Algorithms and Applications of Data Mining

Yijun Lin
yijunlin@usc.edu

02/27

Recommendation System

- Content-Based Recommendations
- Collaborative Filtering
- Hybrid Systems

Content-Based Recommendations

- How to construct item profile?
- How to measure similarity between items?

Content-Based Recommendations

Given four documents A, B, C, and D and their top two TF-IDF words, A: nba, basketball; B: cancer, health; C: vote, democratic; D: basketball, baseball, write the Boolean feature vectors for each document and calculate the cosine similarity between A, D

Content-Based Recommendations

Given four documents A, B, C, and D and their top two TF-IDF words, A: nba, basketball; B: cancer, health; C: vote, democratic; D: basketball, baseball, write the Boolean feature vectors for each document and calculate the cosine similarity between A, D

Feature Vector (nba, basketball, cancer, health, vote, democratic, baseball)

	nba	basketball	cancer	health	vote	democratic	baseball
A	1	1	0	0	0	0	0
B	0	0	1	1	0	0	0
C	0	0	0	0	1	1	0
D	0	1	0	0	0	0	1

Cosine Similarity(A,D) = $1/(\sqrt{2}*\sqrt{2}) = ½$

Content-Based Recommendations

Given a set of document, briefly explain how to calculate TF and IDF in TF-IDF score. You need to describe any preprocessing you need to apply to the words in a document (e.g., stemming) and how to calculate both the TF and IDF components

Content-Based Recommendations

Given a set of document, briefly explain how to calculate TF and IDF in TF-IDF score. You need to describe any preprocessing you need to apply to the words in a document (e.g., stemming) and how to calculate both the TF and IDF components

Preprocessing:

1. Eliminate stop words
2. Remove rare words
3. Stemming

f_{ij} = frequency of term (feature) *i* in document (item) *j*

Term Frequency: $TF_{ij} = \frac{f_{ij}}{\max_k f_{kj}}$

Inverse Document Frequency: $IDF_i = \log_2(N/n_i)$

TF-IDF score: $w_{ij} = TF_{ij} \times IDF_i$