Accuracy of Similarity Measures in Recommender Systems

Seth Sorensen

December 1, 2012

Recommender Systems

Purpose:

- Make personalized recommendations
- Save user time and effort
- Increase consumption and customer loyalty

Process:

- Collect information regarding user preferences
- Predict user ratings based on collected information
- Make recommendations based on predicted ratings

Recommender Systems

amazon.com®

NETFLIX

Recommender Systems

Challenges:

- Calculate accurate predicted ratings
- Adequately handle situations in which data is sparse

Recommender Systems

Collaborative filtering

Users who have exhibited similar preferences in the past are likely to provide similar ratings for co-rated items.

Content-based

Two items that exhibit high levels of similarity are likely to receive similar ratings from a user.

Recommender Systems

Prediction functions

- Calculate predicted rating for user u of some item d
- Denoted by $R(u, d)$

Weighted arithmetic mean (WAM):

$$WAM(\mathbf{X}) = \sum_{j=1}^{k} w_j x_j$$

Assuming:

$$\sum_{j=1}^{k} w_j = 1$$

Recommender Systems

Collaborative filtering:

$$R(u,d) = \sum_{j=1}^{k} sim(u,u_j)R(u_j,d)$$

Content-based:

$$R(u,d) = \sum_{j=1}^{k} sim(d,d_j)R(u,d_j)$$

Recommender Systems

Collaborative filtering:

$$R(u,d) = \sum_{j=1}^{k} sim(u,u_j) R(u_j,d)$$

Content-based:

$$R(u,d) = \sum_{j=1}^{k} sim(d,d_j) R(u,d_j)$$

Cosine Similarity

Given two users, u and u_i, calculate $sim(u, u_i)$.

$D = \{d_1, d_2, ..., d_n\}$ is the set of all items that have been rated by both u and u_i.

Cosine Similarity

Given two users, u and u_i, calculate $sim(u, u_i)$.

$D = \{d_1, d_2, ..., d_n\}$ is the set of all items that have been rated by both u and u_i.

$$\mathbf{u} = \begin{bmatrix} R(u, d_1) \\ R(u, d_2) \\ . \\ . \\ R(u, d_n) \end{bmatrix} \qquad \mathbf{u_i} = \begin{bmatrix} R(u_i, d_1) \\ R(u_i, d_2) \\ . \\ . \\ R(u_i, d_n) \end{bmatrix}$$

Cosine Similarity

$$cos(\theta) = \frac{\mathbf{u} \cdot \mathbf{u_i}}{\|\mathbf{u}\| \, \|\mathbf{u_i}\|}$$

Cosine Similarity

$$cos(\theta) = \frac{\mathbf{u} \cdot \mathbf{u_i}}{\|\mathbf{u}\| \, \|\mathbf{u_i}\|} = sim(u, u_i)$$

Cosine Similarity

$$cos(\theta) = \frac{\mathbf{u} \cdot \mathbf{u_i}}{\|\mathbf{u}\| \|\mathbf{u_i}\|} = sim(u, u_i)$$

$$R(u, d) = \sum_{i=1}^{k} sim(u, u_i) R(u_i, d)$$

Cosine Similarity

$$cos(\theta) = \frac{\mathbf{u} \cdot \mathbf{u_i}}{\|\mathbf{u}\| \|\mathbf{u_i}\|} = sim(u, u_i)$$

$$R(u, d) = \sum_{i=1}^{k} \frac{\mathbf{u} \cdot \mathbf{u_i}}{\|\mathbf{u}\| \|\mathbf{u_i}\|} R(u_i, d)$$

Cosine Similarity

$$cos(\theta) = \frac{\mathbf{u} \cdot \mathbf{u_i}}{\|\mathbf{u}\| \, \|\mathbf{u_i}\|} = sim(u, u_i)$$

$$R(u, d) = \frac{\sum_{i=1}^{k} \frac{\mathbf{u} \cdot \mathbf{u_i}}{\|\mathbf{u}\| \|\mathbf{u_i}\|} R(u_i, d)}{\sum_{i=1}^{k} \frac{\mathbf{u} \cdot \mathbf{u_i}}{\|\mathbf{u}\| \|\mathbf{u_i}\|}}$$

Similarity Measures

Given two users, u and u_i, calculate $sim(u, u_i)$

$D = \{d_1, d_2, ..., d_n\}$ is the set of all items that have been rated by both u and u_i

Correlation Coefficient

$$corr(\mathbf{X}, \mathbf{Y}) = \frac{Cov(\mathbf{X}, \mathbf{Y})}{\sigma_{\mathbf{X}} \sigma_{\mathbf{Y}}}$$

