



Comparison of Various Similarity Measure Techniques for Generating Recommendations for E-commerce Sites and Social Websites

Jyoti¹, Dr. Sanjeev Dhawan², Dr. Kulvinder Singh³

¹M.Tech. (Computer Engineering), ^{2,3}Faculty of Computer Science & Engineering

^{1,2,3}Department of Computer Science & Engineering,
University Institute of Engineering & Technology (U.I.E.T),
Kurukshetra University, Kurukshetra, Haryana, India

Abstract: Recommendation systems play a great role in increasing the efficiency of the website. It not only makes the site user friendly for people, but also, it aids users in making their search process easy, fast and effective. Recommenders also contribute to a good extent the increase in number of users and hence make good profit to website owners. Recommenders provide users with ease to find their item of interest. Item may be a video, song, book, friend list, location etc. the base of a recommender is a similarity calculating algorithm, which calculates the similarity between various items and various users. Better the similarity calculation better will be the results of a recommender according to the taste of users. This paper surveys the existing similarity measure algorithms and implements them on a movie data set to calculate similarity between all pairs of movies. A comparison is made between Pearson correlation, cosine based similarity and Euclidean distance based similarity and their effectiveness in making recommendations to users. In nutshell, an attempt has been made to provide an overview of similarity algorithms and their effectiveness in recommendations systems.

Keywords: Cosine similarity; Euclidean similarity; Pearson coefficient; Recommendation systems.

I. Introduction

The use of internet in all areas of human life, involving all its daily activities has made researchers and website developers face more challenges to make the websites user friendly and efficient. The recommendation systems are an important part of websites which involve customers' interactions [1]. For online shopping and social networking sites, similarity between items and users is necessary. For online shopping sites, user is to be suggested with items which are similar to the item of his taste. And in social websites, similarity between users is to be calculated to provide them suggestion of friends and to check for malicious activities. Similarity can be calculated using various approaches. Three techniques have been implemented and compared in this paper, which are Pearson correlation, cosine based similarity, and Euclidean distance based similarity. The implementation is provided on movie data set. After calculating the similarity, the recommendations are generated for users for a given movie. Similarity between all pair of movies is stored as hash map, using Eclipse IDE for java code. The comparison is made according to the distance between their overall average value for all movies and average of their expected average values for all movie pairs.

The remaining paper is structured as follows: section II introduces the basic Pearson, cosine and Euclidean techniques, section III describes implementation strategy and data sets, section IV presents the comparison between them, section V concludes the paper and section VI gives future scope.

II. Introduction to Similarity Measures

There are several techniques to compute item-to-item similarity. Three of them, Pearson, cosine and Euclidean measures have been discussed.

- A. The Pearson correlation coefficient is a measure of the linear correlation (dependence) between two variables X and Y, giving a value between +1 and -1 inclusive, where 1 is total positive correlation, 0 is no correlation, and -1 is total negative correlation. Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations [2].
- B. Cosine similarity is a measure of similarity between two vectors of an inner product space that measures the cosine of the angle between them [3]. It is thus a judgment of orientation and not magnitude: two vectors with

the same orientation have a cosine similarity of 1, two vectors at 90° have a similarity of 0, and two vectors diametrically opposed have a similarity of -1, independent of their magnitude.

- C. The Euclidean distance or Euclidean metric is the "ordinary" (i.e. straight line) distance between two points in Euclidean space [4]. The distance between two points on the real line is the absolute value of their numerical difference.

II. Implementation Strategy and Data Set

Having introduced the basic definitions of the similarity measures, it can now be deployed on the custom-made search systems and recommendation engines.

- A. Data set: The dataset used in this research is movie data which has three attributes namely, user-id, movie-id, and user-ratings. These ratings are then defined into five class i.e. bad, ok, average, good, and excellent. There are 100000 entries for user ratings. There are 943 users who provide different ratings to 1682 different movies. A subset of this data set is used for computing similarity between movie pairs. The data set has been obtained from MovieLens web site (<http://movielens.org>) which is collected and made available by GroupLens research [5].

B. Implementation Approach:

- First, create a hash-map of movie rating table for all pairs of movies.
- Find users who have rated both of the two movies, for each pair.
- Find average ratings for the two movies, for each pair.
- Compute Pearson based similarity.
 - $$\text{similarity} = \frac{\sum_i [(rating1[i] - average1) \times (rating2[i] - average2)]}{\sqrt{\sum_i (rating1[i] - average1)^2} \times \sqrt{\sum_i (rating2[i] - average2)^2}}$$
- Compute Euclidean distance based similarity.
 - $$\text{distance} = \sqrt{\sum_i (rating1[i] - rating2[i])^2}$$
 - $$\text{similarity} = \frac{1}{(1 + \text{distance})}$$
- Compute cosine based similarity.
 - $$\text{similarity} = \frac{\sum_i (rating1[i] \times rating2[i])}{\sqrt{\sum_i (rating1[i])^2} \times \sqrt{\sum_i (rating2[i])^2}}$$

III. Results

Comparison based on computed values is given in table 1:

Table 1: comparison of expected and obtained

Total average value for 10000 movie pairs	5223.309572
Average RMSE value corresponding to Pearson similarity	1.008200076
Average RMSE value corresponding to Cosine similarity	0.3991831217
Average RMSE value corresponding to Euclidean distance based similarity	0.4894081776
Average expected value	0.522330957
Average Pearson similarity value	1.124032848
Average Cosine similarity value	0.292144858
Average Euclidean similarity value	0.150815165

The above table clearly shows that the average value for Pearson coefficient based similarity is nearest to the expected average value, so it becomes the best method to compute similarity for this movie data.

