Other Topics You May Also Agree or Disagree: Modeling Inter-Topic Preferences using Tweets and Matrix Factorization

Akira Sasaki, Kazuaki Hanawa, Naoaki Okazaki, and Kentaro Inui

Graduate School of Information Sciences Tohoku University

1 引言：本文贡献

- 最先进行inter-topic用户偏好的研究
- 可以准确的预测缺失的用户-话题偏好
- 隐向量能成功编码inter-话题偏好

1 引言

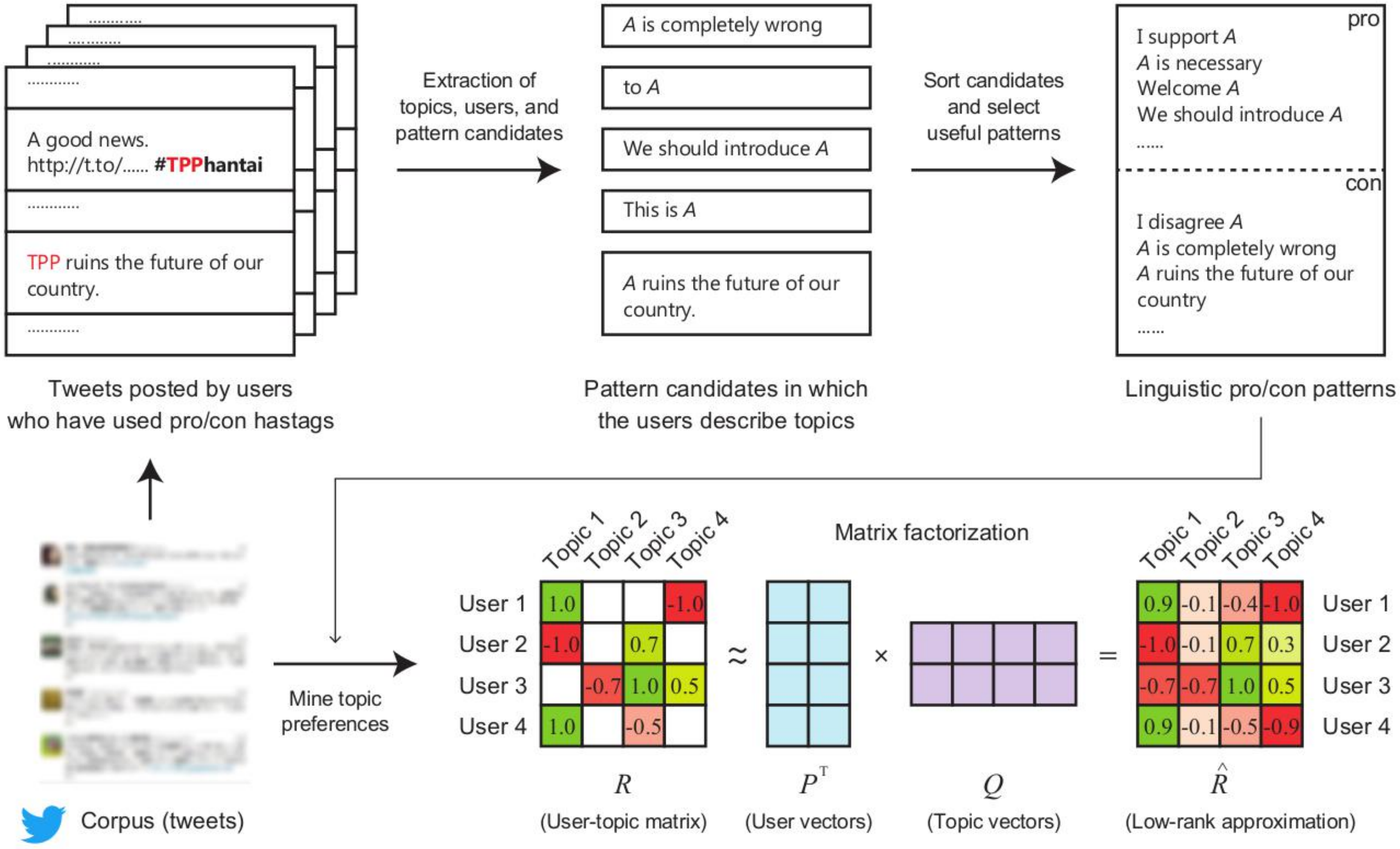

Figure 1: An overview of this study.