Correlation Coefficient

$$corr(\mathbf{X}, \mathbf{Y}) = \frac{Cov(\mathbf{X}, \mathbf{Y})}{\sigma_{\mathbf{X}} \sigma_{\mathbf{Y}}}$$

$$\mathbf{X} = u$$

$$\mathbf{Y} = u_i$$

Correlation Coefficient

$$corr(\mathbf{X}, \mathbf{Y}) = \frac{Cov(\mathbf{X}, \mathbf{Y})}{\sigma_{\mathbf{X}} \sigma_{\mathbf{Y}}}$$

$$\mathbf{X} = u$$

$$\mathbf{Y} = u_i$$

$$corr(u, u_i) = \frac{Cov(u, u_i)}{\sigma_u \sigma_{u_i}}$$

Correlation Coefficient

$$corr(\mathbf{X}, \mathbf{Y}) = \frac{Cov(\mathbf{X}, \mathbf{Y})}{\sigma_{\mathbf{X}} \sigma_{\mathbf{Y}}}$$

$$\mathbf{X} = u$$

$$\mathbf{Y} = u_i$$

$$corr(u, u_i) = \frac{Cov(u, u_i)}{\sigma_u \sigma_{u_i}}$$

Covariance

$$Cov(\mathbf{X}, \mathbf{Y}) = E[(\mathbf{X} - E[\mathbf{X}])(\mathbf{Y} - E[\mathbf{Y}])]$$

Covariance

$$Cov(\mathbf{X}, \mathbf{Y}) = E[(\mathbf{X} - E[\mathbf{X}])(\mathbf{Y} - E[\mathbf{Y}])]$$

$$Cov(u, u_i) = \frac{1}{n} \sum_{j=1}^{n} (R(u, d_j) - \overline{R(u)})(R(u_i, d_j) - \overline{R(u_i)})$$

Correlation Coefficient

$$corr(\mathbf{X}, \mathbf{Y}) = \frac{Cov(\mathbf{X}, \mathbf{Y})}{\sigma_{\mathbf{X}} \sigma_{\mathbf{Y}}}$$

$$\mathbf{X} = u$$

$$\mathbf{Y} = u_i$$

$$corr(u, u_i) = \frac{Cov(u, u_i)}{\sigma_u \sigma_{u_i}}$$

Correlation Coefficient

$$corr(\mathbf{X}, \mathbf{Y}) = \frac{Cov(\mathbf{X}, \mathbf{Y})}{\sigma_{\mathbf{X}} \sigma_{\mathbf{Y}}}$$

$$\mathbf{X} = u$$

$$\mathbf{Y} = u_i$$

$$corr(u, u_i) = \frac{\frac{1}{n} \sum_{j=1}^{n} (R(u, d_j) - \overline{R(u)})(R(u_i, d_j) - \overline{R(u_i)})}{\sigma_u \sigma_{u_i}}$$

Correlation Coefficient

$$corr(\mathbf{X}, \mathbf{Y}) = \frac{Cov(\mathbf{X}, \mathbf{Y})}{\sigma_{\mathbf{X}} \sigma_{\mathbf{Y}}}$$

$$\mathbf{X} = u$$

$$\mathbf{Y} = u_i$$

$$corr(u, u_i) = \frac{\frac{1}{n} \sum_{j=1}^{n} (R(u, d_j) - \overline{R(u)})(R(u_i, d_j) - \overline{R(u_i)})}{\sigma_{\boldsymbol{u}} \sigma_{u_i}}$$

Standard Deviation

$$\sigma_{\mathbf{X}} = \sqrt{E[(\mathbf{X} - E[\mathbf{X}])^2]}$$

Standard Deviation

$$\sigma_{\mathbf{X}} = \sqrt{E[(\mathbf{X} - E[\mathbf{X}])^2]}$$

$$\sigma_u = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(R(u, d_j) - \overline{R(u)})^2}$$

Correlation Coefficient

$$corr(\mathbf{X}, \mathbf{Y}) = \frac{Cov(\mathbf{X}, \mathbf{Y})}{\sigma_{\mathbf{X}} \sigma_{\mathbf{Y}}}$$

$$corr(u, u_i) = \frac{\frac{1}{n} \sum_{j=1}^{n} (R(u, d_j) - \overline{R(u)})(R(u_i, d_j) - \overline{R(u_i)})}{\sigma_u \sigma_{u_i}}$$

