Generalized Vector Space Model (GVSM)

SUBMITTED BY

Nasrin Jaleel	2011A7PS444G
Saransh Varshneya	2011A7PS007G
Kunal Baweja	2011A7PS029G
Ashu Kalra	2011A7PS150G

Motivation

- Classic models enforce independence of index terms
- For the Vector model:
  - Set of term vectors {~k1, ~k2, . . ., ~kt} are linearly independent and form a basis for the subspace of interest
  - Frequently, this is interpreted as: ∀i,j ⇒~ki • ~kj = 0

Key Idea

- In the generalized vector model, two index terms might be non-orthogonal and are represented in terms of smaller components (minterms)
- As before let,
  - wi,j be the weight associated with [ki, dj]
  - {k1, k2, . . ., kt} be the set of all terms
- If these weights are all binary, all patterns of occurrence of terms within documents can be represented by the minterms:
  - m1 = (0, 0, . . . , 0), m2 = (1, 0, . . . , 0)
  - m2t = (1, 1, . . . , 1)
  - In here, m2 indicates documents in which solely the term k1 occurs

Key Idea

- The basis for the generalized vector model is formed by a set of 2t vectors defined over the set of minterms, as follows:

$$\sim m1 = (1, 0, \ldots, 0, 0)$$

$$\sim m2 = (0, 1, \ldots, 0, 0)$$

$$\ldots$$

$$\sim m2t = (0, 0, \ldots, 0, 1)$$

- Notice that,

$\forall i,j \Rightarrow \sim mi \bullet \sim mj = 0$ i.e., pairwise orthogonal

Key Idea

- Minterm vectors are pairwise orthogonal. But, this does not mean that the index terms are independent:
  - The minterm m4 is given by: m4 = (1, 1, 0, . . . , 0)
  - This minterm indicates the occurrence of the terms k1 and k2 within a same document. If such document exists in a collection, we say that the minterm m4 is active and that a dependency between these two terms is induced
  - The generalized vector model adopts as a basic foundation the notion that co-occurence of terms within documents induces dependencies among them

Forming the Term Vectors

- The vector associated with the term ki is computed as:

$$\vec{k}_i = \frac{\sum_{\forall r,\ g_i(m_r)=1} c_{i,r}\ \vec{m}_r}{\sqrt{\sum_{\forall r,\ g_i(m_r)=1} c_{i,r}^2}}$$

$$c_{i,r} = \sum_{d_j \mid g_l(\vec{d}_j)=g_l(m_r)\ for\ all\ l} w_{i,j}$$

- The weight Ci,r associated with the pair [ki,mr] sums up the weights of the term ki in all the documents which have a term occurrence pattern given by mr.
- Notice that for a collection of size N, only N minterms affect the ranking (and not 2t)

Dependency between Index Terms

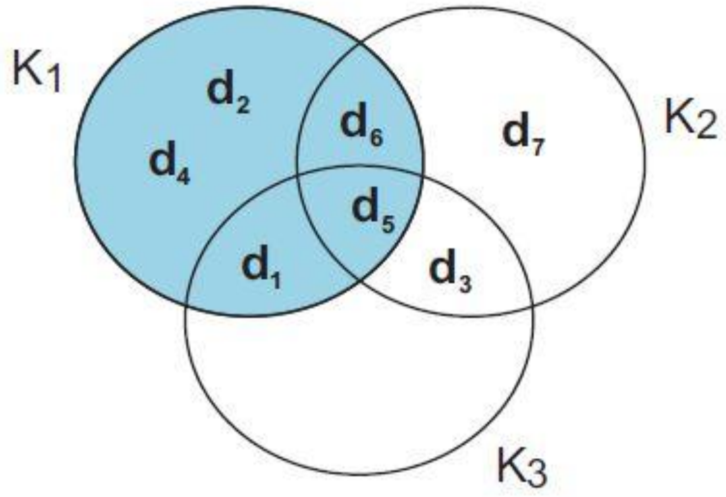
- A degree of correlation between the terms ki and kj can now be computed as:

$$\vec{k}_i \bullet \vec{k}_j = \sum_{\forall r \mid g_i(m_r)=1 \wedge g_j(m_r)=1} c_{i,r} \times c_{j,r}$$

- This degree of correlation sums up (in a weighted form) the dependencies between ki and kj induced by the documents in the collection (represented by the mr minterms).

The Generalized Vector Model

- An Example



	K_1	K_2	K_3
d_1	2	0	1
d_2	1	0	0
d_3	0	1	3
d_4	2	0	0
d_5	1	2	4
d_6	1	2	0
d_7	0	5	0
q	1	2	3

Computation of Ci,r

	K_1	K_2	K_3
d_1	2	0	1
d_2	1	0	0
d_3	0	1	3
d_4	2	0	0
d_5	1	2	4
d_6	0	2	2
d_7	0	5	0
q	1	2	3

	K_1	K_2	K_3
$d_1 = m_6$	1	0	1
$d_2 = m_2$	1	0	0
$d_3 = m_7$	0	1	1
$d_4 = m_2$	1	0	0
$d_5 = m_8$	1	1	1
$d_6 = m_7$	0	1	1
$d_7 = m_3$	0	1	0
$q = m_8$	1	1	1

	$C_{1,r}$	$C_{2,r}$	$C_{3,r}$
m_1	0	0	0
m_2	3	0	0
m_3	0	5	0
m_4	0	0	0
m_5	0	0	0
m_6	2	0	1
m_7	0	3	5
m_8	1	2	4

Computation of Index Term Vectors

$$k_1 = \frac{(3m_2+2m_6+m_8)}{\sqrt{3^2+2^2+1^2}}$$

$$k_2 = \frac{(5m_3+3m_7+2m_8)}{\sqrt{5+3+2}}$$

$$k_3 = \frac{(1m_6+5m_7+4m_8)}{\sqrt{1+5+4}}$$

	$C_{1,r}$	$C_{2,r}$	$C_{3,r}$
m_1	0	0	0
m_2	3	0	0
m_3	0	5	0
m_4	0	0	0
m_5	0	0	0
m_6	2	0	1
m_7	0	3	5
m_8	1	2	4

Computation of Document Vectors

$$d_1 = 2k_1 + k_3$$

$$d_2 = k_1$$

$$d_3 = k_2 + 3k_3$$

$$d_4 = 2k_1$$

$$d_5 = k_1 + 2k_2 + 4k_3$$

$$d_6 = 2k_2 + 2k_3$$

$$d_7 = 5k_2$$

$$q = k_1 + 2k_2 + 3k_3$$

	K_1	K_2	K_3
d_1	2	0	1
d_2	1	0	0
d_3	0	1	3
d_4	2	0	0
d_5	1	2	4
d_6	0	2	2
d_7	0	5	0
q	1	2	3

Conclusions

- Model considers correlations among index terms
- Computation costs are higher
- Model does introduce interesting new ideas