripartite Topic Model (TTM), we include $P(z \mid u)$ into the process by assuming:

$$p(z|\theta_d, \theta_u) \propto \frac{p(z|\theta_d)p(z|\psi_u)}{p(z)}$$

All parameters are calculated from the data by averaging over successive MCMC Gibbs samples (Griffiths 2004).

We use topic models as they:

- are fully Bayesian and cope well with sparse data
- do not require explicit co-occurence between terms (tags) in order for them to share semantic similarity and therefore deal with the ambiguity inherent in social tagging data

Experiment

We evaluated our model's ability to provide good personalised tag suggestions by comparing it to LDA, 2 common recommendation techniques - TopSys and TopUser - and CoTag (Schmitz 2006). We conducted our tests on real social tagging data provided by Bibsonomy with 20% for testing, 80% to train. We conducted the tests over 10 randomly sampled folds of data to ensure statistical accuracy. For each set of annotations in the test set the models tried to suggest the last tag.

Given a pseudo-query q (all other tags except the last one), the suggested tags are ranked via the following formula:

$$P(w|q,u) = \sum_{z}^{Z} P(w|z)\frac{P(z|q)\psi_{z|u}}{P(z)}$$

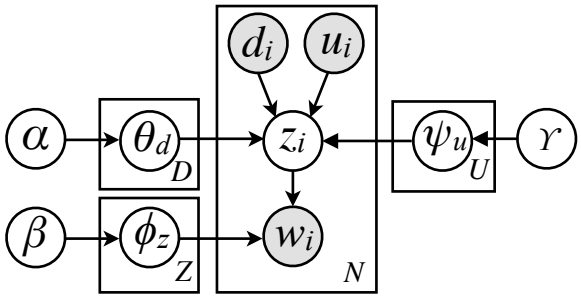
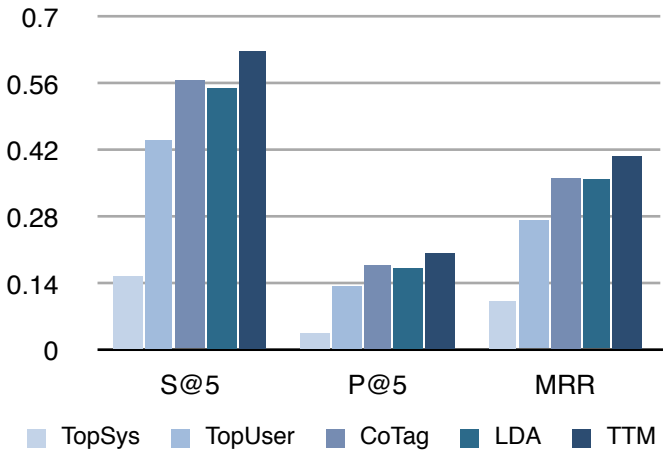


Figure 1: Tripartite Topic Model Diagram

Results



Conclusions

Our model consistently suggests more relevant tags than current systems. In terms of precision, the use of our model significantly improves upon:

- the CoTag method by between 7.87 and 13.6%
- basic LDA by 11.4 to 19.1%
- simpler, more common methods by an even larger margin

TTM provides a complete model of the data collected from a folksonomy and therefore could easily be utilised in future work for other useful estimations and is not merely suited to tag suggestion. For example resource suggestion (collaborative filtering), identification of latent communities of practise and personalised searching and ranking.

References

D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. Journal of Machine Learning Research, 2003.

N. Garg and I. Weber. Personalized tag suggestion for flickr. In WWW, 2008.

T. Griffiths and M. Steyvers. Finding scientific topics. PNAS, 2004.

B. Sigurbjornsson and R. van Zwol. Flickr tag recommendation based on collective knowledge. In WWW, 2008.

P. Schmitz. Inducing ontology from flickr tags. In WWW, 2006.