Does Trust Matter for User Preferences?

A Study on Epinions Ratings

Georgios Pitsilis & Pern Hui Chia

Centre for Quantifiable Quality of Service in Communication Systems (Q2S)

Norwegian University of Science and Technology (NTNU)

IFIPTM 2010, Morioka Japan, 16–18 June

Outline

- Background
- Objective
- Dataset
- Evaluation
- Metrics
- Results
- Conclusions

Background

- Reputation system:
  - Provide global scores about products (user, etc)

Background

- Reputation system:
  - Provide global scores about products (user, etc)
- Recommender system:
  - Provide personalized scores to individual users
  - Employ Collaborative Filtering:
    - Identify influential (similar) users and label them as “neighbors” or “predictors”

Background

- Reputation system:
  - Provide global scores about products (user, etc)
- Recommender system:
  - Provide personalized scores to individual users
  - Employ Collaborative Filtering:
    - Identify influential (similar) users and label them as “neighbors” or “predictors”
- How many predictors are needed?
  - Too many predictors (after a threshold) can reduce prediction accuracy [2]
  - Trick is to select the most suitable predictors

Background (2)

- Similarity value:
  - Measure how closely related two users are based on their preferences towards some common products

Background (2)

- Similarity value:
  - Measure how closely related two users are based on their preferences towards some common products
- Strategies to choose the most suitable predictors:
  - Clustering:
    - Organize the whole population into static groups of similar users

Background (2)

- Similarity value:
  - Measure how closely related two users are based on their preferences towards some common products
- Strategies to choose the most suitable predictors:
  - Clustering:
    - Organize the whole population into static groups of similar users
  - Correlation thresholding:
    - Select predictors who are correlated with a user over a certain threshold

- Similarity value:
  - Measure how closely related two users are based on their preferences towards some common products
- Strategies to choose the most suitable predictors:
  - Clustering:
    - Organize the whole population into static groups of similar users
  - Correlation thresholding:
    - Select predictors who are correlated with a user over a certain threshold
  - Best *k* neighbors:
    - Select *k* most similar neighbors
    - Also known as *k*-Nearest Neighborhood (*k*NN)
    - Outperform correlation thresholding with reasonable *k* values [2]

NTNU
Det skapende universitet

Background (3)

- One may consult people whom he trusts when forming a decision

Background (3)

- One may consult people whom he trusts when forming a decision
- Explicit trust:
  - (Binary) trust value explicitly indicated by user
  - Can form a web of trust

NTNU
Det skapende universitet

Background (3)

- One may consult people whom he trusts when forming a decision
- Explicit trust:
  - (Binary) trust value explicitly indicated by user
  - Can form a web of trust
- Implicit trust:
  - System computed / inferred trust value
    (Eg. Subjective logic [4], Advogato [5], Mole-trust [7])
  - Helpful when explicit trust information is scarce

Background (3)

- One may consult people whom he trusts when forming a decision
- Explicit trust:
  - Binary trust value explicitly indicated by user
  - Can form a web of trust
- Implicit trust:
  - System computed / inferred trust value
    (Eg. Subjective logic [4], Advogato [5], Mole-trust [7])
  - Helpful when explicit trust information is scarce
- Trust-based recommender system:
  - Incorporates trust information into recommendation algorithm
  - Prior research: trust values inferred using various proposed algorithms improve the performance of recommender systems [eg. 1, 6, 7, 8]

Objective

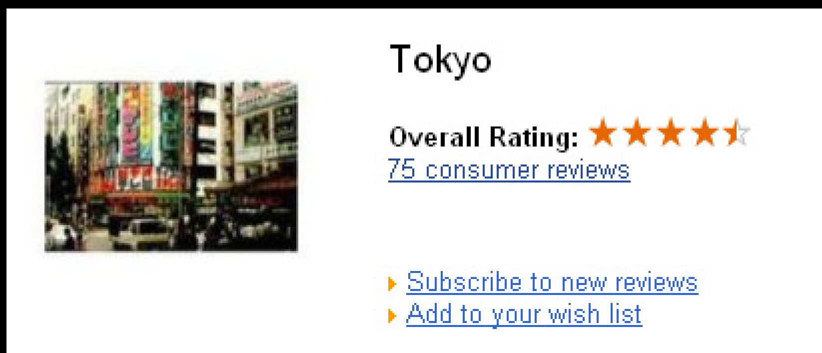
- We do not look into specific trust inference algorithms

Objective

- We do not look into specific trust inference algorithms
- Our work:
  - Explore on using explicit trust to select better predictors
  - Explore on using experience as an additional criterion
  - An experimental study using dataset from *Epinions.com*

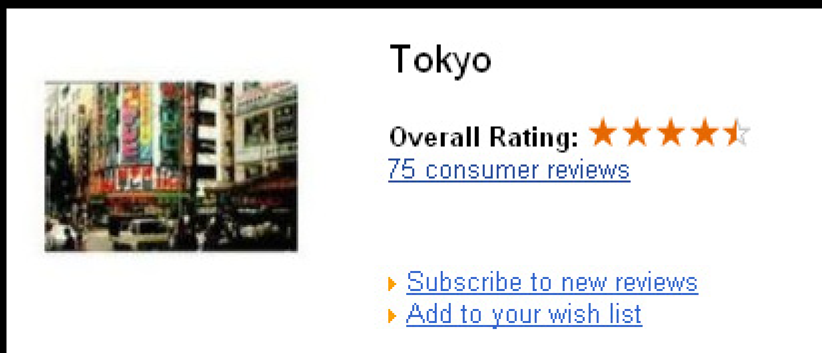
Dataset

- Epinions.com – a reputation system for various products
  - Textual review and rating given in 1 to 5 stars
  - Users indicate explicit trust for other users in the web of trust



Dataset

- Epinions.com – a reputation system for various products
  - Textual review and rating given in 1 to 5 stars
  - Users indicate explicit trust for other users in the web of trust



- Datasets crawled by Paolo Massa [3]
- We use the basic dataset:
  - 664K ratings given by 49K users on 139K products
  - 487K explicit trust links
- Rating & trust links are sparse and has a long-tail

Evaluation

- We use the *k*NN approach & evaluate 3 ways to select the predictor-set:

Evaluation

- We use the *k*NN approach & evaluate 3 ways to select the predictor-set:

  a. Similarity-based (S)

