Algorithms for User Recommendation in Social Networks

Chuang Zhang[1], Bocheng Zhu[1], Ming Wu[1]
Yun Huang[2], Noshir Contractor[2]

[1] PRIS LAB. Beijing University of Posts and Telecommunications
[2] SONIC LAB. Northwestern University

This research was supported by grants from the National Science Foundation (PetaApps: OCI-0904356 & NetSe: CNS-1010904).

NORTHWESTERN UNIVERSITY

SONIC
advancing the science of networks in communities

Motivations

- Online social networks attract tremendous interests from various disciplines.
  - SNA, data mining, business, etc.
- Business practices
  - How the operator provider better services, meet the users' social needs?
  - Probably recommend users form better relationships with others.
- Capture users' attributes, interests and relation in an integrated space and serve them with potentially interesting items.

Outline

- Item recommendation in microblog system
  - Task, dataset, & evaluation metric
- Algorithms predicting item clicks
  - Linear models, reasoning, & statistical models
  - Approaches, strategies, & results
- Conclusion and next step

Goal Description

- KDD CUP 2012 from Tencent
  - which users (or information sources) one user might follow in Tencent Weibo.
- Predicting whether or not a user will follow an item that has been recommended to the user. Items can be persons, organizations, or groups.
- More information:
  - http://www.kddcup2012.org/c/kddcup2012-track1

Dataset

- Supplementary Data Files
  - User_profile (56MB): UserId , Yearofbirth, Gender, Tweets count, Tag-Ids
  - Item (1.2MB): ItemId , ItemCategory, ItemKeyword
  - User_action (217MB): UserId , ActionDestinationUserId, Number- of-at-action, Number-of-retweet, Number-of-comment
  - User_sns (740MB): FollowerUserid, FolloweeUserid
  - User_key_word (182MB): UserId , Keywords
- Training and Testing files
  - Rec_log_train (2GB) & Rec_log_train (900MB): UserId, ItemId, Result, UnixTimestamp

Evaluation Process and Metric

- Average precision at n
  - ap@n = $\Sigma_{k=1,\ldots,n}$ P(k) / (number of items clicked)
    - P(k) the precision at cut-off k in the item list
  - Examples
    - If among the 5 items predicted for a user, the user clicked #1, #3, #4, then ap@3 = (1/1 + 2/3)/3 ≈ 0.56
    - If among the 4 items predicted for a user, the user clicked #1, #2, #4, then ap@3 = (1/1 + 2/2)/3 ≈ 0.67
    - If among the 3 items predicted for a user, the user clicked #1, #3, then ap@3 = (1/1 + 2/3)/2 ≈ 0.83
- Average precision for N users at position n
  - AP@n = $\Sigma_{k=1,\ldots,N}$ ap@n / N

Overview of Our Approaches

- Linear Models for Classification and Regression
- Assumption-based Reasoning
  - Naïve Bayes
  - Probabilistic Relational Models (PRM)
- Statistical Methods
  - Factorization Machines (FM)

Linear Model

- Logistic regression
- $score = \alpha \cdot S_{common\ neighbors} + \beta \cdot S_{fans} + \gamma \cdot S_{catagory} + \theta \cdot S_{acceptance}$
- Feature used
  - #common neighbors, item's # fans, # common categories, & user acceptance rate.
- Best score 0.33

Naïve Bayes

- Conditional probability

$$P(y_i|x)=\frac{P(x|y_i)P(y_i)}{P(x)},$$

x=$\{a_1, a_2, \cdots, a_m\}$ is an unclassified item

Since p(x) is a constant for all item, we have:

