

# MemoryLeak

## First Iteration Report

In this iteration, we had three objectives, which were collecting papers for improving mathematical model, collecting papers to improve the speed of processing raw data and determining the similarity function.

According to one of the papers that we studied, which is “Performance Comparison of Combined Collaborative Filtering Algorithms for Recommender Systems”, Tanimoto Coefficient Similarity is mostly used for non-normalized sparse datasets. After considering this paper and the results of our first evaluation, we decided to use the Tanimoto Coefficient Similarity as our similarity function. Formulas for Tanimoto Coefficient and some other algorithms are given below:

- Pearson Correlation

$$r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}}$$

- Spearman Correlation

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2} \sqrt{\sum_i (y_i - \bar{y})^2}}$$

- Tanimoto Coefficient

$$T(a, b) = \frac{N_c}{N_a + N_b - N_c}$$

Another paper that we studied, “I like it... I like it not: Evaluating User Ratings Noise in Recommender Systems”, taught us that the rating scale, the order of the items rated, and the time of the ratings are given are three important factors in calculating ratings with a mathematical model. Therefore, we decided to extend our current mathematical model by taking these into account on the next iteration, which is improving the math model. In our case, especially the time when the user listened to a song is of great importance in giving recommendations.

The final paper that inspires us is “Item-Based Collaborative Filtering Recommendation Algorithms”. In that paper, it is stated that the Model-Based Approach can be used when computing recommendations to reduce the complexity of the process. If we divide n items into groups of k ( $k \ll n$ ), the complexity will be  $O(k^2)$  which is much more smaller than  $O(n^2)$ . The important thing is to decide on the k value because as k increases, the accuracy of the

recommendations gets better but the efficiency gets worse and vice versa. Therefore, we have to consider this tradeoff of the Model-Based Approach.

Links to the papers mentioned above are given below:

- <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6341589>
- [http://www.nuriaoliver.com/RecSys/Likelt\\_umap09.pdf](http://www.nuriaoliver.com/RecSys/Likelt_umap09.pdf)
- <http://www.ra.ethz.ch/cdstore/www10/papers/pdf/p519.pdf>