  - Select best predictors based on Pearson's similarity

$$w_{a,b} = \frac{\sum_k (r_{a,k} - \bar{r}_a)(r_{b,k} - \bar{r}_b)}{\sqrt{\sum_k (r_{a,k} - \bar{r}_a)^2 \sum_k (r_{b,k} - \bar{r}_b)^2}}$$

  - Similarity is computable (meaningful) only if both users *a* and *b* have rated some number of common items *k*

Evaluation

- We use the *k*NN approach & evaluate 3 ways to select the predictor-set:

  a. **Similarity-based (S)**
     - Select best predictors based on Pearson's similarity

$$w_{a,b} = \frac{\sum_k (r_{a,k} - \bar{r}_a)(r_{b,k} - \bar{r}_b)}{\sqrt{\sum_k (r_{a,k} - \bar{r}_a)^2 \sum_k (r_{b,k} - \bar{r}_b)^2}}$$

     - Similarity is computable (meaningful) only if both users *a* and *b* have rated some number of common items *k*

  b. **Trust-based (T)**
     - Predictors must also be explicitly trusted

Evaluation

- We use the *k*NN approach & evaluate 3 ways to select the predictor-set:
  a. **Similarity-based (S)**
     - Select best predictors based on Pearson's similarity

$$w_{a,b} = \frac{\sum_k (r_{a,k} - \bar{r}_a)(r_{b,k} - \bar{r}_b)}{\sqrt{\sum_k (r_{a,k} - \bar{r}_a)^2 \sum_k (r_{b,k} - \bar{r}_b)^2}}$$

     - Similarity is computable (meaningful) only if both users *a* and *b* have rated some number of common items *k*
  b. **Trust-based (T)**
     - Predictors must also be explicitly trusted
  c. **Trust-Experience-based (TE)**
     - Predictors must also be explicitly trusted and of an equal or higher experience level
     - Users are categorized into 5 experience levels

Evaluation

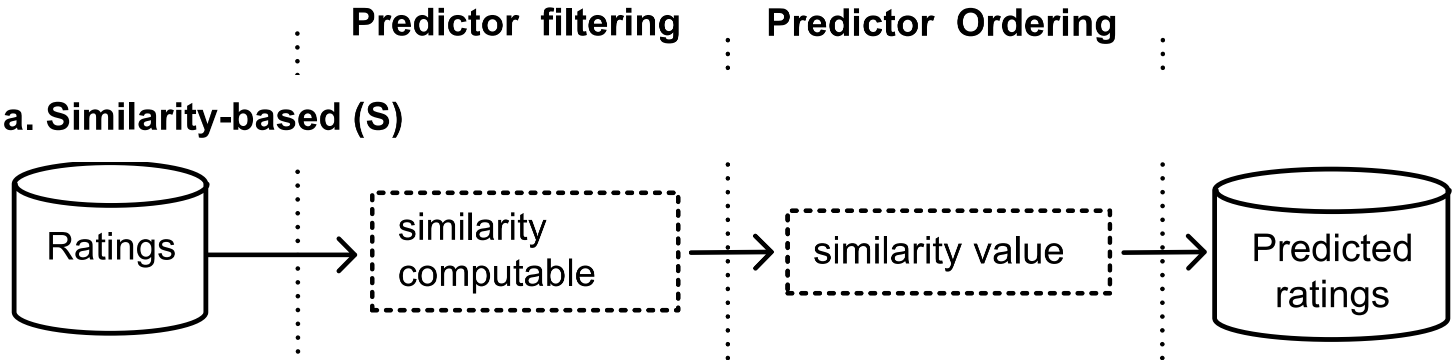
- We use the *k*NN approach & evaluate 3 ways to select the predictor-set:
  a. **Similarity-based (S)**
     - Select best predictors based on Pearson's similarity

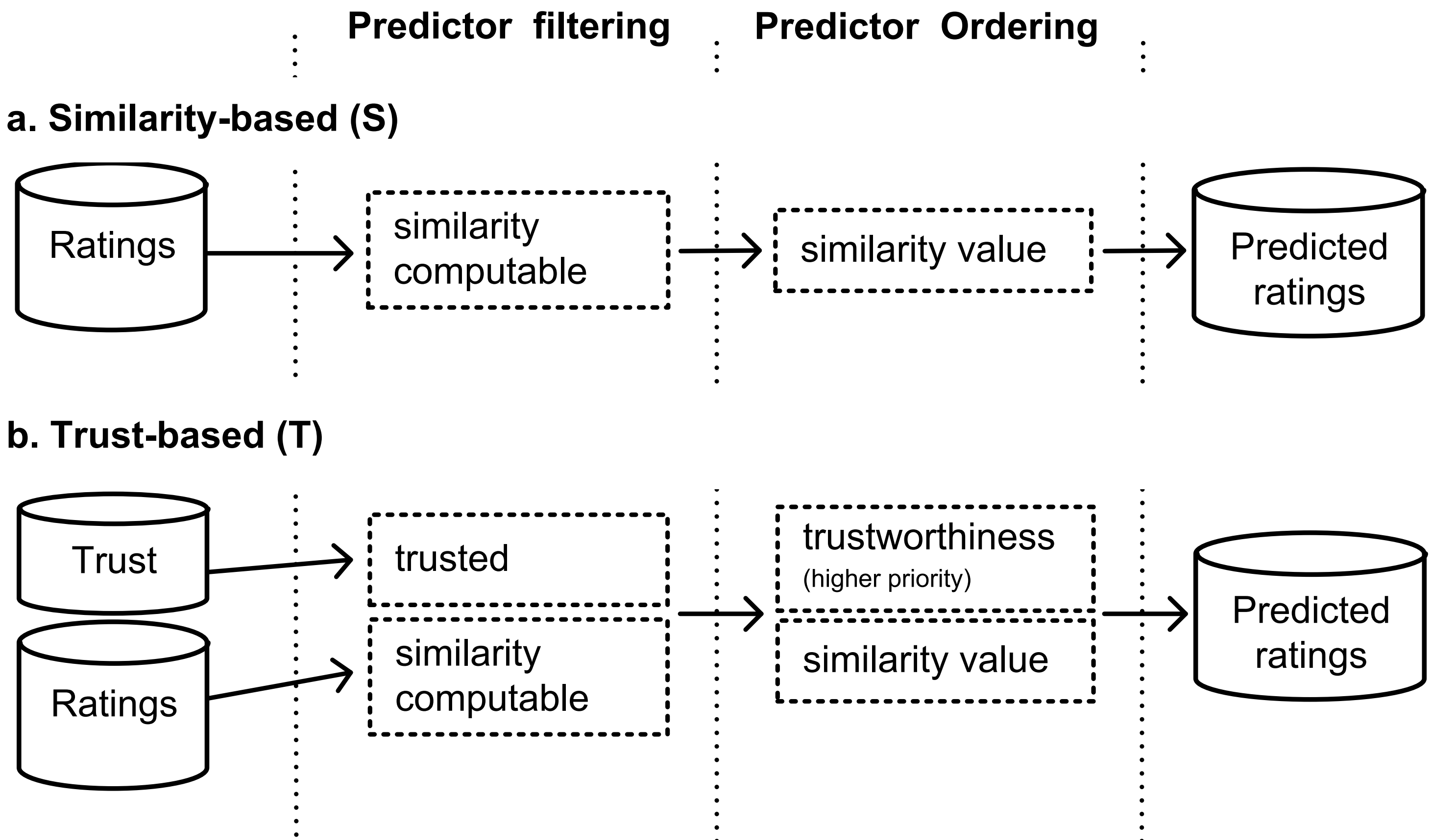
$$w_{a,b} = \frac{\sum_k (r_{a,k} - \bar{r}_a)(r_{b,k} - \bar{r}_b)}{\sqrt{\sum_k (r_{a,k} - \bar{r}_a)^2 \sum_k (r_{b,k} - \bar{r}_b)^2}}$$