2 挖掘用户话题偏好

- 使用语言模式提取元组(u, t, v) ，其中u表示用户，t表示话题，v表示是否同意{1，-1}。

- 当一个推特信息与一条语言模式（如 $\underline{t}$ is necessary）匹配时，认为用户 u 认同话题。

2.1 挖掘语言模式

- 使用预料中含有情感、情绪、讽刺的标签，如 #immigrantWelcome #StopAbortion等。
- 使用正则表达来匹配正负标签，如 “#(.+)sansei” “#(.+)hantai”等。
- 在匹配到标签中的正（负）话题后，从文中挖掘语言模式。
- 由于会出现许多类似“this is A”或“to A”等无用模式，对模式按照使用次数降序排序，对出现次数较多的无用模式人工筛选。

2.2 挖掘话题偏好实例

- 使用语言模式提取实例(u, t, v)。
- When a sentence in a tweet whose author is user u matches one of the pro patterns (e.g., “t is necessary”) and the topic t is included in the set of target topics T , we recognize this as an instance of (u, t, +1).
- When a sentence matches one of the con patterns (e.g., “I don’t want t”) and the topic t is included in the set of target topics T , we recognize this as an instance of (u, t, −1).

3 矩阵分解

- 用户对话题的评分 $$r_{u,t} = \frac{\#(u,t,+1) - \#(u,t,-1)}{\#(u,t,+1) + \#(u,t,-1)}$$

- 目标函数

$$\min_{P,Q} \sum_{(u,t)\in R} \Big((r_{u,t} - \boldsymbol{p}_u{}^\intercal \boldsymbol{q}_t)^2 + \lambda_P \|\boldsymbol{p}_u\|^2 + \lambda_Q \|\boldsymbol{q}_t\|^2 \Big).$$

4 评价：确定隐向量维度参数k

- 使用均方误差 $$RMSE = \sqrt{\frac{\sum_{(u,t)\in R} (\boldsymbol{p}_u{}^\top \boldsymbol{q}_t - r_{u,t})^2}{N}}.$$

- 不同维度对比

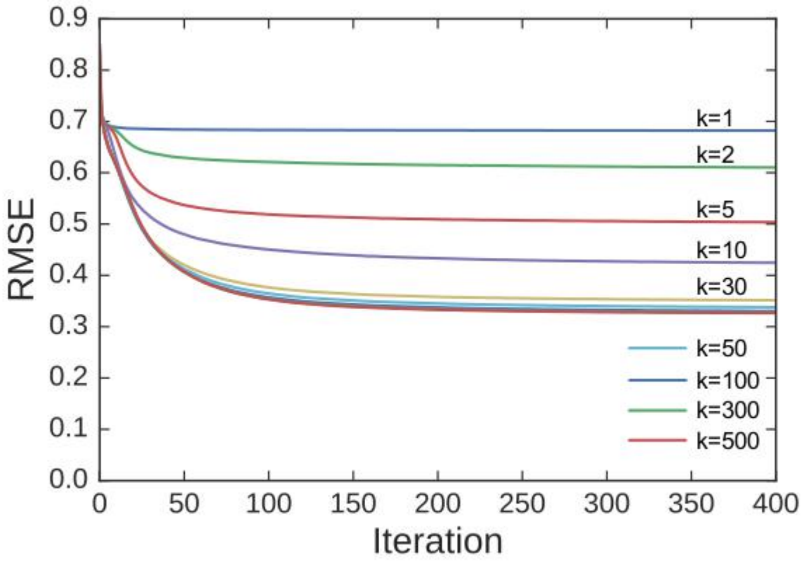

Figure 2: Reconstruction error (RMSE) of matrix factorization with different k.

