Trust in Recommender Systems

John O’Donovan & Barry Smyth
Overview

Recommender Systems, circa 2005

On Trust in Recommender System

Beyond Trust → Trust, Reputation, Social, Explanations

Conclusions & Wrapup
Remembering 2005 ...
Top Movie

STAR WARS
EPISODE III
REVENGE OF THE SITH
Recommender Systems (c. 2005)

(Single-Shot) Collaborative vs Content

Relevance & Accuracy vs Diversity & Novelty

Movies, Books, Music (Atomic Items)
Correlations between the ratings patterns of users denote user similarity ...

People like you have also liked ...
Conversely, **correlations** between the ratings patterns of **items** denote **item similarity** ...

*If you liked X then you might like Y*...
The Rise of the Social Web


The Promise of the Social Graph

Similarity vs Relationship

Trust & Reputation
Trust me I’m a Recommender

User Similarity ~ Ratings Correlations

Is Similarity Sufficient?

Trust, Reputation, Expertise
Modeling Trust

Consider two types of users in a recommendation session.

Producer
Users (profiles) selected to participate in the recommendation session.

Consumer
The user profile that is receiving the item rating from the producer profiles.
Correct Predictions

We say that a ratings prediction for an item, $i$, by a producer $p$ for a consumer $c$, is correct if the predicted rating, $p(i)$, is within $\epsilon$ of $c$’s actual rating $c(i)$.
Calculating Trust Scores from Ratings Data

To evaluate the correctness of $p$’s recommendation we separately perform recommendation by using $p$ as $c$’s sole recommendation partner.

We say a trust score for item $i_1$ is generated for producer $b$ by using the information in profile $b$ only to generate predictions for each consumer profile.
Binary Trust Scores

\[ T_p(i, c) = \text{Correct}(i, p, c) \]

The trust score for producer \( p \) is simply whether or not the prediction for item \( i \) is correct.
RecSet is the set of recommendations that a producer $p$ has participated in ... that is the consumer-item pairs for which ratings have been generated.

CorrectSet is the subset of these for which correct predictions have been generated (as per above).
Profile Level Trust

The percentage of overall recommendations that the producer has correctly contributed.

\[
Trust^P(p) = \frac{|CorrectSet(p)|}{|RecSet(p)|}
\]

Coarse-grained trust metric; some users may be better at making predictions for certain types of items which will not be captured by this profile-level metric.
Item-Level Trust

The percentage of $p$’s recommendations for item $i$ that have proven to be correct.

$$\text{Trust}^I_{(p, i)} = \frac{|\{(c_k, i_k) \in \text{CorrectSet}(p) : i_k = i\}|}{|\{(c_k, i_k) \in \text{RecSet}(p) : i_k = i\}|}$$

For example, if user $p$ has been involved in 20 recommendation sessions for item $i$ and only 5 of these have been correct, then $p$’s item-level trust will be 0.25.
Adding Trust to Recommendation
Trust-Based Weighting

The basic idea with trust-based weighting is to combine the trust and similarity at prediction time.

\[
c(i) = \bar{c} + \frac{\sum_{p \in P(i)} (p(i) - \bar{p})w(c, p, i)}{\sum_{p \in P(i)} |w(c, p, i)|}
\]

\[
w(c, p, i) = \frac{2(sim(c, p))(trust^I(p, i))}{sim(c, p) + trust^I(p, i)}
\]

Standard Resnick Prediction

New weighting formula combining conventional similarity and trust.

(harmonic mean)
Trust-based Filtering

Alternatively …

… eliminate untrustworthy neighbours from consideration at prediction time.

Only consider users with a trust score > T.

\[
c(i) = \bar{c} + \frac{\sum_{p \in P^T(i)} (p(i) - \bar{p}) \text{sim}(c, p)}{\sum_{p \in P^T(i)} |\text{sim}(c, p)|}
\]

\[P^T_i = \{p \in P(i) : \text{Trust}^I(p, i) > T\} \]
Evaluation

(MovieLens)
Recommendation Strategies

Standard CF (Resnick)

Profile Level Trust
- WProfile (weighting)
- FProfile (filtering)
- CProfile \((w + f)\)

Item Level Trust
- WItem (weighting)
- FItem (filtering)
- CItem \((w + f)\)
MAE Results

![Graph showing MAE Results]
TO BE CONTINUED....
Google Scholar Citations
A few things about trust...

