3-3

SIMILARITY

Similarity between Two Sets

- Jaccard Index

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}.$$

(If A and B are both empty, we define $J(A,B) = 1$.)

https://en.wikipedia.org/wiki/Jaccard_index

Similarity between Two Real-valued Vectors

- Euclidean distance

The **Euclidean distance** between points **p** and **q** is the length of the line segment connecting them ($\overline{\mathbf{pq}}$).

In Cartesian coordinates, if **p** = (p_1, p_2, ..., p_n) and **q** = (q_1, q_2, ..., q_n) are two points in Euclidean n-space, then the distance (d) from **p** to **q**, or from **q** to **p** is given by the Pythagorean formula:

$$\begin{aligned} \mathrm{d}(\mathbf{p},\mathbf{q}) = \mathrm{d}(\mathbf{q},\mathbf{p}) &= \sqrt{(q_1-p_1)^2 + (q_2-p_2)^2 + \cdots + (q_n-p_n)^2} \\ &= \sqrt{\sum_{i=1}^{n}(q_i-p_i)^2}. \end{aligned} \quad (1)$$

https://en.wikipedia.org/wiki/Euclidean_distance

Similarity between Two Real-valued Vectors

- Cosine Similarity

Given two vectors of attributes, A and B, the cosine similarity, $\cos(\theta)$, is represented using a dot product and magnitude as

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}$$

, where A_i and B_i are components of vector A and B respectively.

The resulting similarity ranges from −1 meaning exactly opposite, to 1 meaning exactly the same, with 0 indicating orthogonality (decorrelation), and in-between values indicating intermediate similarity or dissimilarity.

https://en.wikipedia.org/wiki/Cosine_similarity

Similarity between Two Real-valued Vectors

- Cosine Similarity

In the case of information retrieval, the cosine similarity of two documents will range from 0 to 1, since the term frequencies (tf-idf weights) cannot be negative. The angle between two term frequency vectors cannot be greater than 90°.

If the attribute vectors are normalized by subtracting the vector means (e.g., $A - \bar{A}$), the measure is called centered cosine similarity and is equivalent to the Pearson Correlation Coefficient.

Similarity between Two Real-valued Vectors

- Pearson Correlation Coefficient

$$W_{a,u} = \frac{\sum_{i=1}^{m}(r_{a,i} - \bar{r}_a)(r_{u,i} - \bar{r}_u)}{\sqrt{\sum_{i=1}^{m}(r_{a,i} - \bar{r}_a)^2}\sqrt{\sum_{i=1}^{m}(r_{u,i} - \bar{r}_u)^2}}$$

The Cosine similarity of user-mean norm
is Pearson correlation.

https://en.wikipedia.org/wiki/Pearson_product-moment_correlation_coefficient#For_a_sample

Similarity between Two Real-valued Vectors

- Pearson Correlation Coefficient (expressed by z-score)

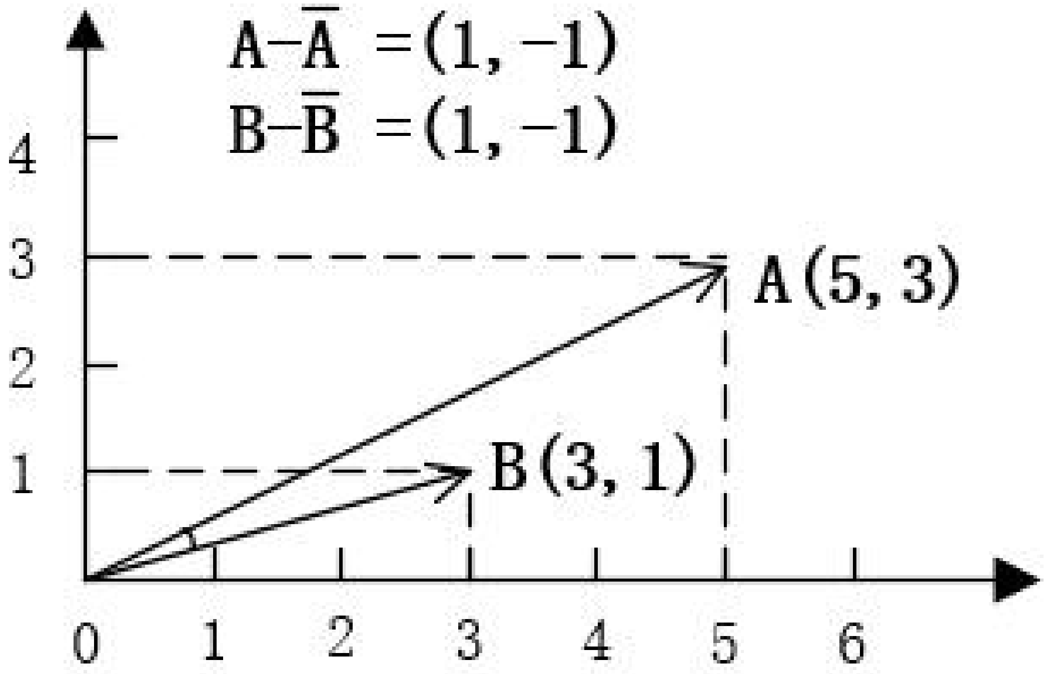
$$W_{a,u} = \frac{\sum_{i=1}^{m} \left(\frac{r_{a,i} - \bar{r}_a}{\sigma_a}\right)\left(\frac{r_{u,i} - \bar{r}_u}{\sigma_u}\right)}{m}$$