Correlation Coefficient

$$corr(\mathbf{X}, \mathbf{Y}) = \frac{Cov(\mathbf{X}, \mathbf{Y})}{\sigma_{\mathbf{X}} \sigma_{\mathbf{Y}}}$$

$$corr(u, u_i) = \frac{\frac{1}{n} \sum_{j=1}^{n} (R(u, d_j) - \overline{R(u)})(R(u_i, d_j) - \overline{R(u_i)})}{\sqrt{\frac{1}{n} \sum_{j=1}^{n} (R(u, d_j) - \overline{R(u)})^2} \sigma_{u_i}}$$

Correlation Coefficient

$$corr(\mathbf{X}, \mathbf{Y}) = \frac{Cov(\mathbf{X}, \mathbf{Y})}{\sigma_{\mathbf{X}} \sigma_{\mathbf{Y}}}$$

$$corr(u, u_i) = \frac{\frac{1}{n} \sum_{j=1}^{n} (R(u, d_j) - \overline{R(u)})(R(u_i, d_j) - \overline{R(u_i)})}{\sqrt{\frac{1}{n} \sum_{j=1}^{n} (R(u, d_j) - \overline{R(u)})^2} \sqrt{\frac{1}{n} \sum_{j=1}^{n} (R(u_i, d_j) - \overline{R(u_i)})^2}}$$

Correlation Coefficient

$$corr(\mathbf{X}, \mathbf{Y}) = \frac{Cov(\mathbf{X}, \mathbf{Y})}{\sigma_{\mathbf{X}} \sigma_{\mathbf{Y}}}$$

$$corr(u, u_i) = \frac{\sum_{j=1}^{n} (R(u, d_j) - \overline{R(u)})(R(u_i, d_j) - \overline{R(u_i)})}{\sqrt{\sum_{j=1}^{n} (R(u, d_j) - \overline{R(u)})^2} \sqrt{\sum_{j=1}^{n} (R(u_i, d_j) - \overline{R(u_i)})^2}}$$

Correlation Coefficient

$$corr(\mathbf{X}, \mathbf{Y}) = \frac{Cov(\mathbf{X}, \mathbf{Y})}{\sigma_{\mathbf{X}} \sigma_{\mathbf{Y}}}$$

$$corr(u, u_i) = \frac{\sum_{j=1}^{n} (R(u, d_j) - \overline{R(u)})(R(u_i, d_j) - \overline{R(u_i)})}{\sqrt{\sum_{j=1}^{n} (R(u, d_j) - \overline{R(u)})^2} \sqrt{\sum_{j=1}^{n} (R(u_i, d_j) - \overline{R(u_i)})^2}}$$

$$R(u, d) = \sum_{i=1}^{k} sim(u, u_i) R(u_i, d)$$

Correlation Coefficient

$$corr(\mathbf{X}, \mathbf{Y}) = \frac{Cov(\mathbf{X}, \mathbf{Y})}{\sigma_{\mathbf{X}} \sigma_{\mathbf{Y}}}$$

$$corr(u, u_i) = \frac{\sum_{j=1}^{n} (R(u, d_j) - \overline{R(u)})(R(u_i, d_j) - \overline{R(u_i)})}{\sqrt{\sum_{j=1}^{n} (R(u, d_j) - \overline{R(u)})^2} \sqrt{\sum_{j=1}^{n} (R(u_i, d_j) - \overline{R(u_i)})^2}}$$

$$R(u, d) = \sum_{i=1}^{k} corr(u, u_i) R(u_i, d)$$

Correlation Coefficient

$$corr(\mathbf{X}, \mathbf{Y}) = \frac{Cov(\mathbf{X}, \mathbf{Y})}{\sigma_{\mathbf{X}} \sigma_{\mathbf{Y}}}$$

$$corr(u, u_i) = \frac{\sum_{j=1}^{n} (R(u, d_j) - \overline{R(u)})(R(u_i, d_j) - \overline{R(u_i)})}{\sqrt{\sum_{j=1}^{n} (R(u, d_j) - \overline{R(u)})^2} \sqrt{\sum_{j=1}^{n} (R(u_i, d_j) - \overline{R(u_i)})^2}}$$

$$R(u, d) = \frac{\sum_{i=1}^{k} corr(u, u_i) R(u_i, d)}{\sum_{i=1}^{k} corr(u, u_i)}$$

Orkut

- Orkut
- Orkut Büyükökten
- Communities - groups of users with a shared interest
- Four months, 50,000 communities

Preparation

Initial data set

- All pairs of the form (u, c), where c is a community to which at least 20 members belong, and u is a user that belongs to at least one such community
- 19,792 communities, 181,160 users

Initial calculations

- Similarities were calculated for each pair of communities, using 6 different measures, including the cosine similarity measure
- Calculations were based on the number of overlapping members
- Similarities were calculated once at the beginning of the experiment