- After computing similarity between all movie pairs, average similarity value for each pair is computed.
- Using this average value and pearson similarity value, for each movie pair, Root mean square error is calculated for each movie pair.
- Similarly Root mean square error is calculated [6] for each pair of movie, corresponding to average value and cosine based similarity value.

- Then again, Root mean square error is calculated for each movie pair, for Euclidean distance based similarity value.

Then, for the count of 100 movies, there are 10000 pairs for comparison. And the sum of RMSE for pearson coefficient based similarity, cosine based similarity and Euclidean distance based similarity is computed. And finally, their average values are compared so as to find the best similarity measure technique for movie data set.

A snapshot of first few rows of data is shown in fig 1:

	A	B	C	D	E	F	G	H	I
1	Movie ID	Movie ID	Pearson Similarity	Cosine Similarity	Euclidean Distance Measure	average	Pearson RMSE	Cosine RMSE	Euclidean RMSE
2	1	1	1.165667944	1	1	1.055222648	0.495264494	0.336889948	0.336889948
3	1	2	1.526745114	0.402382178	0.035596773	0.654908022	1.379146883	0.51671375	0.653939895
4	1	3	1.516086456	0.330244787	0.040579177	0.62897014	1.379461745	0.53529601	0.627659755
5	1	4	1.401078709	0.454937918	0.030549964	0.62885553	1.252023271	0.434155236	0.628113029
6	1	5	1.430685608	0.286713512	0.048918361	0.588772494	1.303920342	0.514245478	0.586736775
7	1	6	1.291866335	0.116343979	0.09174737	0.499985895	1.19118963	0.48626122	0.491495997
8	1	7	1.150045811	0.620978588	0.024563109	0.598529169	0.982022505	0.165460691	0.598024933
9	1	8	1.340854694	0.481113887	0.0308031	0.61759056	1.190156801	0.387230587	0.61682191
10	1	9	1.228897966	0.496288432	0.02634391	0.583843436	1.081349644	0.307523901	0.583248794
11	1	10	1.42261058	0.27393511	0.044548171	0.580364621	1.298844783	0.511646996	0.57865236
12	1	11	1.347879051	0.468290931	0.029664036	0.615278006	1.199254315	0.399087244	0.614562503
13	1	12	1.299561857	0.460391973	0.028369529	0.596107786	1.154779861	0.378660434	0.595432333
14	1	13	1.355894799	0.41750874	0.032628627	0.602010722	1.214921313	0.433708844	0.601125845
15	1	14	1.405478122	0.347678268	0.035620181	0.596258857	1.272730972	0.484401122	0.595193941
16	1	15	1.224030538	0.574376678	0.029813119	0.609406778	1.061543281	0.203637063	0.60867709
17	1	16	1.426358969	0.224119569	0.063569073	0.571349204	1.306927693	0.525557163	0.567801801
18	1	17	1.444680148	0.273915881	0.046630811	0.588408947	1.319422466	0.520764034	0.586558314
19	1	18	1.614325768	0.046232337	0.148267583	0.602941896	1.497500836	0.601166783	0.584427629
20	1	19	1.307667326	0.191771749	0.064987838	0.521475638	1.199190141	0.484933435	0.517410303
21	1	20	1.460516329	0.232930087	0.049229054	0.580891823	1.340027103	0.532145549	0.578802048
22	1	21	1.505052851	0.239195619	0.042644074	0.628964165	1.367320847	0.529662395	0.627516858

Fig. 1: similarity values for first few movie pairs

IV. Conclusion

The measure of item-to-item similarity is very useful for the collaborative filtering in generating recommendations for users. Similarity can be measured using various techniques. The most renowned methods are Pearson coefficient, cosine based similarity and Euclidean distance based similarity. In this paper, an attempt has been made to give a comparative view of the above mentioned techniques. The implementation of these approaches on movie data set to build a collaborative filtering recommender gives a hash map of values. These values are then compared against the average expected values and root mean square error is computed. Pearson coefficient comes out to be the best approach for this subset of data.

V. Future Scope

In future, this similarity measure can be deployed for calculating user-to-user similarity instead of item-to-item-similarity. The computation of similarity with best techniques is not only useful for users to get best matches, but also, for developers to interact more customers by satisfying their needs. This approach can be combined with any searching algorithm, ranking algorithm to get online content more effective. In future, a combination of these approaches can be helpful to get more efficient results for recommendation systems.

References

- [1] A. T. Mulik and S. Z. Gawali, "Recommendation System: Online Movie Store", pp. 207-211, International Journal of Application or Innovation in Engineering & Management (IAJEM), vol. 2, issue 6, 2013.
- [2] <http://www.statisticshowto.com/what-is-the-pearson-correlation-coefficient/>, accessed on 13-May-2015 at 11:00 pm.
- [3] http://en.wikipedia.org/wiki/Cosine_similarity, accessed on 13-May-2015 at 06:00 pm.
- [4] http://en.wikipedia.org/wiki/Euclidean_distance, accessed on 14-May-2015 at 09:00 pm.
- [5] <http://grouplens.org/datasets/movielens/> accessed on 12-january-2015 at 05:00 pm.
- [6] http://en.wikipedia.org/wiki/Root_mean_square accessed on 12-May-2015 at 06:45 pm.