Recuperação de Informação e Web Mining

1

Melhorando a Recuperação de Informação

- O modelo de espaço vetorial
  - Utiliza pesos para termos
  - Permite o ranqueamento dos resultados
  - Pode reduzir a dimensão do espaço selecionando termos

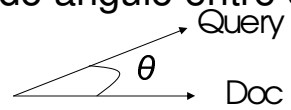
2

O Modelo de Representação Vetorial

- Os documentos e a query são ambos documentos do tipo:

$$d_i = (w_{i1}, w_{i2}, w_{i3}, \ldots)$$

- cada w_{ij} representa o peso do termo j dentro do documento i
- Um documento é um "bag" de termos
- Similaridade entre um documento e a query = coseno do ângulo entre eles



3

Medida de Similaridade do Coseno

Produto Vetorial

$$\cos(\vec{d_j}, \vec{q}) = \frac{\vec{d_j} \bullet \vec{q}}{\left|\vec{d_j}\right| \cdot \left|\vec{q}\right|} = \frac{\sum_{i=1}^{t} (w_{ij} \cdot w_{iq})}{\sum_{i=1}^{t} w_{ij}^{2} \cdot \sum_{i=1}^{t} w_{iq}^{2}}$$

Normalização

- Coseno é o produto vetorial normalizado
- Documentos são ranqueados com a ordem decrescente do coseno
  - sim(d,q) = 1 quando d = q
  - sim(d,q) = 0 quando d e q não compartilham termos

4

Ponderação de Termos

- Alto peso = maior impacto no coseno
- É necessário atribuir maior peso aos termos mais “importantes”
- O que é um termo importante?
  - Se o termo está na query, então a presença desse termo aumenta a relevância de um documento
  - Como podemos modelar isso?

5

Analogia a Clustering

- Os documentos formam uma coleção C objetos
- Uma query é uma coleção vaga de um subconjunto A de C
- Problema de IR: particionar C em A e ¬A
- Nós queremos determinar
  - quais características descrevem melhor os membros de A
  - quais características discriminam melhor os membros de A e ¬A
- Para documentos
  - Freqüência de um termo em um documento
  - Freqüência de um termo na coleção

6

Fator de Freqüência (TF)

- Quão bem um termo descreve um documento
  - Se um termo *t aparece* com freqüência em um documento, então uma query contendo *t* deve recuperar aquele documento
  - Palavras freqüentes (non-stop) são temáticas
    - Ex.: fluxo, markoviano, camada, pressão

f_{ij} = frequência do termo *i* no documento *j*

- Frequência pode estar normalizada ou não:

$tf_{ij} = f_{ij} / max\{f_{ij}\}$ ou $tf_{ij} = f_{ij}$

7

Fator de Freqüência Inversa no Documento (IDF)

- A escassez de um termo em relação a coleção de documentos é uma medida de importância
  - Importância é inversamente proporcional à freqüência de ocorrência nos documentos
- Termos que aparecem em muitos documentos diferentes são menos indicativos de um assunto geral.

df_i = freqüência de documento do termo *i*
= número de documentos que contém o termo *i*

idf_i = freqüência inversa do document do termo *i,*
= $\log_2 (N/ df_i)$
(N: Número total de documentos)

- Uma indicação do poder discriminante do termo
- Usada para diminuir o efeito de *tf*.

8

Ponderação TF-IDF

- Um esquema de ponderação onde

$$w_{d,t} = tf_{d,t} \times idf_t$$

- É o chamado esquema *tf-idf*
- Um termo ocorrendo frequentemente em poucos documentos tem um peso alto
- Muitas outras formas de determinação do peso de termos foram desenvolvidas.
- Experimentalmente, *tf-idf* tem trabalhado bem para coleções de documentos em geral

9

Exemplo de ponderação TF-IDF

Dado um documento contendo termos com as seguintes freqüências:

A(3), B(2), C(1)

Assumindo que a coleção tem 10.000 docs e freqüência de documentos destes termos são:

A(50), B(1300), C(250)

Então:

