A list of similarity metrics for KNN Recommender Systems

Maurizio Ferrari Dacrema
Politecnico di Milano, Italy
maurizio.ferrari@polimi.it

ABSTRACT

This paper provides a list of some similarity heuristics that can be used in a KNN Recommender system, both Item-based and User-based.

ACM Reference Format:
Maurizio Ferrari Dacrema. 2019. A list of similarity metrics for KNN Recommender Systems. In *Recommender Sytems course at Politecnico di Milano (RecSys@Polimi).* 1 page.

1 SIMILARITY HEURISTICS

Vectors $\vec{r}_i, \vec{r}_j \in \mathbb{R}^{|U|}$ represent the ratings of a user for items i and j, respectively, and $|U|$ is the number of users. In case of CBF recommender vectors $\vec{r}_i, \vec{r}_j \in \mathbb{R}^{|F|}$ describe the features of items i and j, respectively, and $|F|$ is the number of features.

Parameter h (the *shrink term*) is used to lower the similarity between items having only few interactions [2].

Cosine

$$s_{ij} = \frac{\vec{r}_i \cdot \vec{r}_j}{\|\vec{r}_i\|\|\vec{r}_j\| + h} \tag{1}$$

Asymmetric Cosine Described in [1]

$$s_{ij} = \frac{\vec{r}_i \cdot \vec{r}_j}{\|\vec{r}_i\|^{\alpha}\|\vec{r}_j\|^{1-\alpha} + h} \tag{2}$$

Dice - Sørensen Is a similarity measure on sets, therefore defined on boolean vectors. Given two sets or verctors the respective Dice similarity can be computed as a linear function of Jaccard similarity and vice versa. The Dice coefficient doesn't satisfy the triangle inequality, it can be considered a semimetric version of the Jaccard index.

$$s_{AB} = 2\frac{A \cap B}{\|A\| + \|B\| + h} \tag{3}$$

$$s_{ij} = 2\frac{\vec{r}_i \cdot \vec{r}_j}{\|\vec{r}_i\| + \|\vec{r}_j\| + h} \tag{4}$$

Jaccard - Tanimoto Is a similarity measure on sets, therefore defined on boolean vectors. The term Tanimoto is sometimes used to refer to its formulation on vectors, however the corresponding similarity is not a proper similarity since its distance does not preserve the triangle inequality.

$$s_{AB} = \frac{A \cap B}{\|A\| + \|B\| - \|A \cap B\| + h} \tag{5}$$

$$s_{ij} = \frac{\vec{r}_i \cdot \vec{r}_j}{\|\vec{r}_i\| + \|\vec{r}_j\| - \vec{r}_i \cdot \vec{r}_j + h} \tag{6}$$

Tversky Is an asymmetric similarity measure on sets which is a generalisation of Dice/Sørensen and Tanimoto/Jaccard coefficients.

$$s_{AB} = \frac{A \cap B}{\|A \cap B\| + \alpha\|A - B\| + \beta\|B - A\| + h} \tag{7}$$

Where $-$ denotes the *relative complement* of two sets, therefore $\|B - A\|$ means the number of elements in B that are not in A and can be represented as $\|B\| - \|A \cap B\|$. If $\alpha = \beta = 1$ we get the Tanimoto coefficient, if $\alpha = \beta = 0.5$ we get the Dice coefficient.

In case of boolean attributes it can be calculated as follows:

$$s_{ij} = \frac{\vec{r}_i \cdot \vec{r}_j}{\vec{r}_i \cdot \vec{r}_j + \alpha(\|\vec{r}_i\| - \vec{r}_i \cdot \vec{r}_j) + \beta(\|\vec{r}_j\| - \vec{r}_i \cdot \vec{r}_j) + h} \tag{8}$$

Euclidean The Euclidean similarity, e, is can be defined via the euclidean distance ed which can be efficiently computed from the dot product of the vectors:

$$\begin{aligned} ed_{ij} &= \sqrt{\sum_{u \in U} \left(\vec{r}_{iu} - \vec{r}_{ju}\right)^2} && (9) \\ &= \sqrt{\sum_{u \in U} \left(\vec{r}_{iu}^2 + \vec{r}_{ju}^2 - 2\vec{r}_{iu} \cdot \vec{r}_{ju}\right)^2} && (10) \\ &= \sqrt{\|\vec{r}_i\| + \|\vec{r}_j\| - 2\vec{r}_i \cdot \vec{r}_j} && (11) \end{aligned}$$

$$e_{ij} = \begin{cases} \frac{1}{e^{ed_{ij}}+h} & \text{exponential} \\ \frac{1}{log(ed_{ij}+1)+h} & \text{logarithmic} \\ \frac{1}{ed_{ij}+h} & \text{linear} \end{cases}$$

Usage notes. Cosine based similarities can handle all types of data and are sensitive to the values of the attributes; set based similarities are not sensitive to the value but only to the presence or absence of a certain feature, for this reason, they should not be applied on dense features. Both cosine and set similarities fail when the data has only one dense feature, in which case euclidean similarity is a good solution.

REFERENCES

[1] Fabio Aiolli. 2013. Efficient top-n recommendation for very large scale binary rated datasets. In *Proceedings of the 7th ACM conference on Recommender systems.* ACM, 273–280.

[2] Robert M Bell and Yehuda Koren. 2007. Improved neighborhood-based collaborative filtering. In *KDD cup and workshop at the KDD '07.* Citeseer, 7–14.

RecSys@Polimi, A.A. 2019-20, Milano, Italy, EU