Generalized Vector Space Model In Information Retrieval

SIGIR 1985 by S.K.M. Wong, Wojciech Ziarko and Patrick C.N. Wong

報告：志豪

Outline

- Introduction
- Vector Space Model
- Generalized Vector Space Model
- Experimental

Introduction

- Vector Space Model 假設term vector間爲orthogonal.
- 計算term correlations的方法很多, 但不是計算量太大, 就是要對於現存的系統做大幅度的修改
- 在最小修改的限度下, 加入correlations between term vector

Vector Space Model

- The documents and the query are represented by a set of vectors.

$$\vec{d}_\alpha = \sum_{i=1}^{n} a_{\alpha i}\vec{t}_i , \qquad (\alpha = 1, 2, \ldots, p)$$

$$\vec{q} = \sum_{j=1}^{n} q_j \vec{t}_j$$

- The similarity of documents to the query

$$\vec{d}_\alpha \cdot \vec{q} = \sum_{i=1,\, j=1}^{n} a_{\alpha i} q_j \vec{t}_i \cdot \vec{t}_j , \quad (\alpha = 1, 2, \ldots, p)$$

$$\vec{t}_i \cdot \vec{t}_j \begin{cases} 1, \text{ if } i == j \\ 0, \text{ if } i != j \end{cases}$$

$$= a_{\alpha i} q_1 \vec{t}_1 \vec{t}_1 + a_{\alpha i} q_1 \vec{t}_2 \vec{t}_1 + \ldots\ldots + a_{\alpha i} q_1 \vec{t}_n \vec{t}_1 +$$

$$a_{\alpha i} q_2 \vec{t}_1 \vec{t}_2 + a_{\alpha i} q_2 \vec{t}_2 \vec{t}_2 + \ldots\ldots + a_{\alpha i} q_2 \vec{t}_n \vec{t}_2 +$$

.

.

$$a_{\alpha i} q_n \vec{t}_1 \vec{t}_n + a_{\alpha i} q_n \vec{t}_2 \vec{t}_n + \ldots\ldots + a_{\alpha i} q_n \vec{t}_n \vec{t}_n$$

Generalized Vector Space Model

- Key idea
  - 整體與Vector Space Model相似, 但拿掉term vector間爲independent的限制. 把term vector用minterm展開, 而minterm vector間爲independent.
  - Minterm : Each represent one kind of co-occurrence of index terms in aspecific document

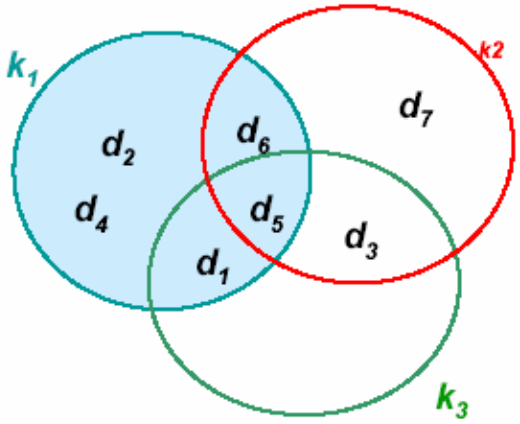
$$\vec{k}_i = \frac{\sum_{\forall r, g_i(m_r)=1} c_{i,r} \vec{m}_r}{\sqrt{\sum_{\forall r, g_i(m_r)=1} c_{i,r}^2}}$$

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

Generalized Vector Space Model

- **Example** (a system with three index terms)

minterm	k_1	k_2	k_3
m_1	0	0	0
m_2	1	0	0
m_3	0	1	0
m_4	1	1	0
m_5	0	0	1
m_6	1	0	1
m_7	0	1	1
m_8	1	1	1



$$\vec{k}_1 = \frac{c_{1,2}\vec{m}_2 + c_{1,4}\vec{m}_4 + c_{1,6}\vec{m}_6 + c_{1,8}\vec{m}_8}{\sqrt{c_{1,2}^{\ 2} + c_{1,4}^{\ 2} + c_{1,6}^{\ 2} + c_{1,8}^{\ 2}}}$$

$$\vec{k}_2 = \frac{c_{2,3}\vec{m}_3 + c_{2,4}\vec{m}_4 + c_{2,7}\vec{m}_7 + c_{2,8}\vec{m}_8}{\sqrt{c_{2,3}^{\ 2} + c_{2,4}^{\ 2} + c_{2,7}^{\ 2} + c_{2,8}^{\ 2}}}$$

$$\vec{k}_3 = \frac{c_{3,5}\vec{m}_5 + c_{3,6}\vec{m}_6 + c_{3,7}\vec{m}_7 + c_{3,8}\vec{m}_8}{\sqrt{c_{3,5}^{\ 2} + c_{3,6}^{\ 2} + c_{3,7}^{\ 2} + c_{3,8}^{\ 2}}}$$

	k_1	k_2	k_3	minterm
d_1	2	0	1	m_6
d_2	1	0	0	m_2
d_3	0	1	3	m_7
d_4	2	0	0	m_2
d_5	1	2	4	m_8
d_6	1	2	0	m_4
d_7	0	5	0	m_3
q	1	2	3	

