

3-3: TFIDF and More!



UNIVERSITY OF MINNESOTA

Introduction to Recommender Systems

Learning Objectives

- To understand the problem that requires a weighting for search or filtering
- To understand TFIDF weighting in detail, and how it is used in both search and filtering
- To understand the range of variants and alternatives to TFIDF
- To appreciate the similarities and differences between content filtering and search

Introduction to Recommender Systems



UNIVERSITY OF MINNESOTA

The search problem ...

- Why do primitive search engines fail?
- What would a primitive search engine do?
 - Return all documents that contain search terms?
 - More frequent occurrence ranked higher?
- At a minimum, need to consider two factors
 - Term frequency may be significant
 - Not all terms equally relevant
- Actually, much harder than this, more later ...

Introduction to Recommender Systems



TFIDF weighting

- Term Frequency * Inverse Document Frequency
- Term Frequency =
 - Number of occurrences of a term in the document (can be a simple count)
- Inverse Document Frequency =
 - How few documents contain this term
 - Typically

$$\log (\# \text{documents} / \# \text{documents with term})$$

Introduction to Recommender Systems



What does TFIDF do?

- Automatic demotion of stopwords, common terms
- Promotes core terms over incidental ones

But where does it fail?

- If core term/concept isn't actually used (much) in document (e.g., legal contracts)
- Poor searches (other techniques for that)

Introduction to Recommender Systems



How does TFIDF apply to CBF?

- TFIDF concept can be used to create a profile of a document/object
 - A movie could be described as a weighted vector of its tags (details next lecture)
- These TFIDF profiles can be combined with ratings to create user profiles, and then matched against future documents

Introduction to Recommender Systems



Variants and Alternatives

- Some applications use variants on TF
 - 0/1 boolean frequencies (occurs above threshold)
 - Logarithmic frequencies ($\log (tf+1)$)
 - Normalized frequency (divide by document length)
- BM25 (aka Okapi BM25) is a ranking function used by search engines:
 - Includes frequency in query, in document, number of documents, length
 - Variants with different weights: BM11, BM15, ...

Introduction to Recommender Systems



Actually much harder, as we said

- Phrases and n-grams
 - “computer science” != “computer” and “science”
 - Adjacency
- Significance in Documents
 - Titles, headings, ...
- General Document Authority
 - Pagerank and similar approaches
- Implied Content
 - Links, usage ...

Introduction to Recommender Systems



Take-Away and Moving Forward

- You should
 - Understand TFIDF and why it is needed
 - Also understand its limitations
- Next
 - Building and applying content profiles

Introduction to Recommender Systems



3-3: TFIDF and More!



UNIVERSITY OF MINNESOTA

Introduction to Recommender Systems