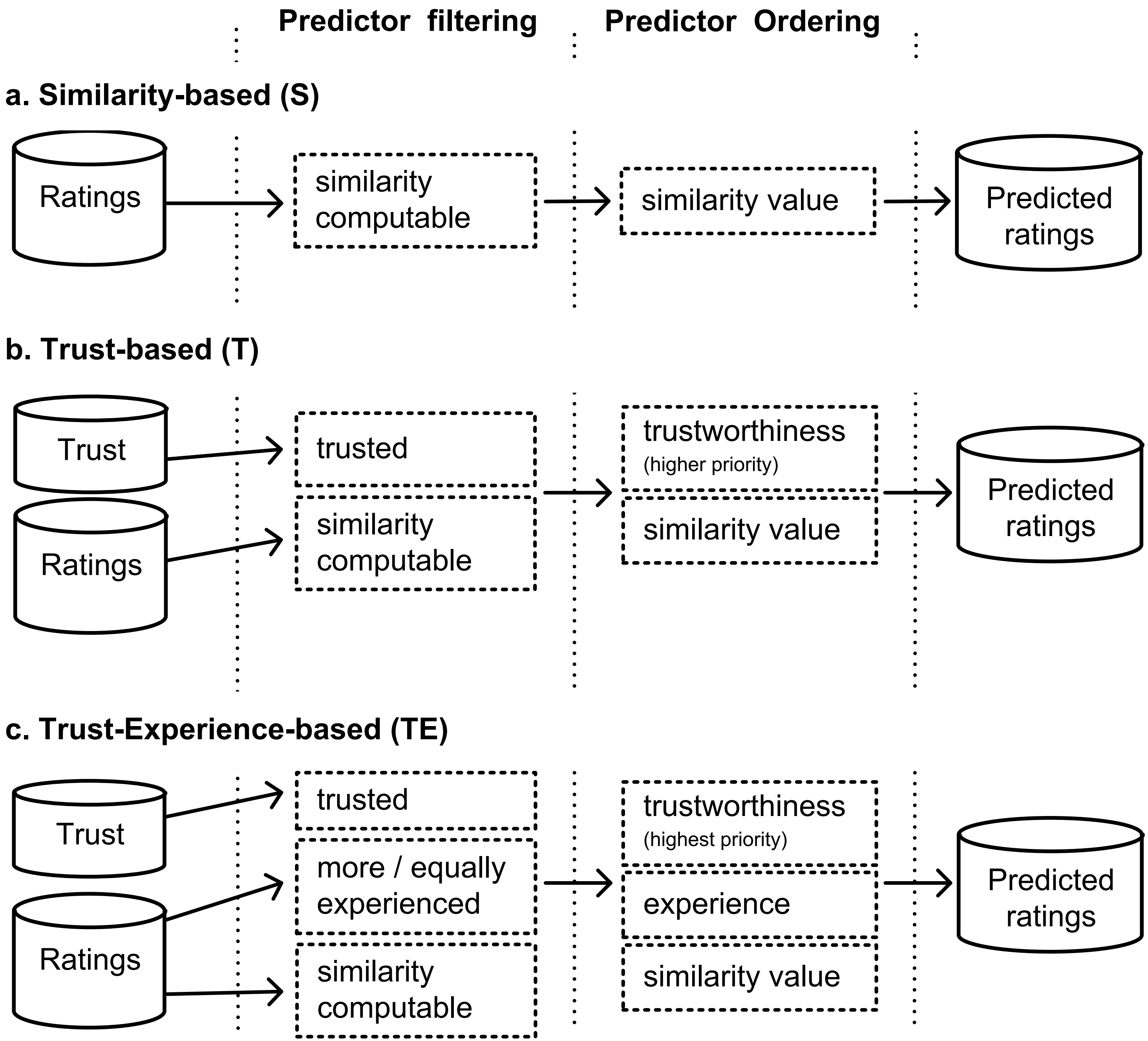
     - Similarity is computable (meaningful) only if both users *a* and *b* have rated some number of common items *k*
  b. **Trust-based (T)**
     - Predictors must also be explicitly trusted
  c. **Trust-Experience-based (TE)**
     - Predictors must also be explicitly trusted and of an equal or higher experience level
     - Users are categorized into 5 experience levels

- Predictors must have rated an item of interest
  – Predictor-set is thus item-dependent and not static (refer to paper for formal desc.)

a. Similarity-based (S)




Predictor filtering
Predictor Ordering
a. Similarity-based (S)
Ratings
similarity computable
similarity value
Predicted ratings
b. Trust-based (T)
Trust
Ratings
trusted
similarity computable
trustworthiness
(higher priority)
similarity value
Predicted ratings