- Context-dependent
- Subjective
- Dynamic
Information Retrieval

Visualization and Interaction

Cognition and Understanding

Trust
Computational Models
Trust-aware Recommender Systems (Massa 2007)

Figure 1: Trust-Aware Recommender System Architecture.

Figure 2: MAE on cold start users for some representative algorithms.

Figure 3: Ratings coverage on cold start users for some representative algorithms.
FilmTrust (Golbeck & Hendler, ’05, ’06, ‘07)
Gradual Trust and Distrust in Recommender Systems (Victor et al. 2009)

Fig. 4. Trust score space $\mathcal{BL}$

Models, State Of The Art

<table>
<thead>
<tr>
<th></th>
<th>trust only</th>
<th>trust and distrust</th>
</tr>
</thead>
<tbody>
<tr>
<td>probabilistic</td>
<td>Kamvar et al. [14]</td>
<td>Jøsang et al. [13]</td>
</tr>
<tr>
<td></td>
<td>Richardson et al. [21]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Zaihrayeu et al. [27]</td>
<td></td>
</tr>
<tr>
<td>gradual</td>
<td>Abdul-Rahman et al. [1]</td>
<td>De Cock et al. [7]</td>
</tr>
<tr>
<td></td>
<td>Falcone et al. [17]</td>
<td>Guha et al. [12]</td>
</tr>
<tr>
<td></td>
<td>Almenárez et al. [2]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Tang et al. [25]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Golbeck [11]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Massa et al. [15]</td>
<td></td>
</tr>
</tbody>
</table>

$\leq_k$ Completely inconsistent $(1,1)$

$\leq_t$ Very reliable $(1,0.2)$

Complete Distrust $(0,1)$

(0.5,0.5) Complete trust $(1,0)$

Ignorance $(0,0)$
Sequence-based Trust in Recommender Systems (Liu et al. 2011)

\[ \hat{P}_{c,d} = \bar{r}_c + \frac{\sum_{p \in NS} H(UPS(c,p), T_{c,p}) (r_{p,d} - \bar{r}_p)}{\sum_{p \in NS} |H(UPS(c,p), T_{c,p})|}, \]

\[ H(UPS(c,p), Trust^H_{c,p}) = \frac{2(UPS(c,p))(Trust^H_{c,p})}{UPS(c,p) + Trust^H_{c,p}}. \]

Fig. 2. Illustration of sequence-based trust with time factor.
Comparative Recommender System Evaluation: Benchmarking Recommendation Frameworks.[Bellogin’14]

Table 2. Comparing the inferred trust scores (implicit) with the ground trust scores (explicit)

<table>
<thead>
<tr>
<th>Trust metric</th>
<th>nDCG@10</th>
<th>nDCG</th>
<th>P@10</th>
<th>R@10</th>
<th>MRR</th>
<th>Cvg</th>
</tr>
</thead>
<tbody>
<tr>
<td>O'Donovan &amp; Smyth [9] (TM1)</td>
<td>0.007</td>
<td>0.008</td>
<td>0.007</td>
<td>0.001</td>
<td>0.022</td>
<td>98.8%</td>
</tr>
<tr>
<td>Lathia et al. [7] (TM2)</td>
<td>0.004</td>
<td>0.008</td>
<td>0.004</td>
<td>0.001</td>
<td>0.014</td>
<td>99.7%</td>
</tr>
<tr>
<td>Hwang &amp; Chen [4] (TM3)</td>
<td>0.006</td>
<td>0.009</td>
<td>0.005</td>
<td>0.001</td>
<td>0.020</td>
<td>100%</td>
</tr>
<tr>
<td>Shambour &amp; Lu [12] (TM4)</td>
<td>0.006</td>
<td>0.009</td>
<td>0.005</td>
<td>0.001</td>
<td>0.017</td>
<td>100%</td>
</tr>
<tr>
<td>Papagelis et al. [10] (TM5)</td>
<td>0.028</td>
<td>0.007</td>
<td>0.024</td>
<td>0.003</td>
<td>0.071</td>
<td>9.5%</td>
</tr>
</tbody>
</table>

Table 3. Performance comparison of the SocialMF using implicit trust against the baselines (the lower, the better); lowest values for each $k$ in bold face and best values underlined.