Recommendation

- July 1 2004 - July 18 2004
- New users
- Base community
- Select two similarity measures "in a deterministic manner so that a given user always saw the same recommendations for a given community"
- Interleave the top six results returned by each measure
- Display recommendations on base community page in rows of three with name, link, and picture (optional)

Observation

M $\rightarrow$ M - Member of base community, already member of recommended community

n $\rightarrow$ M - Not member of base community, already member of recommended community

M $\rightarrow$ n - Member of base community, not member of recommended community, doesn't join

n $\rightarrow$ n - Not member of base community, not member of recommended community, doesn't join

M $\rightarrow$ j - Member of base community, not member of recommended community, does join

n $\rightarrow$ j - Not member of base community, not member of recommended community, does join

Observation

M $\rightarrow$ M - Member of base community, already member of recommended community

n $\rightarrow$ M - Not member of base community, already member of recommended community

M $\rightarrow$ n - Member of base community, not member of recommended community, doesn't join

n $\rightarrow$ n - Not me **Conversions** community, not member of recommended community, doesn't join

M $\rightarrow$ j - Member of base community, not member of recommended community, does join

n $\rightarrow$ j - Not member of base community, not member of recommended community, does join

Results

Measures		M → j			n → j		
		Wins	Losses	Ties	Wins	Losses	Ties
Cosine	Sim2	6899	4993	2977	2600	1853	1073
Cosine	Sim3	6940	5008	2743	2636	1872	1078
Cosine	Sim4	6929	5064	2697	2610	1865	1064
Cosine	Sim5	7039	4834	2539	2547	1983	941
Cosine	Sim6	8186	4442	1638	2852	1655	564

Table : Wins, losses, and ties for the L2 similarity measure when compared to 5 other similarity measures [16]

Experiment

- Lathia *et al.*, 2008
- MovieLens data set - movie ratings on a scale of 1 to 5
- Compare the accuracy of predicted ratings calculated using 7 different measures of similarity
- Each measure returns a value in the range [-1.0, 1.0]

Similarity Measures

Co-rated:

- Based on the quantity of co-rated items
- R_u is the set of items rated by user u
- R_{u_i} is the set of items rated by user u_i

$$sim(u, u_i) = \frac{|R_u \cap R_{u_i}|}{|R_u \cup R_{u_i}|}$$

Similarity Measures

Concordance:

- Concordant - Both users rate the item higher than their respective average ratings or both users rate the item lower than their respective average ratings
- Discordant - One user rates the item higher than their average rating and the other rates it lower than their average rating
- Tied - The rating of one or both users is the same as their average rating

$$sim(u, u_i) = \frac{|C| - |D|}{|N| - |T|}$$

Similarity Measures

Remaining 5 measures:

- Pearson Correlation Coefficient (PCC)
- Weighted PCC
- R(0.5, 1.0)
- R(-1.0, 1.0)
- Constant(1.0)

Calculations

- Divide data set into $s1$, $s2$, $s3$, $s4$, and $s5$
- Calculate a predicted rating for every actual rating provided in the initial data set by substituting each of the 7 similarity measures into the equation:

$$R(u,d) = \overline{R(u)} + \frac{\sum_{i=1}^{k} sim(u,u_i)[R(u_i,d) - \overline{R(u_i)}]}{\sum_{i=1}^{k} sim(u,u_i)}$$

  and varying k.
- Calculate the mean average error for each measure of similarity

Results

Dataset	Co-Rated	Concordance	PCC	R(0.5, 1.0)	R(-1.0, 1.0)	Constant(1.0)
s1	0.7718	0.7992	0.8073	0.7773	0.7812	0.7769
s2	0.7559	0.7825	0.7953	0.7630	0.7666	0.7628
s3	0.7490	0.7706	0.7801	0.7554	0.7563	0.7551
s4	0.7463	0.7666	0.7792	0.7534	0.7554	0.7531
s5	0.7501	0.7715	0.7824	0.7573	0.7595	0.7573
Average	**0.7548**	**0.7781**	**0.7889**	**0.7613**	**0.7638**	**0.7610**

Table : The average error of the predicted rating, ordered by dataset [11]