Predictor filtering
Predictor Ordering
a. Similarity-based (S)
Ratings
similarity computable
similarity value
Predicted ratings
b. Trust-based (T)
Trust
Ratings
trusted
similarity computable
trustworthiness
(higher priority)
similarity value
Predicted ratings
c. Trust-Experience-based (TE)
Trust
Ratings
trusted
more / equally experienced
similarity computable
trustworthiness
(highest priority)
experience
similarity value
Predicted ratings

Evaluation (3)

- Rating & trust links are sparse and has a long-tail
  - Evaluate on 1500 users with most number of ratings only
    - To ensure that Pearson's similarity is computable (meaningful)

Evaluation (3)

- Rating & trust links are sparse and has a long-tail
  - Evaluate on 1500 users with most number of ratings only
    - To ensure that Pearson's similarity is computable (meaningful)
- Users are divided into 3 communities (500 each):
  - Most active, Medium active, Least active

Community	Avg Trust link	Avg Rating Count
Most active	40	261
Medium active	9	114
Least active	4	83

Evaluation (3)

- Rating & trust links are sparse and has a long-tail
  - Evaluate on 1500 users with most number of ratings only
    - To ensure that Pearson's similarity is computable (meaningful)
- Users are divided into 3 communities (500 each):
  - Most active, Medium active, Least active

Community	Avg Trust link	Avg Rating Count
Most active	40	261
Medium active	9	114
Least active	4	83

- We perform 5-fold cross-validation:
  - Further divide each community into 5 fifths
  - Repeat 5 times, each time:
    1 fifth as test-set and 4 others as training-sets (where predictors are selected from)
  - Compute the average performance

Evaluation (4)

For each predictor-selection scheme { *S, T, TE* }

For different community { *most, medium, least active* }

For different *k* in [3,13]

Evaluation (4)

For each predictor-selection scheme { *S, T, TE* }

For different community { *most, medium, least active* }

For different *k* in [3,13]

For each of the five-fold cross-validation

Evaluate average performance

Evaluation (4)

For each predictor-selection scheme { *S, T, TE* }

For different community { *most, medium, least active* }

For different *k* in [3,13]

For each of the five-fold cross-validation

For each user *a* in test-set

For each item *i* user *a* is interested in

Choose *k* best predictors

Compute rating prediction

Evaluate average performance

Metrics

- Prediction computed using Resnick's formula

$$\hat{p}_{a,i} = \bar{r}_a + \frac{\sum_b \{w_{a,b} \cdot (r_{b,i} - \bar{r}_b)\}}{\sum_b |w_{a,b}|}$$

  - Predicted ratings rounded to closest integer [1,5]
  - We evaluate also real-value prediction
    - Found no significant difference in results

Metrics

- Prediction computed using Resnick's formula

$$\hat{p}_{a,i} = \bar{r}_a + \frac{\sum_b \{w_{a,b} \cdot (r_{b,i} - \bar{r}_b)\}}{\sum_b |w_{a,b}|}$$

  - Predicted ratings rounded to closest integer [1,5]
  - We evaluate also real-value prediction
    - Found no significant difference in results

- We measure:
  1. Predictive accuracy
     a. Mean Absolute Error *(MAE)*
     b. Root Mean Square Error *(RMSE)*

2. Classification accuracy

An item is of user's interest if he has given it a rating ≥ 4

a. Precision *(P)*

- Relative success in recommending items that are of user's interest

b. Recall *(R)*

- Relative success in retrieving all items of user's interest

	Predicted Value ≥ 4	Predicted Value < 4
User Rating ≥ 4	***TP***	***FN***
User Rating < 4	***FP***	***TN***

$$P = \frac{TP}{TP + FP} \quad R = \frac{TP}{TP + FN}$$

NTNU
Det skapende universitet

2. Classification accuracy

An item is of user's interest if he has given it a rating ≥ 4

a. Precision *(P)*

- Relative success in recommending items that are of user's interest

b. Recall *(R)*

- Relative success in retrieving all items of user's interest

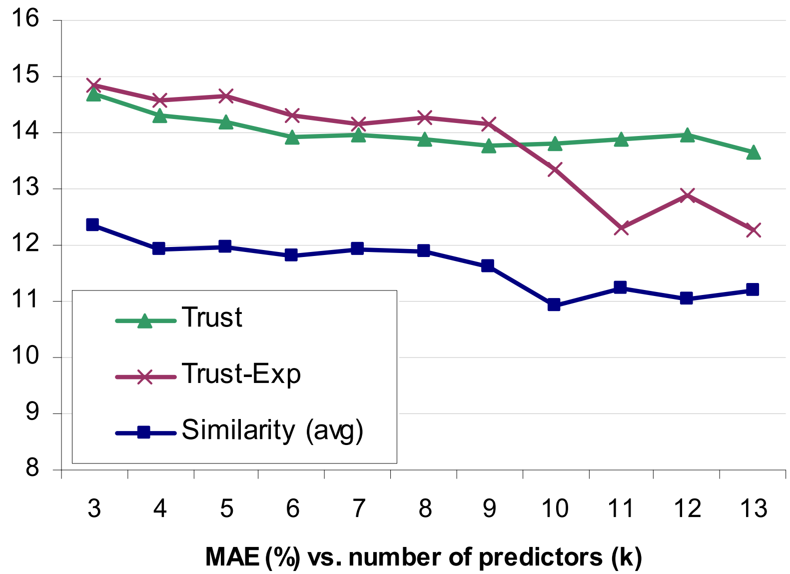
	Predicted Value ≥ 4	Predicted Value < 4
User Rating ≥ 4	*TP*	*FN*
User Rating < 4	*FP*	*TN*

