Affiliation recommendation using auxiliary friendship networks

Vishvas Vasuki, Nagarajan Natarajan, Zhengdong Lu, Inderjit Dhillon

July 8, 2010

1 Abstract

1. Introduce and motivate the problem

    Introduce social network analyis.

    Introduce the affiliation recommendation problem.

    Introduce affiliation networks.

    We consider a particular case: community recommendation.

    In general, affiliations can be items. Also, there are applications in biology: gene - disease networks.

2. Key contributions

    Simple way of combining the networks. Based on this, we explore two classes of algorithms.

    Graph proximity based approaches common in link prediction. The use of such link prediction techniques in recommender systems is, to our knowledge, novel.

    Approaches based on the latent factor model.

    Graph proximity based approaches do better.

    Propose a way of evaluating affiliation recommendations. We demonstrate the importance of designing the right evaluation strategy.

3. State that you evaluate on Youtube and Orkut, by seeing how good the top 50 recommendations are.

2 Introduction

1. Introduce and motivate problem, as in abstract, but in more detail.
2. Describe key contributions, as in abstract, but in more detail.

2.1 Overview

1. The table of contents as a paragraph.

3 Models

1. Overview of the section.
2. Establish notation.
3. Pose the affiliation recommendation problem as a ranking problem. The methods we describe to solve this problem rely on assigning scores to various items. Describe the score matrix.

3.1 Prediction on the combined graph

1. This is our way of combining the social and affiliation networks. Describe the joint adjacency matrix.
2. Briefly introduce the scoring approaches based on graph proximity and based on latent factors approach.

3.2 Graph proximity model

1. Introduce the basic Katz measure for general graphs.
2. Extend Katz measure to bipartite graphs. Interpret it as a weighted combination of paths.
3. Extend Katz measure to the combined graph C. Interpret it as a weighted combination of paths.
4. Introduce truncated Katz measure.
5. Analysis of computational cost.

3.3 Latent factors model

1. Overview of the section.

2. Say that the 0 entries in A are actually unknown, but that there is a huge prior on them actually being 0.

3. Model the adjascency matrix as arising out of inner products between user and group factors, which are low dimensional representations of users and groups. Users and groups with 'high' inner products are likely to be connected to each other.

4. Build the combined matrix C with unknown entries in the A part, and model this as arising out of interactions between user and group factors.

5. Describe the objective being minimized.

6. Describe role of λ in C.

7. Describe various potential choices for D. Observe that choice of D is not very important, according to experiments.

8. Describe SVD(C) as the solution to the above optimization problem.

9. Analysis of computational cost.

4 Related work

1. Overview of the section.

2. Describe the area of recommendation systems research.

   Community recommendation is less studied.

3. Probabilistic collaborative filtering: Prior work.

   Community recommendation using LDA by Chen et al: They have not used the social network in making recommendations.

   Relate their LDA based approach to the latent factors approach. Describe LDA's connections with pLSA and LSA, which is just SVD.

   Combinatorial collaborative filtering is also based on LSA, but uses text descriptions rather than social networks.

4. Prior work in joint matrix factorization.

   Compare with Linked Matrix Factorization by Tang et al.

   Compare with Collective Matrix Factorization by Singh et al, which generalizes this.

   Latent factors approach we propose are much more efficient than them: involve computing SVD, rather than using optimization techniques based on alternating least squares.

5. Prior work in modelling and studying the co-evolution of social and affiliation networks.

   Inspires our community recommendation effort.

6. Prior work in social network analysis.

   Inspires our graph proximity based scoring models.

7. Our attempt to use link prediction/ graph proximity based techniques is novel.

5 Experimental evaluation

1. Overview of the section.

5.1 Data

1. Describe datasets.

   Present some statistics. Use bar graphs like those in Orkut and Youtube. Make observations about them.

2. Describe the test, training and validation sets.

   Describe how the (per-user) test set is created.

   Describe how validation set is created. Mention the number of predictions made during validation, in comparing various parameters.

5.2 Evaluation method

1. Overview of the subsection.

2. Describe sensitivity, specificity, precision.

3. Describe AUC, ROC, the use of the appropriate slice of ROC in evaluating the performance of a predictor in making the first 50 predictions.

4. Note the robustness of the results: sensitivities and specificities were averaged over 9500 users in Orkut and 16000 users in Youtube.

5. Note the importance of using the right evaluation method. Contrast with results seen while using link prediction-style evaluation methods : Include the bar graph.

   Show by algebra that different quantities are being measured in the two evaluation strategies. Note that global sensitivity measurement is same as taking a weighted average over sensitivities.

   Emphasize this as a contribution of this paper.

5.3 Results and discussion

1. Overview of the subsection.

2. Describe performance of graph proximity based methods.

   For the average user, graph proximity methods are much better.

   Note that use of social network in recommendation is important.

   Specify learned parameters.

3. Discuss performance of methods based on the latent factors model.

   Note that using C makes a big difference. Note that use of social network in recommendation is important.

   Note that using various choices for the group-group part does not make significant difference.

   Specify learned parameters.

4. Conclude.

   Note general consistency of the results across different datasets.

   Remind that we explored two classes of algorithms, and that the graph-proximity based ones are better.

6 Conclusion and future work

1. Mention the problem again.

   Motivate the problem and its general applicability. Eg: gene-disease networks.

2. Outline key contributions.

3. Describe future work.

   Using the affiliation network in social network link prediction: Early experiments indicate that this is hard.

   Mention application in areas beyond community recommendation: biology example.

   Community recommendation using even more sources of information: Chen et al used text information.

   Combining predictors.

7 Acknowledgements

1. Acknowledge Prateek, Berkant, Alan Mislove.

8 Corrections to be made

Flawed claim: 'The inherent low rank nature of the user and group factors conforms to the huge prior of the unobserved entries in A being 0.' Instead, say that most unknown entries in A are 0.

We sould probably have made this point in the paper.

There were some other points we wished to make:
Eg: Showing computation-cost data, mentioning that the different recommenders could be combined together to make better predictions, adding more pictures explaining features of the datasets etc..

In Figure 3: 'Comparison of latent factors based algorithms', dataset is not identified. Make caption more descriptive. Refer to the right section for discussion in the caption. Fix its discussion: ambiguous description.

In Table 2: 'best parameters learned', no learned parameter is specified for SVD(C) on orkut.

Fix / complete related work section to cite recommender systems literature.

Story about validation is incomplete: how many predictions are being made in comparing different parameters?