TF-IDF

Assumptions of TF-IDF are well exemplified by the function:

$$TF-IDF\left(t_k,\, d_j\right) = \underbrace{TF\left(t_k,\, d_j\right)}_{\text{TF}} \cdot \underbrace{\log \frac{N}{n_k}}_{\text{IDF}}$$

where N denotes the number of documents in the corpus, and n_k denotes the number of documents in the collection in which the term t_k occurs at least once.

$$TF\left(t_k,\, d_j\right) = \frac{f_{k,\,j}}{\max_z f_{z,\,j}}$$

where the maximum is computed over the frequencies $f_{z,\,j}$ of all terms t_z that occur in document d_j. In order for the weights to fall in the [0, 1] interval and for the documents to be represented by vectors of equal length, weights are usually normalized by cosine normalization:

$$w_{k,\,j} = \frac{TF-IDF\left(t_k,\, d_j\right)}{\sqrt{\sum_s^{|T|} TF-IDF\left(t_k,\, d_j\right)^2}}$$

which enforces the normalization assumption.

A similarity measure is required to determine the closeness between two documents. Many similarity measures have been derived to describe the proximity of two vectors; among those measures, cosine similarity is the most widely used:

$$sim\left(d_i,\, d_j\right) = \frac{\sum_k w_{ki} w_{kj}}{\sqrt{\sum_k w_{ki}^2}\sqrt{\sum_k w_{kj}^2}}$$

In content-based recommender systems relying on VSM, both user profiles and items are represented as weighted term vectors. Predictions of a user's interest in a particular item can be derived by computing the cosine similarity