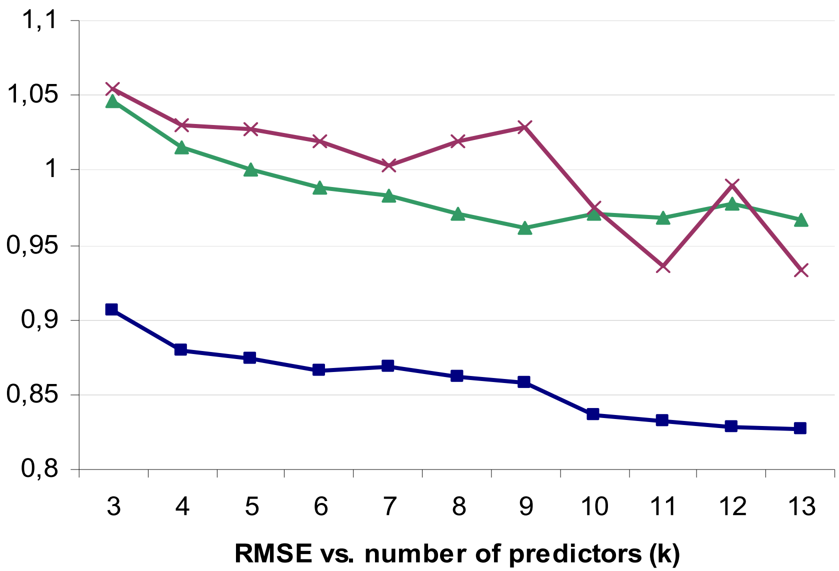
$$P = \frac{TP}{TP + FP} \qquad R = \frac{TP}{TP + FN}$$

3. Coverage

$$C_a = \frac{\text{Number of items the predictors can recommend}}{\text{Total number of items user } a \text{ is interested in}}$$

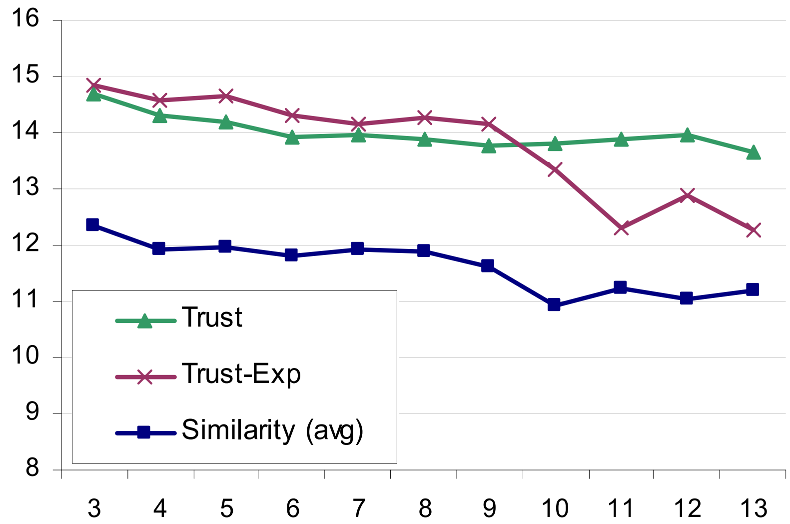
MAE & RMSE (most active)



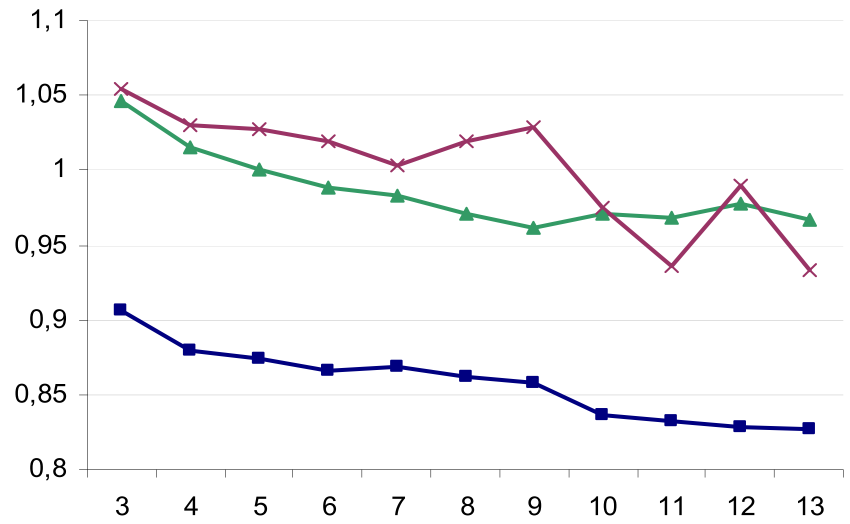


- MAE and RMSE using T and T-E are higher than S

MAE & RMSE (most active)



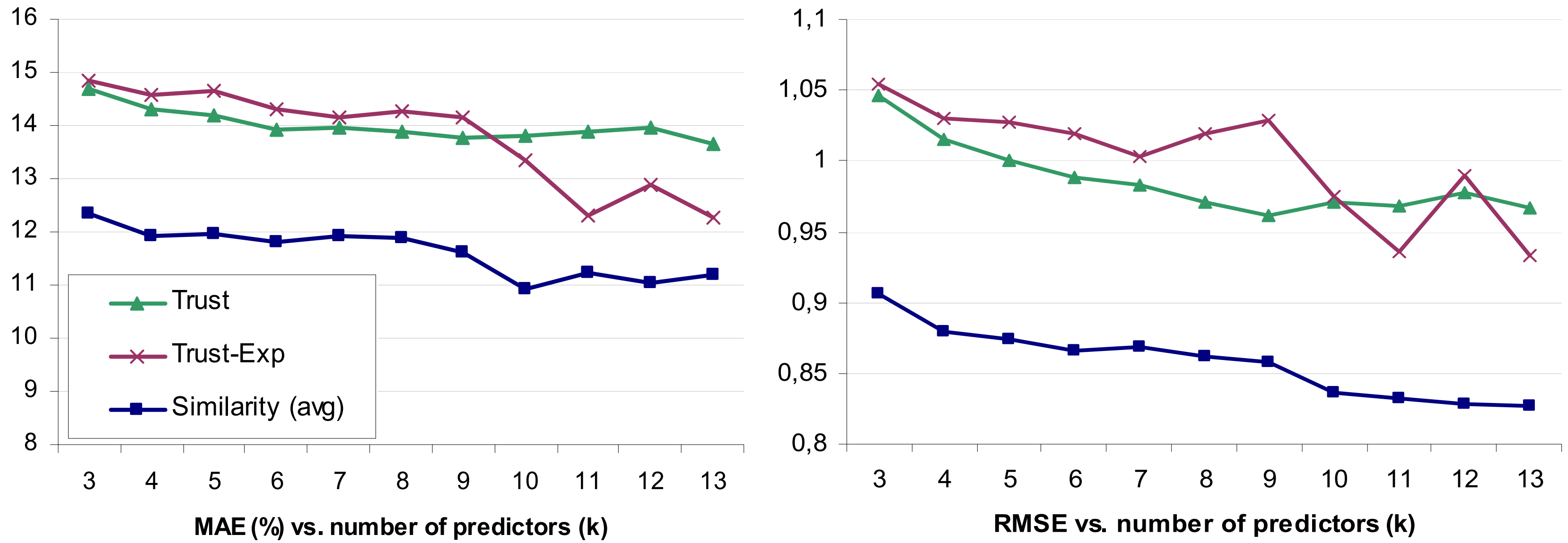
MAE (%) vs. number of predictors (k)



RMSE vs. number of predictors (k)

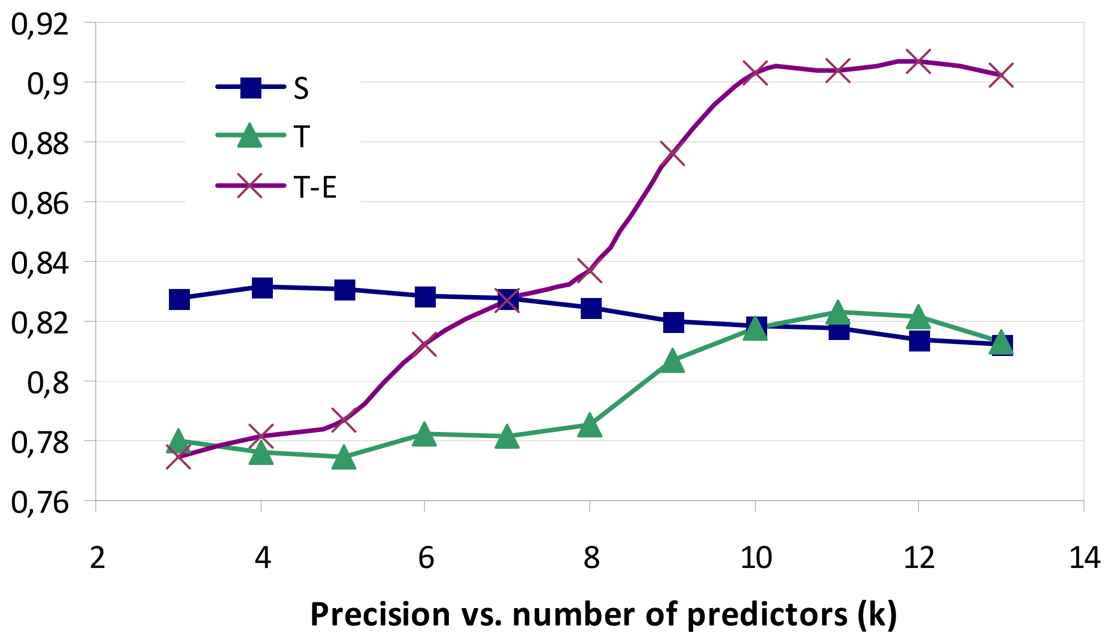
- MAE and RMSE using T and T-E are higher than S
  - Explicit trust does not help to select better predictors
  - Highly-active users may have stronger personal opinions
    - May not rely on or be influenced easily by even those they trust

MAE & RMSE (most active)



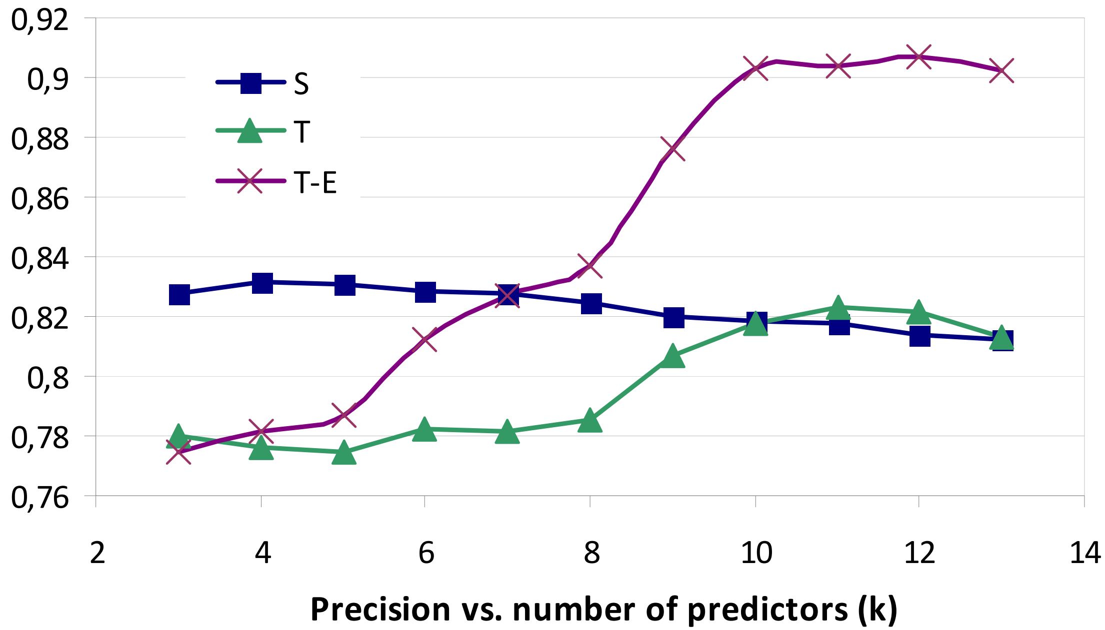
- MAE and RMSE using T and T-E are higher than S
  - Explicit trust does not help to select better predictors
  - Highly-active users may have stronger personal opinions
    - May not rely on or be influenced easily by even those they trust
- Predictive accuracy improves with an increasing number of predictors

Precision (most active)



- Precision improves with TE when there is a large number of trusted and more experienced predictors

Precision (most active)



- Precision improves with TE when there is a large number of trusted and more experienced predictors
  - Opinions from trusted and experienced sources can help to reliably recommend items of user's interest when there are sufficient inputs

Precision, MAE (medium, least active)

- Few trusted neighbors in "medium" and "least active" communities
  - We present evaluation results for number of predictors, *k* = 3,4,5

NTNU
Det skapende universitet

Precision, MAE (medium, least active)

- Few trusted neighbors in “medium” and “least active” communities
  - We present evaluation results for number of predictors, *k* = 3,4,5