$$P(x|y_i)p(y_i)=P(a_1|y_i)P(a_2|y_i)\cdots p(a_m|y_i)P(y_i)$$

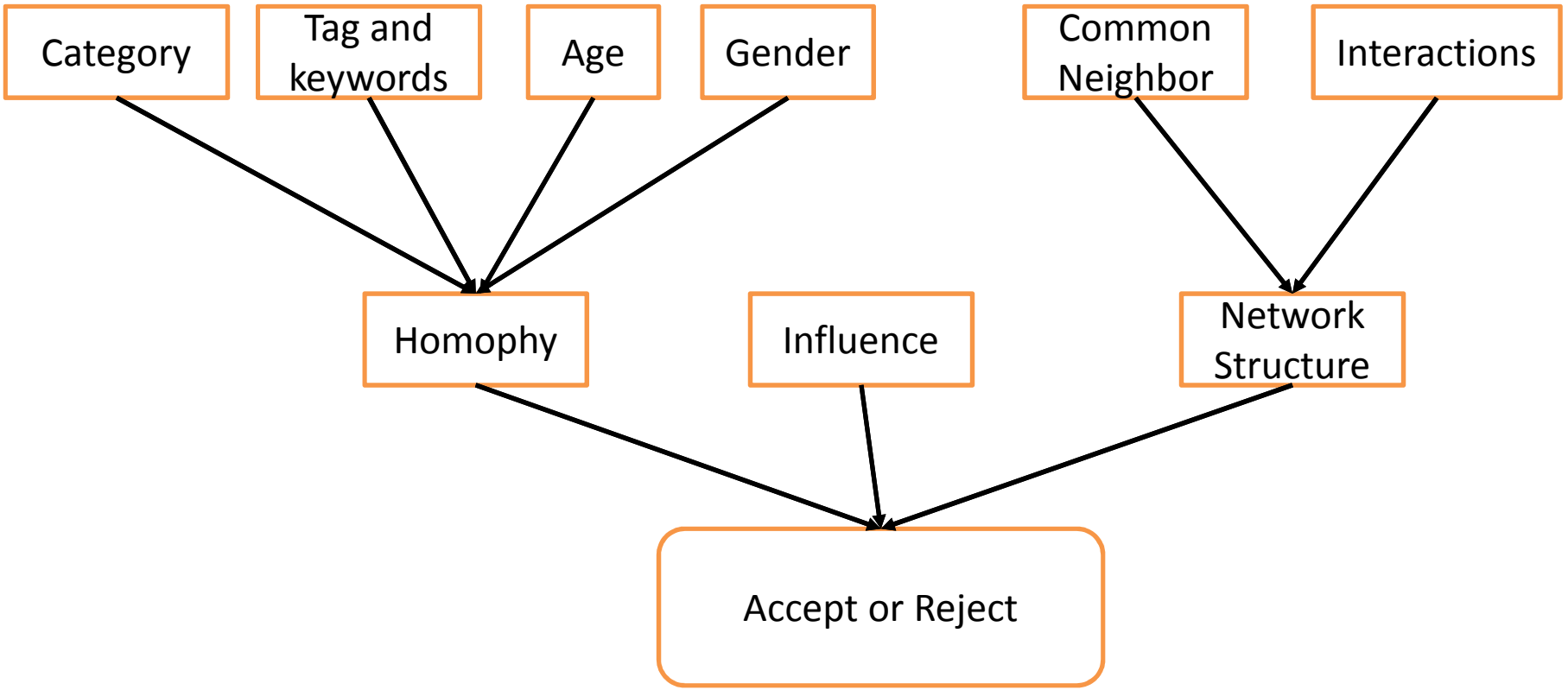
*e.g. P(cn=0|$y_i = 1$)*P(10000<fans<100000|y_i=1)* P(cata_match=3|y_i=1)*P($y_i = 1$)*

- *Best score 0.20*
- *Naïve Bayes assumes that each feature is independent.*
  - *The features' value partition is not calculated by Gaussian distribution, which makes the distribution of features a piecewise function not a Gaussian function.*

Probabilistic Relational Models (PRM)

- Assumptions
  - Users' attributes may affect their choices.
  - Historical behaviors represent the future.
- Five groups of features used
  - User attributes: gender, age, activity
  - Homophylic interests: tag, category
  - Network structure: common neighbors, social preference
  - Influence: based on the items' information
  - Historical interactions: comment, retweet, mention

PRM Model


Category
Tag and keywords
Age
Gender
Common Neighbor
Interactions
Homophy
Influence
Network Structure
Accept or Reject

PRM: Results and Conclusion

- Best score 0.29
- Major problems of PRM
  - Different users have different patterns of choices, thus can not be taken as a whole to form the training set.
  - A better choice is to do “personalized training”, that is to use enough samples to train and predict each user. This won’t work because data sparsity.
  - Another possible way is to do clustering on users to deal with the problem of data sparsity.

Factorization Machines (FM)

- Factorization Machines are a new model class that combines the advantages of Support Vector Machines(SVM) with factorized models.
  - factorization models??
  - FMs model all interactions between variables using factorized parameters, they are able to estimate interaction even in problems with huge sparsity where SVMs fail.
- Features used
  - Previous relationship between users and items
  - Influence of items from rec_log_train
  - Interaction of users and items in user_action
- FM model + linear model, best score 0.35553, 120/677

Factorization Machines (FM)

Model Equation: The model equation for a factorization machine of degree=2 is defined as:

$$\hat{y}(\mathbf{x}) := w_0 + \sum_{i=1}^{n} w_i \, x_i + \sum_{i=1}^{n} \sum_{j=i+1}^{n} \langle \mathbf{v}_i, \mathbf{v}_j \rangle \, x_i \, x_j$$

where the model parameters that have to be estimated are:

$$w_0 \in \mathbb{R}, \quad \mathbf{w} \in \mathbb{R}^n, \quad \mathbf{V} \in \mathbb{R}^{n \times k}$$

The 2-way FM captures all single and pairwise interactions between variables:

- ω_0 is the global bias.
- ω_i models the strength of the i-th variable.
- <v_i,v_j> models the interaction between the i-th and j-th variable ,instead of using an own model parameter $\omega(i,j) \in R$ for each interaction, the FM models the interaction by factorizing it.

Conclusions

- Link prediction based on attribute and relations information
- Use data mining algorithms to study the relation building
- Convert a network graph into a network space, i.e. from relation to vector distance