<table>
<thead>
<tr>
<th>Method/k</th>
<th>RMSE</th>
<th>MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>$k=5$</td>
<td>$k=10$</td>
<td>$k=5$</td>
</tr>
<tr>
<td>PMF</td>
<td>1.1741</td>
<td>1.1705</td>
</tr>
<tr>
<td>SocialMF-explicit trust</td>
<td>1.0956</td>
<td>1.0934</td>
</tr>
<tr>
<td>SocialMF-TM1: O'Donovan &amp; Smyth [9]</td>
<td>$\underline{1.0926}$</td>
<td>1.1003</td>
</tr>
<tr>
<td>SocialMF-TM2: Lathia et al. [7]</td>
<td>1.0968</td>
<td>1.1005</td>
</tr>
<tr>
<td>SocialMF-TM3: Hwang &amp; Chen [4]</td>
<td>1.0947</td>
<td>1.1006</td>
</tr>
<tr>
<td>SocialMF-TM4: Shambour &amp; Lu [12]</td>
<td>1.0952</td>
<td>1.0990</td>
</tr>
<tr>
<td>SocialMF-TM5: Papagelis et al. [10]</td>
<td>1.0970</td>
<td>1.1065</td>
</tr>
</tbody>
</table>

Table 3 presents the results of comparing the SocialMF on implicit trust scores, explicit trust scores, and PMF. Based on the results, all SocialMFs that incorporates implicit trust outperforms the PMF; the largest difference is 8.2% and smallest difference is 7.7%.
that accompanies the recommended items. The variations among the recommended items (here, aggregated star rating of the movies) according to the genre of the movies), use of a human model (here, group-based profiling), and factors are described in Table 1.

Table 1. Dimensions, factors, and debrief texts.

<table>
<thead>
<tr>
<th>Factor</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Star</td>
<td>This recommender favours movies popular with the user, based on the aggregated recommendations of similar users.</td>
</tr>
<tr>
<td>Quality</td>
<td>This recommender focuses on movies with high quality, based on the aggregated recommendations of similar users.</td>
</tr>
<tr>
<td>Genre</td>
<td>This recommender recommends movies in the user's favorite genres.</td>
</tr>
</tbody>
</table>

When reviewing these findings, we hypothesise that the changes in the content presentation phase, the explanations used to tailor the personalised explanations, and the priority of the constructs of trust are due to the identical presence of new high-extraversion users. As earlier, statistically significant differences are marked by *. The changes can be attributed to the identical presentation of the three recommendation lists, the content perception towards all trust constructs, during the ranking phase.

Figure 1. Presentation dimension: genre (A), human (B), star (C).

Figure 5. Low- vs high-extraversion users: presentation (top), explanation (middle), and priority (bottom). Significant differences are marked by *.
Abstract Approaches with Human-in-the-loop
Trust, Cognition and UI in the Diner’s Dilemma Problem (Onal et al. IEEE CogSima, best paper 2014)
Dynamics of Human Trust in Recommender Systems (Harman et al. 2014)
Yu et al. IUI’17: User Trust Dynamics: An Investigation Driven by Differences in System Performance

Figure 2: Upon decision from the user, the actual glass is shown and fictitious earnings updated.

Figure 4. Implications of system failures and successes for user trust change.
Figure 2. We designed three visualizations for autonomous driving cars. *Left:* A world in miniature, *Middle:* A chauffeur avatar, *Right:* Basic car indicators as the baseline. The world in miniature and the chauffeur avatar are intelligent autopilot visualizations that interpret the current driving situation and react accordingly. The car indicator visualization only visualizes the basic turn intentions and does not show any intelligence. This corresponds to what drivers currently see in cars. We ensured that the indicators light up and sound for the same duration in all autopilot variants.

We measured the participants' attitudes and behaviors by several questionnaires (see section Questionnaires). In general, our participants seemed to be safe drivers. However, many of them would cross an intersection although the traffic light is about to turn red and exceed the speed limit for up to 15 km/h but not 30 km/h. They would not do risky overtaking maneuvers or race for fun or out of boredom.

Overall, our participants trust autonomous systems until they fail or show deficits in competence. They are also confident that autonomous systems have a high competence. In addition, all of our participants are thrilled about and like to use novel technologies. Comparing the attitude towards autonomous driving of our participants with the results of Kyriakidis et al. [19], we found that our participants were slightly more fascinated about fully autonomous driving. In both studies, autonomous and manual driving are rated as equally comfortable. We therefore think that our set of participants is representative but, probably due to the technical background of many of them, slightly positively biased about autonomous driving.

**Study Design**

We designed a within-subjects experiment with 3 x 3 conditions: We counterbalanced three driving videos and three autopilot visualizations (chauffeur, world in miniature, indicators) throughout the study, resulting in three test phases. We defined six groups that represent the presentation orders of the autopilot visualization (at fixed order of the driving videos) and randomly assigned five participants to each group.

The driving videos were recorded by means of a GoPro HERO 4 (1280 x 960 px resolution) which was placed inside the car below the rear view mirror. We drove the same track several times during different times of day and weather conditions to obtain similar but not identical footage. We then cut three driving videos with 7 min each out of two different recordings: The first part shows urban two-lane roads with little to medium traffic density. It contains many maneuvers such as stopping at a traffic light and turns and situations such as crossing pedestrians and overtaking cyclists. The second part shows high-density traffic on an urban multi-lane road with a lot of lane change traffic and many stops at traffic lights. It also contains one unpredictable event that requires fast action such as a suddenly stopping lead car.

For each driving video, we then designed corresponding simulations for each autopilot visualization. The visualizations chauffeur and world in miniature present their intelligence by their understanding of and reaction to objects, events and situations. To ensure comparability, both visualizations react to the same events within each driving video. The baseline visualization only communicates the car’s turn intentions by means of the indicators; this visualization does not convey intelligence or understanding of the situation.

In order to create a high feeling of realism, we used the videos of real driving along with a real car test setup. Participants even started the automated drive themselves by pressing a button on a smartphone next to the steering wheel, which also aimed to increase the feeling of control and the interactivity. In order to keep participants involved in the situation and prevent distraction, the experimenter frequently asked the participant to judge the current feeling of trust on a scale from 1 to 10 (low to high trust). Overall, the study lasted about 75 min.

**Questionnaires**

Since we investigated subjective feelings, we based our evaluation on questionnaires. Prior to the test, we used introductory questionnaires to gather demographic data as well as to get insights into the participants' attitudes and tendencies to trust people or systems. After each test phase, participants had to fill out the intermediate questionnaires to report their feeling of trust in autonomous driving supported by one particular autopilot visualization. After the last test phase and its intermediate questionnaire, the participants had to fill out a closing questionnaire which compares the three visualizations directly.
Human-in-the-loop Recommenders
PeerChooser [CHI 08]
Trust Through Explanation in Recommenders: A Brief History

1995
- Collaborative Filtering of Netnews (Resnick et al)
- Trust in RecSys [IUI’05]
- Trust in RecSys [IUI’05]

2000
- Google PageRank
- Herlocker’s work on Explanations for Recommenders
- Tintarev & Mashhoff on Explaining Recommendations
- TopicNets [ACM TIST ’10]

2005
- BellKor’s Pragmatic Chaos wins the 1M Netflix Prize
- Garcia-Molina Keynote on Explanation and Control (ACM RecSys)
- Ed Chi’s Keynote on Exp. and Control (ACM IUI, March 2016)
- Xavier Amatrain Keynote on Exp. and Control (ACM IUI, March 2016)

2010
- Amazon’s: People also bought...
- Google Personalized Search
- 1st ACM Conference on Recommender Systems
- 1st ACM Workshop on Interfaces for Recommender Systems
- Tintarev & Masthoff on Explaining Recommendations
- LinkedVis [IUI’13]
- SmallWorlds [EuroVis’10]
- WiGis Framework [GraphVis 2010]

2015
- Amazon’s: People also bought...
- Google Personalized Search
- Amazon’s: People also bought...
- Google Personalized Search
- Amazon’s: People also bought...
- Google Personalized Search
- Amazon’s: People also bought...
- Google Personalized Search

2017
- Jeff Dean, Keynote at UCSB, March 2017: talked about visualizations for deep learning.
- LeARn: Augmented Reality RecSys (2017…)
- Xavier Amatrain Keynote on Exp. and Control (ACM IUI, March 2016)
- Xavier Amatrain Keynote on Exp. and Control (ACM IUI, March 2016)
- Xavier Amatrain Keynote on Exp. and Control (ACM IUI, March 2016)
MoodPlay [UMAP’16]
MoodPlay

Music recommendation system that enables exploration and discovery of new artists through an interactive mood space.
Thoughts and Challenges…
The User Is Important  [Knijnenburg ’12]

Inspectability and Control in TasteWeights
How is trust influenced by explanation and control?

Risk, Reward etc.
(Willingness to engage)

Complexity.
(functionality Etc.)
Measuring Challenge

Real World Problems
- Crisis Detection
- Tweet Credibility Modeling

Automated (Data-Driven) Studies
- Music Exploration
- Trust Perceptions

Abstract Problems

User Studies
- Diner’s Dilemma
- Trust Dynamics Game

Career Recommendation

…in the “Intelligent Systems” session!