$$\sigma_a = \sqrt{\frac{\sum_{i=1}^{m} (r_{a,i} - \bar{r}_a)^2}{m}}$$

https://en.wikipedia.org/wiki/Pearson_product-moment_correlation_coefficient#For_a_sample

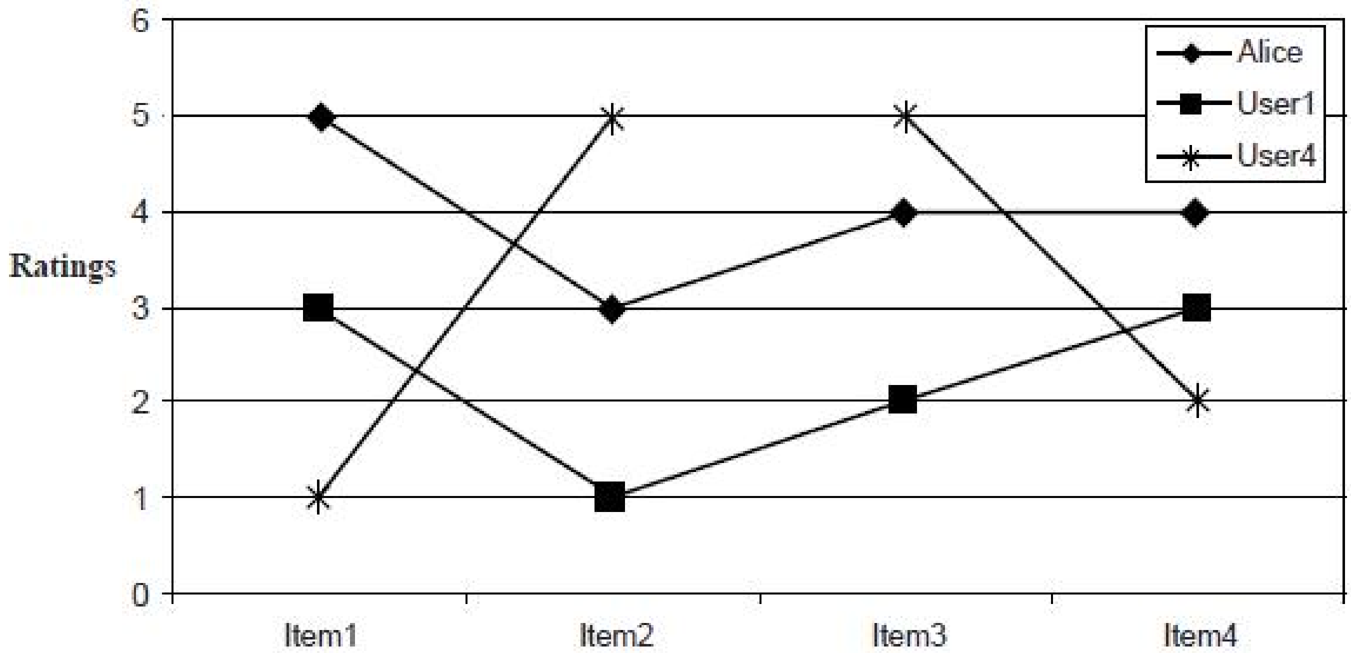
Similarity between Two Real-valued Vectors

- Pearson Correlation Coefficient vs. Cosine Similarity



Similarity between Two Real-valued Vectors

- Pearson Correlation Coefficient vs. Cosine Similarity



Similarity between Two Real-valued Vectors

- **Spearman's rank correlation coefficient**
  - Covert the raw scores to their ranks
  - Compute the Pearson Correlation Coefficient based on the vectors of ranks
- Examples

V1	**5**	**10**	**3**	**8**
V2	**3**	**6**	**9**	**1**

Rank(V1)	**3**	**1**	**4**	**2**
Rank(V2)	**3**	**2**	**1**	**4**

https://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient