# CS 399 Paper

**Mar 21, 2014**

**Bill Chickering (bchick)**
**Jamie Irvine (jirvine)**

## Abstract

*Will write this when the paper is done*

## Introduction

Determining the similarity of two items is super important because... With large amounts of rating data, collaborative filtering is a good method for calculating similarity. However, with sparse rating data, this technique gives noisy and unreliable approximations for similarity. Even worse, low confidence measurements of similarity are indistinguishable from high confidence ones. In this paper, we develop a novel method which incorporates confidence into the similarity score to produce more accurate estimates of similarity.

Item-based collaborative filtering is a common technique for measuring the similarity of two items. There are different forms of collaborative filtering. For this paper, we use a traditional approach. Each item is represented as a vector of ratings. The similarity score of two items is computed by measuring the similarity of the two rating vectors.

There are a few ways to compute the similarity of two vectors. One approach is to use Cosine-similarity, defined as the cosine of the angle between the two vectors:

$$PearsSim(A, B) = ... \tag{1}$$

where this = that ... Another popular approach is to use Pearson correlation:

$$CosSim(A, B) = ... \tag{2}$$

where this = that. Other similarity functions exist, such as one-sided similarities, but Pearson correlation and Cosine-similarity are the most popular and will be the two measurements used in this paper. [1]

Both similarity measurements primarily leverage information from common users, that is users who have rated both items. Because of this, they perform well with a large number common users, but give unreliable results when the items have few common users. In an extreme case, if only one user has rated item $A$ and item $B$, $PearsSim(A, B) = 1$ or $-1$. Not only is this unlikely to be an accurate assessment of the true similarity of $A$ and $B$, it also gives the most extreme results possible, without giving any indication that this is a low-confidence calculation.

In this paper, we construct a more accurate similarity score by incorporating the number of common users, $n$, into the similarity function.

## Problem

The goal is to calculate a modified similiary function that leverages the number of commonn users to better estimate the true similarity. Formally, we design $ModSim$ such that for two items $A$ and $B$ with $n$ users in common

$$ModSim(A, B, n) \approx TrueSim(A, B) \tag{3}$$

[1] Note that the difference between the two similarities is how they handle unpaired ratings; that is ratings from a user who has not rated the other item. Cosine-similarity considers the unknown rating from an unpaired rating to be 0 and then calculates the cosine of the two vectors. Pearson correlation simply throws away all unpaired ratings and calculates the cosine of the two modified vectors.

where $TrueSim$ is the true similarity of the two items. Of course, there is no way to actually know the true similarity of two items, but for a good enough similarity function, such as $PearsSim$ or $CosSim$ and a large enough $n$, we can get a good estimate. Thus, we model $TrueSim$ as

$$TrueSim(A, B) = lim_{n\to\infty} Sim(A, B) \tag{4}$$

where $Sim$ is the most appropriate generic similarity function for the dataset.

Since $TrueSim$ dependends on the choice of $Sim$, we abstract away the details of $Sim$ and use its output directly. In all, the goal is to construct $ModSim$ such that

$$ModSim(Sim(A, B), n) \approx lim_{n\to\infty} Sim(A, B) \tag{5}$$

# Model

We model the problem probabilistically. Let $Y$ be a random variable representing the $TrueSim$ of a randomly chosen pair of items. Let $X_n$ be a random variable representing the $Sim$ of a pair of items with $n$ common users.

We model $Y$ as a Normal distribution

$$Y \sim N(\mu, \sigma_1^2) \tag{6}$$

where $\mu$ and $\sigma_1^2$ are the average similarity score and the variance of similarity scores of all pairs of items, respectively. Since $X_n$ represents a noisy reading of the true similarity $Y$, we model $X_n|Y$ as a Normal distribution around $Y$:

$$(X_n|Y = y) \sim N(y, \sigma_{2,n}^2) \tag{7}$$

Note that $\sigma_{2,n}^2$ represents how noisy the estimation of $Sim$ is when there are $n$ common users. This should decrease as $n$ increases.

- illustration

In probabilistic terms, for a calculated value $X_n$, we want to find the value of $Y$ that most likely produced $X_n$. Thus we desire the Maximum Likelihood Estimate of $Y|X_n$

$$\begin{aligned}
\underset{y}{\operatorname{argmax}}\, P(Y = y|X_n = x) &= \underset{y}{\operatorname{argmax}} \left[\frac{P(X_n = x|Y = y)P(Y = y)}{P(X_n = x)}\right] && (8)\\
&= \underset{y}{\operatorname{argmax}} \left[P(X_n = x|Y = y)P(Y = y)\right] && (9)\\
&= \underset{y}{\operatorname{argmax}} \left[\frac{1}{\sigma_{2,n}\sqrt{2\pi}} \exp\left(\frac{-(x-y)^2}{2\sigma_{2,n}^2}\right) \frac{1}{\sigma_1\sqrt{2\pi}} \exp\left(\frac{-(y-\mu)^2}{2\sigma_1^2}\right)\right] && (10)\\
&= \underset{y}{\operatorname{argmin}} \left[\frac{(x-y)^2}{2\sigma_{2,n}^2} + \frac{(y-\mu)^2}{2\sigma_1^2}\right] && (11)\\
&= \underset{y}{\operatorname{argmin}} \left[\left(\sigma_1^2 + \sigma_{2,n}^2\right) y^2 - 2\left(\sigma_1^2 x + \sigma_{2,n}^2 \mu\right) y\right] && (12)
\end{aligned}$$

We find the exact minimum by taking the derivative with respect to $y$ and setting it to zero:

$$\begin{aligned}
\frac{d}{dy}\left[\left(\sigma_1^2 + \sigma_{2,n}^2\right) y^2 - 2\left(\sigma_1^2 x + \sigma_{2,n}^2 \mu\right) y\right] &= 2\left(\sigma_1^2 + \sigma_{2,n}^2\right) y - 2\left(\sigma_1^2 x + \sigma_{2,n}^2 \mu\right) && (13)\\
&= 0 && (14)
\end{aligned}$$

Solving for y:

$$y = \frac{\sigma_1^2 x + \sigma_{2,n}^2 \mu}{\sigma_1^2 + \sigma_{2,n}^2} \tag{15}$$

which can be rewritten as:

$$(y - \mu) = \frac{\sigma_1^2}{\sigma_1^2 + \sigma_{2,n}^2} (x - \mu) \tag{16}$$

- MLE with bias
- Features - linear, mu is fixed point
- Slope - between 0 and 1. As sigma2 changes.
- Model sigma2 as a function of users in common
- Show final parameterized model - discuss params, speculation in model (?)

## Technique

- Discussion of abandonment of parameterization in exchange for linear regression
- Pick data that's big enough to have gold standard
- Slice data and group by n
- Run linear regression on each group of n
- For predicting true similarity, take a given point and project on appropriate
- regression.

## Experiment

## Results

## Conclusion