4 评价：对比Majority baseline

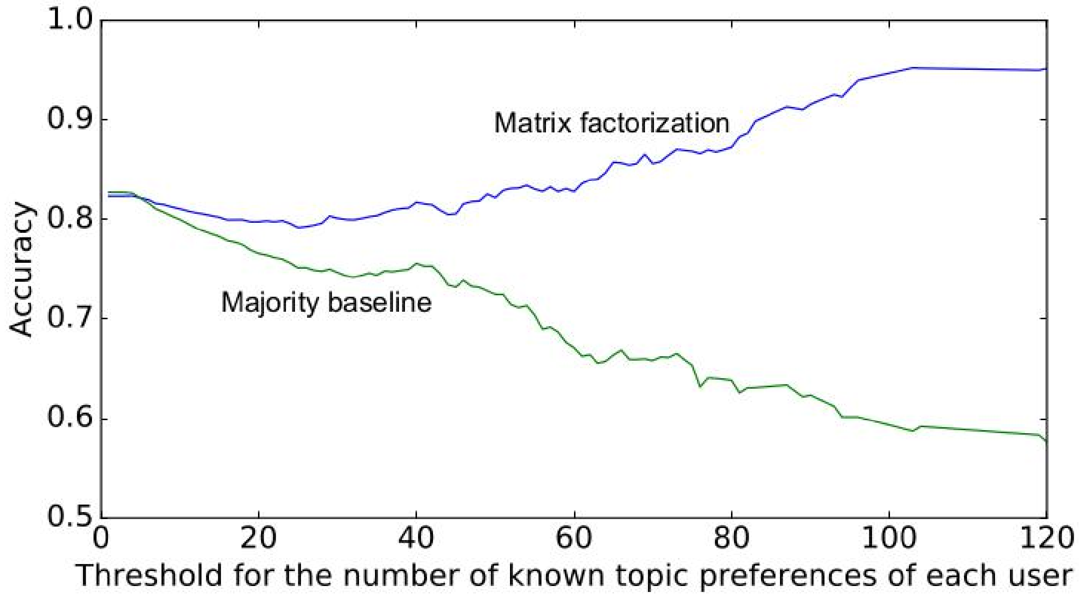

Figure 3: Prediction accuracy when changing the threshold for the number of known topic preferences of each user.

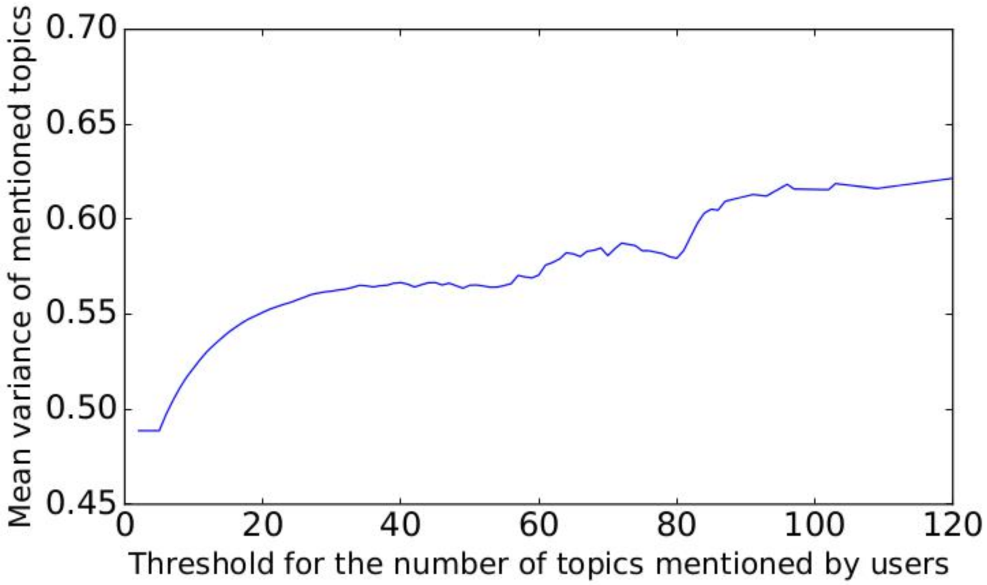

Figure 4: Mean variance of preference values of self-declared topics when changing the threshold for the number of self-declared topics.

5 总结

- 提出了一种建模用户话题偏好的新颖方法。
- 设计了语言模式来识别“支持”和“反对”评论。
- 建立矩阵分解的模型来预测缺失信息。