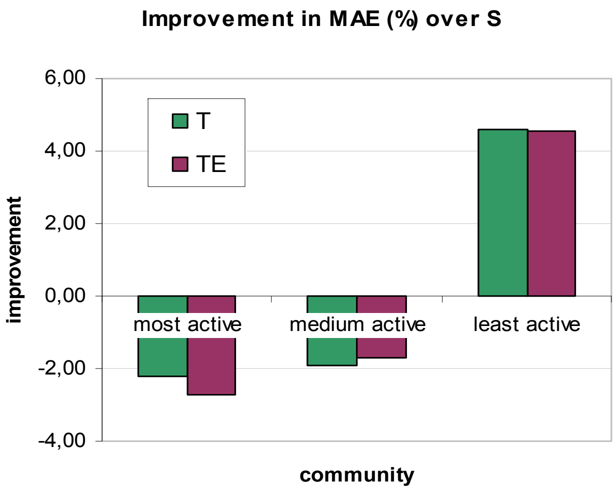
Precision (k=3,4,5 medium active)		
S	**T**	**TE**
0.88	0.91	0.89

- Precision is slightly higher in T and TE schemes

Precision, MAE (medium, least active)

- Few trusted neighbors in “medium” and “least active” communities
  - We present evaluation results for number of predictors, k = 3,4,5

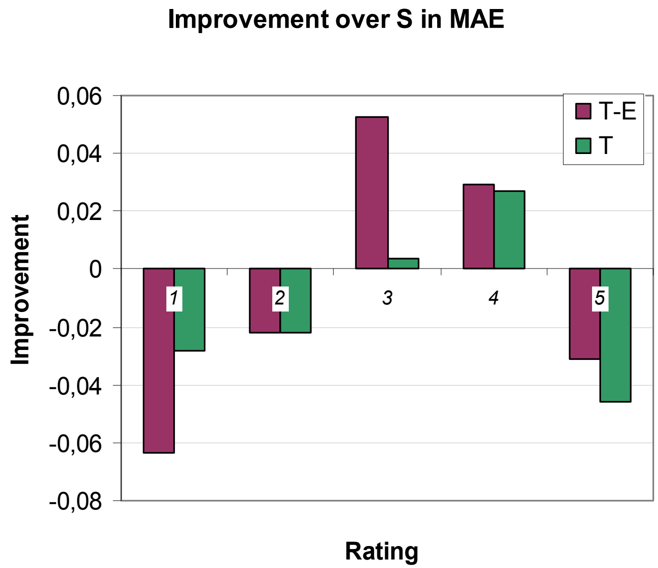
Precision (k=3,4,5 medium active)		
S	**T**	**TE**
0.88	0.91	0.89

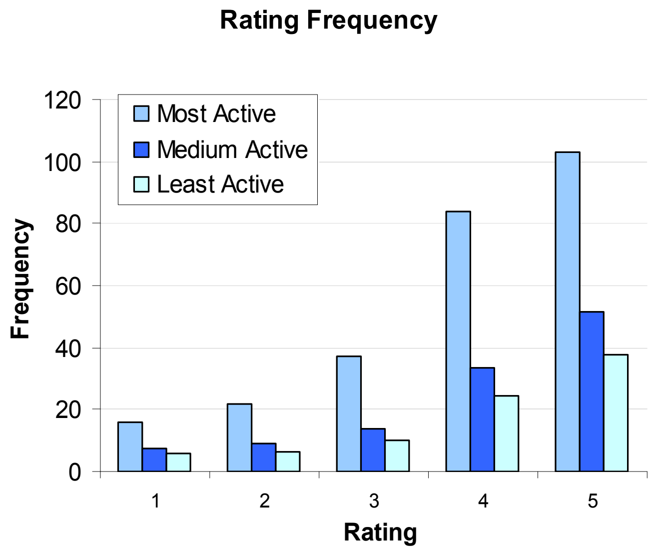


- Precision is slightly higher in T and TE schemes
- MAE with T and TE schemes (compared to S) improve from the “most active” to “least active” community
  - Using trust to select better predictors can be helpful for the less experienced though not for the more experienced users

NTNU
Det skapende universitet

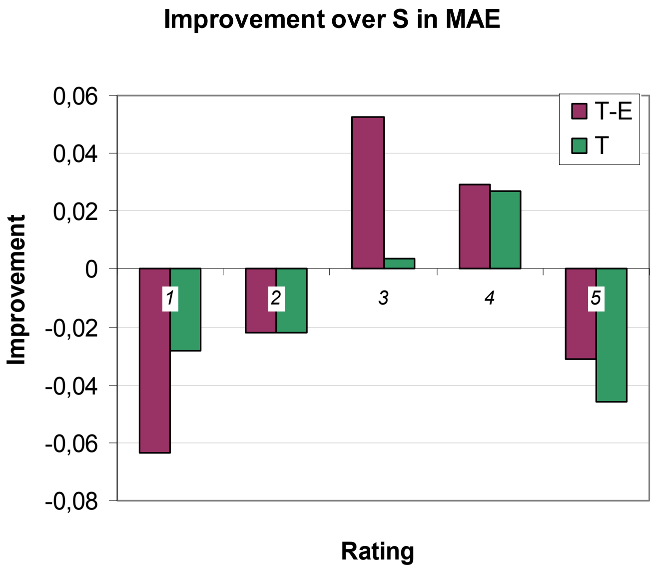
Robustness check

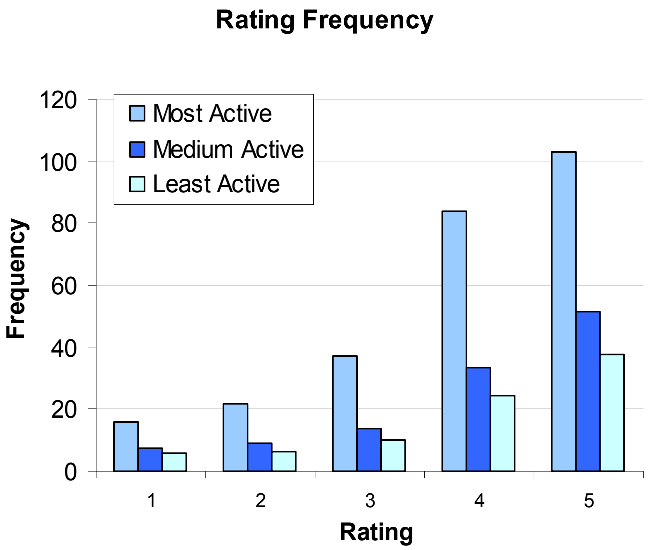




- Better predictive accuracy for less experienced users (excluding the “most active”) is not due to the rating distribution itself