Results

k	Co-Rated	Concordance	PCC	R(0.5, 1.0)	R(-1.0, 1.0)	Constant(1.0)
1	0.9449	0.9492	1.1150	1.0665	1.0341	1.0406
10	0.8498	0.8355	1.0455	0.9595	0.9689	0.9495
30	0.7979	0.7931	0.9464	0.8903	0.8848	0.9108
50	0.7852	0.7817	0.9007	0.8584	0.8498	0.8922
100	0.7759	0.7728	0.8136	0.8222	0.8153	0.8511
153	0.7725	0.7727	0.7817	0.8053	0.8024	0.8243
229	0.7717	0.7771	0.7716	0.7919	0.8058	0.7992
459	0.7718	0.7992	0.8073	0.7773	0.7812	0.7769

Table : The average error of the predicted rating, ordered by neighborhood size, for $s1$ [11]

Remarks

- Orkut *et al.* find that certain similarity measures perform better than others.
- Lathia *et al.* find that choice of similarity measure does not affect the accuracy of predicted ratings and that predicted ratings are most accurate when k approaches the size of the data set.
- The experiments considered two different sets of similarity measures, used different types of test data, and based conclusions on different types of observations.

Acknowledgements

- Kristin Lamberty
- Nic McPhee
- Elijah Mayfield

Questions?

Questions?

Comments?

References I

G. Adomavicius and Y. Kwon.

New recommendation techniques for multicriteria rating systems.
IEEE Intelligent Systems, 22(3):48–55, May 2007.

X. Amatriain, A. Jaimes*, N. Oliver, and J. M. Pujol.

Data mining methods for recommender systems.
In F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, editors, *Recommender Systems Handbook*, pages 39–71. Springer US, 2011.

Amazon.

Company facts, 2012.
[Online; accessed 10-October-2012].

R. Andersen, C. Borgs, J. Chayes, U. Feige, A. Flaxman, A. Kalai, V. Mirrokni, and M. Tennenholtz.

Trust-based recommendation systems: an axiomatic approach.
In *Proceedings of the 17th international conference on World Wide Web*, WWW '08, pages 199–208, New York, NY, USA, 2008. ACM.

G. Beliakov, T. Calvo, and S. James.

Aggregation of preferences in recommender systems.
In F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, editors, *Recommender Systems Handbook*, pages 705–734. Springer US, 2011.

References II


Google.
2012 financial tables, 2012.
[Online; accessed 10-October-2012].


Google.
Powerful targeting technology to reach the right audience, 2012.
[Online; accessed 10-October-2012].


J. L. Herlocker, J. A. Konstan, A. Borchers, and J. Riedl.
An algorithmic framework for performing collaborative filtering.
In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '99, pages 230–237, New York, NY, USA, 1999. ACM.


J. A. Jacobi, E. A. Benson, and G. D. Linden.
U.S. Patent 7,908,183 B2, 03 2011.


Y. Koren, R. Bell, and C. Volinsky.
Matrix factorization techniques for recommender systems.
Computer, 42(8):30 –37, aug. 2009.


N. Lathia, S. Hailes, and L. Capra.
The effect of correlation coefficients on communities of recommenders.
In *Proceedings of the 2008 ACM symposium on Applied computing*, SAC '08, pages 2000–2005, New York, NY, USA, 2008. ACM.

References III

Pandora.

Pandora announces september 2012 audience metrics, 2012.
[Online; accessed 10-October-2012].

F. Ricci, L. Rokach, and B. Shapira.

Introduction to recommender systems handbook.
In F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, editors, *Recommender Systems Handbook*, pages 1–35. Springer US, 2011.

N. Rubens, D. Kaplan, and M. Sugiyama.

Active learning in recommender systems.
In F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, editors, *Recommender Systems Handbook*, pages 735–767. Springer US, 2011.

J. Schafer, D. Frankowski, J. Herlocker, and S. Sen.

Collaborative filtering recommender systems.
In P. Brusilovsky, A. Kobsa, and W. Nejdl, editors, *The Adaptive Web*, volume 4321 of *Lecture Notes in Computer Science*, pages 291–324. Springer Berlin / Heidelberg, 2007.

E. Spertus, M. Sahami, and O. Buyukkokten.

Evaluating similarity measures: a large-scale study in the Orkut social network.
In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, KDD '05, pages 678–684, New York, NY, USA, 2005. ACM.

References IV

E. W. Weisstein.

Least squares fitting.
From MathWorld–A Wolfram Web Resource. http://mathworld.wolfram.com/LeastSquaresFitting.html.
[Online; accessed 3-October-2012].

Wikipedia.

Correlation and dependence — wikipedia, the free encyclopedia, 2012.
[Online; accessed 14-October-2012].

Wikipedia.

Least squares — wikipedia, the free encyclopedia, 2012.
[Online; accessed 3-October-2012].

Wikipedia.

Weighted mean — wikipedia, the free encyclopedia, 2012.
[Online; accessed 4-October-2012].