SNS COLLEGE OF ENGINEERING

Kurumbapalayam (Po), Coimbatore – 641 107

An Autonomous Institution

Accredited by NAAC – UGC with ‘A’ Grade
Approved by AICTE, New Delhi & Affiliated to Anna University, Chennai

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

COURSE NAME : 19CS732 INFORMATION RETRIEVAL TECHNIQUES

IVYEAR / VIII SEMESTER

Unit 2- MODELING AND RETRIEVAL EVALUATION

Topic 2 : Vector Space Model

Problem

- ➢How to determine important words in a document?
  - ➢Word sense?
  - ➢Word *n*-grams (and phrases, idioms,…) → terms
- ➢How to determine the degree of importance of a term within a document and within the entire collection?
- ➢How to determine the degree of similarity between a document and the query?
- ➢In the case of the web, what is the collection and what are the effects of links, formatting information, etc.?

Vector Space Model

- The Vector Space Model (VSM) is a way of representing documents through the words that they contain
- It is a standard technique in Information Retrieval
- The VSM allows decisions to be made about which documents are similar to each other and to keyword queries

How it works: Overview

- Each document is broken down into a word frequency table
- The tables are called vectors and can be stored as arrays
- A vocabulary is built from all the words in all documents in the system
- Each document is represented as a vector based against the vocabulary

Vector Space Model -Cont..

- Document A
  - “A dog and a cat.”

a	dog	and	cat
2	1	1	1

- Document B
  - “A frog.”

a	frog
1	1

- ➢The vocabulary contains all words used
  - ➢a, dog, and, cat, frog
- ➢The vocabulary needs to be sorted
  - ➢a, and, cat, dog, frog

Vector Space Model -Cont..

Document A: “A dog and a cat.”

a	and	cat	dog	frog
2	1	1	1	0

Vector: (2,1,1,1,0)

Document B: “A frog.”

a	and	cat	dog	frog
1	0	0	0	1

Vector: (1,0,0,0,1)

- Queries can be represented as vectors in the same way as documents:
  - Dog = (0,0,0,1,0)
  - Frog = (0,0,0,0,1)
  - Dog and frog = (0,0,0,1,1)

Vector Space Model -Cont..

- Define:
  - ❑*wij > 0* whenever *ki* $\in$ *dj*
  - ❑*wiq >= 0* associated with the pair *(ki,q)*
  - ❑ *vec(dj) = (w1j, w2j, ..., wtj)* *vec(q) = (w1q, w2q, ..., wtq)*
  - ❑To each term *ki,* associate a unit vector *vec(i)*
  - ❑ The *t* unit vectors, *vec(1), ..., vec(t)* form an *orthonormal basis* (embodying independence assumption) for the t-dimensional space for representing queries and documents

Vector Space Model -Cont..

- How to compute the weights *wij* and *wiq* ?
  - ❑quantification of intra-document content (similarity/semantic emphasis)
    - *tf* factor, the *term frequency* within a document
  - ❑quantification of inter-document separation (dis-similarity/significant discriminant)
    - *idf* factor, the *inverse document frequency*
  - ❑*wij = tf(i,j) * idf(i)*

Vector Space Model -Cont..

Let,

N be the total number of docs in the collection

ni be the number of docs which contain *ki*

freq(i,j) raw frequency of *ki* within *dj*

A normalized *tf* factor is given by

$f(i,j) = freq(i,j) / max(freq(l,j))$

where the maximum is computed over all terms which occur within the document *dj*

The *idf* factor is computed as

$idf(i) = log (N/ni)$

the *log* makes the values of *tf* and *idf* comparable.

Vector Space Model -Cont..

Represent documents and queries as

- Vectors of term-based features
  - Features: tied to occurrence of terms in collection
- E.g. $\vec{d}_j = (t_{1,j}, t_{2,j}, \ldots, t_{N,j}); \vec{q}_k = (t_{1,k}, t_{2,k}, \ldots, t_{N,k})$

Solution 1: Binary features: t=1 if presense, 0 otherwise

- Similiarity: number of terms in common
  - Dot product

$$sim(\vec{q}_k, \vec{d}_j) = \sum_{i=1}^{N} t_{i,k} t_{i,j}$$

Vector Space Model -Cont..

- Problem: Not all terms equally interesting
  - E.g. the vs dog vs Levow

$$\vec{d}_j = (w_{1,j}, w_{2,j}, \ldots, w_{N,j}); \vec{q}_k = (w_{1,k}, w_{2,k}, \ldots, w_{N,k})$$

- Solution: Replace binary term features with weights
  - Document collection: term-by-document matrix
  - View as vector in multidimensional space
    - Nearby vectors are related
  - Normalize for vector length

Vector Space Model -Cont..

Similarity = Dot product

$$sim(\vec{q}_k,\vec{d}_j)=\vec{q}_k \bullet \vec{d}_j = \sum_{i=1}^{N} w_{i,k}w_{i,j}$$

Normalization:

Normalize weights in advance

Normalize post-hoc

$$sim(\vec{q}_k,\vec{d}_j)=\frac{\sum_{i=1}^{N} w_{i,k}w_{i,j}}{\sqrt{\sum_{i=1}^{N} w_{i,k}^2}\sqrt{\sum_{i=1}^{N} w_{i,j}^2}}$$

Activity

Disadvantages

- assumes independence of index terms; not clear that this is bad though

Advantages

- ➢term-weighting improves answer set quality
- ➢partial matching allows retrieval of docs that approximate the query conditions
- ➢cosine ranking formula sorts documents according to degree of similarity to the query

Assessment 1

1. List out the Advantages of Vector Space Model

   a)______________________________

   b)______________________________

   c)______________________________

   d)______________________________

2. Identify the disadvantages of Vector Space Model

   a)______________________________

   b)______________________________

   c)______________________________

   d)______________________________

TEXT BOOKS:

1. Ricardo Baeza-Yates and Berthier Ribeiro-Neto, —Modern Information Retrieval: The Concepts and Technology behind Search, Second Edition, ACM Press Books, 2011.

2. Ricci, F, Rokach, L. Shapira, B.Kantor, —Recommender Systems Handbook||, First Edition, 2011.

REFERENCES:

1. C. Manning, P. Raghavan, and H. Schütze, —Introduction to Information Retrieval, Cambridge University Press, 2008.

2. Stefan Buettcher, Charles L. A. Clarke and Gordon V. Cormack, —Information Retrieval: Implementing and Evaluating Search Engines, The MIT Press, 2010.

THANK YOU