A: tf = 3/3; idf = $\log_2(10000/50)$ = 7,64; tf-idf = 7,64

B: tf = 2/3; idf = $\log_2(10000/1300)$ = 2,94; tf-idf = 1,96

C: tf = 1/3; idf = $\log_2(10000/250)$ = 5,32; tf-idf = 1,77

10

Exemplo detalhado de representação vetorial

Doc	Conteúdo
1	Apple ballon ballon elephant apple apple
2	Chocolate ballon ballon chocolate appel chocolate duck
3	Ballon ballon ballon ballon elephant ballon
4	Chocolate ballon elephant
5	Ballon apple chocolate ballon
6	Elephant elephant elephant chocolate elephant

Corpus

Valores TF

Docs	Termos				
	a	b	c	d	e
1	3	2	0	0	1
2	1	2	3	1	0
3	0	5	0	0	1
4	0	1	1	0	1
5	1	2	1	0	0
6	0	0	1	0	4
IDF_t:	$\log_2(6/3)=1.0$	$\log_2(6/5)=0.26$	$\log_2(6/4)=0.58$	$\log_2(6/1)=2.58$	$\log_2(6/4)=0.58$

11

Exemplo detalhado de representação vetorial

Doc	Queries				
	d	c	c,d	a,b,e	a,b,c,d,e
1	0	0	0	**0.95**	0.39
2	**0.78**	0.53	**0.88**	0.29	**0.90**
3	0	0	0	0.40	0.16
4	0	**0.67**	0.15	0.40	0.30
5	0	0.46	0.1	0.76	0.40
6	0	0.24	0.05	0.48	0.24

$$\cos(Q=(a,b,c,d,e), Doc=2) = \frac{Q \cdot Doc}{|Q| \times |Doc|}$$

$$Q = ((1\times1),(1\times0.26),(1\times0.58),(1\times2.58),(1\times0.58))$$

$$Doc = ((1\times1.0),(2\times0.26),(3\times0.58),(1\times2.58),(0\times0.58))$$

$$Q \cdot Doc = 1+0.26\times(2\times0.26)+0.58\times(3\times0.58)+2.58\times(1\times2.58)+0.58\times(0\times0.58) = 8,8088$$

$$|Q| = \sqrt{1+0.26^2+0.58^2+2.58^2+0.58^2} = 2,8977$$

$$|Doc| = \sqrt{1+0.52^2+1.74^2+2.58^2+0.58^2} = 3,36$$

$$\cos(Q, Doc) = 0,90$$

12

Exemplo detalhado de representação vetorial

Doc	Queries				
	d	c	c,d	a,b,e	a,b,c,d,e
1	0	0	0	**0.95**	0.39
2	**0.78**	0.53	**0.88**	0.29	**0.90**
3	0	0	0	0.40	0.16
4	0	**0.67**	0.15	0.40	0.30
5	0	0.46	0.1	0.76	0.40
6	0	0.24	0.05	0.48	0.24

$$\cos(Q = \{a,b,c,d,e\}, Doc = 3) = \frac{Q \cdot Doc}{|Q| \times |Doc|}$$

$$Q = \{(1\times 1),(1\times 0.26),(1\times 0.58),(1\times 2.58),(1\times 0.58)\}$$

$$Doc = \{(0\times 1.0),(5\times 0.26),(0\times 0.58),(0\times 2.58),(1\times 0.58)\}$$

$$Q \cdot Doc = 0 + (5\times 0.26\times 0.26)\times 0 + 0 + (1\times 0.58)\times(1\times 0.58) = 0{,}67$$

$$|Q| = \sqrt{1 + 0.26^2 + 0.58^2 + 2.58^2 + 0.58^2} = 2{,}8977$$

$$|Doc| = \sqrt{0 + 1{,}30^2 + 0 + 0 + 0{,}58^2} = 1{,}42$$

$$\cos(Q, Doc) = 0{,}16$$

13

Modelo vetorial

- Vantagens
  - Recuperação é ranqueada
  - Termos são ponderados por importância
  - Permite matchings parciais
- Desvantagens
  - Assume que termos são independentes
  - Ponderação é intuitiva mas não é um modelo formal

14

Algoritmo básico de recuperação

- Converta todos os documentos na coleção D para vetores ponderados $\boldsymbol{d_j}$ com TF-IDF usando cada termo de V (coleção de documentos).
- Converta a query para um vetor ***q*** ponderado com TF-IDF.
- Para cada $\boldsymbol{d_j}$ em D faça
- Calcule o score $s_j = \cos(\boldsymbol{q}, \boldsymbol{d_j})$
- Ordene os documentos em ordem decrescente.
- Apresente os *n* primeiros documentos.

15