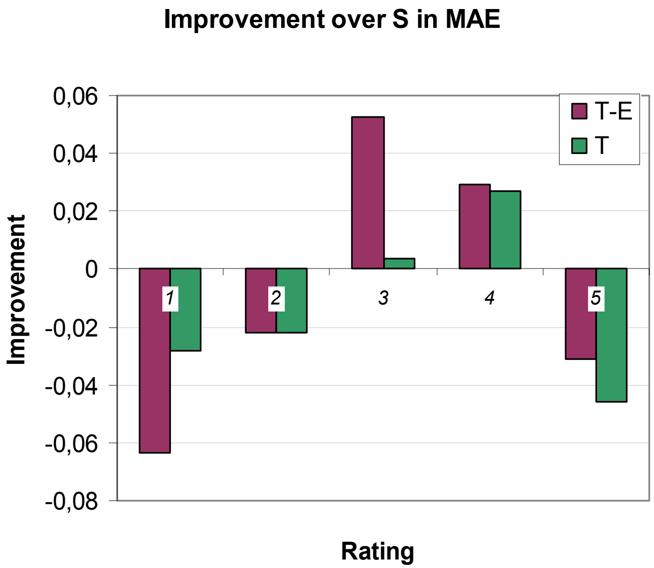
Robustness check

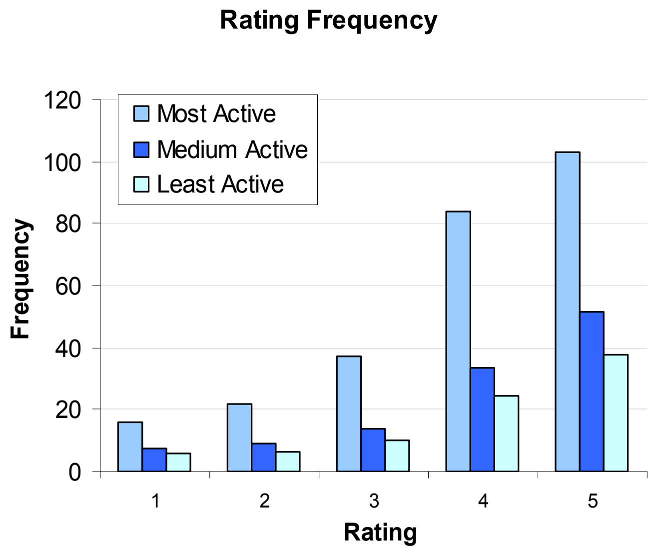




- Better predictive accuracy for less experienced users (excluding the "most active") is not due to the rating distribution itself
- Higher predictive accuracy for rating 3 and 4:

Robustness check





- Better predictive accuracy for less experienced users (excluding the “most active”) is not due to the rating distribution itself
- Higher predictive accuracy for rating 3 and 4:

  Possible explanation:
  - Easier for users to believe in non-extreme ratings and trust these reviewers
  - As more non-extreme reviewers being trusted, predictive accuracy for moderate ratings improves
  - Interesting to explore this from the perspective of behavioral science

Conclusions

- We construct the Trust-based & Trust-Experience-based schemes to study if explicit trust is useful for selecting better predictors

Conclusions

- We construct the Trust-based & Trust-Experience-based schemes to study if explicit trust is useful for selecting better predictors
- Explicit trust can
  - improve predictive accuracy for the less experienced but not for highly active users
  - improve precision to recommend items of users' interest when coupled with the experience criterion

Conclusions

- We construct the Trust-based & Trust-Experience-based schemes to study if explicit trust is useful for selecting better predictors
- Explicit trust can
  - improve predictive accuracy for the less experienced but not for highly active users
  - improve precision to recommend items of users' interest when coupled with the experience criterion
- Coverage is strongly affected due to the scarce trust links
  - Algorithm to infer trust would be important

Conclusions

- We construct the Trust-based & Trust-Experience-based schemes to study if explicit trust is useful for selecting better predictors
- Explicit trust can
  - improve predictive accuracy for the less experienced but not for highly active users
  - improve precision to recommend items of users' interest when coupled with the experience criterion
- Coverage is strongly affected due to the scarce trust links
  - Algorithm to infer trust would be important
- Potential future work: to learn from psychology and behavioral science

NTNU
Det skapende universitet

Reference

1. Massa, P., Avesani, P.: Trust-aware recommender systems. In: Proc. RecSys 2007 (2007)
2. Herlocker, J.L., Konstan, J.A., Borchers, A., Riedl, J.: An algorithmic framework for performing collaborative filtering. In: Proc. 22nd ACM SIGIR 1999 Conference on Research and Development in Information Retrieval (1999)
3. A cooperative environment for the scientific research of trust metrics on social networks, http://www.trustlet.org/
4. Jøsang, A.: A Logic for Uncertain probabilities. International Journal of Uncertainty, Fuzziness & Knowledge Based Systems 9(3) (2001)
5. Levien, R.: Advogato’s trust metric, White Paper (2009), http://www.advogato.org/trust-metric.html
6. Lathia, N., Hailes, S., Capra, L.: Trust-Based Collaborative Filtering. In: Karabulut, Y., Mitchell, J., Herman, P., Jensen, C.D. (eds.) IFIP, Trust Management II, Trondheim, Norway (2008)
7. Massa, P., Avesani, P.: Controversial Users demand Local Trust Metrics: an Experimental Study on Epinions.com Community. In: Proc. of 20th National conf. AAAI 2005 (2005)
8. O’Donovan, J., Smyth, B.: Trust No One: Evaluating Trust-based Filtering for Recommenders. In: Proc. of the 19th International Joint Conference on Artificial Intelligence IJCAI, (2005)

NTNU
Det skapende universitet

Thank you. Questions?

Georgios Pitsilis
pitsilis@q2s.ntnu.no

<u>Pern Hui Chia</u>
chia@q2s.ntnu.no
www.q2s.ntnu.no/people/chia/

NTNU
Det skapende universitet