$$c_{1,2} = w_{1,2} + w_{1,4} = 1 + 2 = 3$$
$$c_{1,4} = w_{1,6} = 1$$
$$c_{1,6} = w_{1,1} = 2$$
$$c_{1,8} = w_{1,5} = 1$$

$$\vec{k}_1 = \frac{3\vec{m}_2 + 1\vec{m}_4 + 2\vec{m}_6 + 1\vec{m}_8}{\sqrt{3^2 + 1^2 + 2^2 + 1^2}}$$

$$c_{2,3} = w_{2,7} = 5$$
$$c_{2,4} = w_{2,6} = 2$$
$$c_{2,7} = w_{2,3} = 1$$
$$c_{2,8} = w_{2,5} = 2$$

$$\vec{k}_2 = \frac{5\vec{m}_3 + 2\vec{m}_4 + 1\vec{m}_7 + 2\vec{m}_8}{\sqrt{5^2 + 2^2 + 1^2 + 2^2}}$$

$$c_{3,5} = 0$$
$$c_{3,6} = w_{3,1} = 1$$
$$c_{3,7} = w_{3,3} = 3$$
$$c_{3,8} = w_{3,5} = 4$$

$$\vec{k}_3 = \frac{0\vec{m}_5 + 1\vec{m}_6 + 3\vec{m}_7 + 4\vec{m}_8}{\sqrt{0^2 + 1^2 + 3^2 + 4^2}}$$

Generalized Vector Space Model

- Problem : minterm 無法計算

$$\overline{d_1} = 0\overline{k_1} + 2\overline{k_2} + \overline{6k_3} + 0\overline{k_4} + \overline{k_5} + \overline{2k_6}$$

$$=> 0*2^1 + 1*2^2 + 1*2^3 + 0*2^4 + 1*2^5 + 1*2^6 + 1 = 108$$

$$=> m_{108}$$

- Solution : LogAdd()

$$\log(0*2^1 + 1*2^2 + 1*2^3 + 0*2^4 + 1*2^5 + 1*2^6 + 1) = \log(108)$$

$$=> \log(2^2) + \log(2^3) + \log(2^5) + \log(2^6) = \log(108)$$

$$\log(x + y) = \log(x) + \log(1.0 + \exp(\log(x) - \log(y)))$$

- Solution : term vector 比對
  - minterm間爲indeptendent
  - 重複的minterm
  - Minterm與term的關係

Generalized Vector Space Model

$$\vec{d_1} = \vec{k_1} + \vec{k_2}$$
$$\vec{q} = \vec{k_1} + 2\vec{k_3}$$

$$\vec{k_1} = c_{11}\vec{m_1} + c_{12}\vec{m_2}$$
$$\vec{k_2} = c_{22}\vec{m_2} + c_{23}\vec{m3}$$
$$\vec{k_3} = c_{31}\vec{m1} + c_{33}\vec{m_3}$$

$$\begin{aligned}\vec{d_1} &= c_{11}\vec{m_1} + c_{12}\vec{m_2} + c_{22}\vec{m_2} + c_{23}\vec{m3} \\ &= c_{11}\vec{m_1} + (c_{12} + c_{22})\vec{m_2} + c_{23}\vec{m3}\end{aligned}$$

$$\begin{aligned}\vec{q} &= c_{11}\vec{m_1} + c_{12}\vec{m_2} + 2c_{31}\vec{m1} + 2c_{33}\vec{m_3} \\ &= (c_{11} + 2c_{31})\vec{m_1} + c_{12}\vec{m_2} + 2c_{33}\vec{m_3}\end{aligned}$$

$$\begin{aligned}\vec{d_1} * \vec{q} &= \vec{k_1}\vec{k_1} + 2\vec{k_1}\vec{k_3} + \vec{k_2}\vec{k_1} + 2\vec{k_2}\vec{k_3} \\ &= \frac{c_{11}^2 + c_{12}^2}{\sqrt{c_{11}^2 + c_{12}^2}} + \frac{2c_{11}c_{31}}{\sqrt{c_{11}^2 + c_{12}^2}\sqrt{c_{31}^2 + c_{33}^2}} + \frac{c_{12}c_{22}}{\sqrt{c_{11}^2 + c_{12}^2}\sqrt{c_{22}^2 + c_{23}^2}} + \frac{2c_{23}c_{33}}{\sqrt{c_{22}^2 + c_{23}^2}\sqrt{c_{31}^2 + c_{33}^2}}\end{aligned}$$

$$\begin{aligned}\vec{d_1} * \vec{q} &= \frac{c_{11}(c_{11} + 2c_{31}) + c_{12}(c_{12} + c_{22}) + 2c_{23}c_{33}}{\sqrt{c_{11}^2 + (c_{12} + c_{22})^2 + c_{23}^2}\sqrt{(c_{11} + 2c_{31})^2 + c_{12}^2 + 2c_{33}^2}} \\ &= \frac{c_{11}^2 + c_{12}^2 + 2c_{11}c_{31} + c_{12}c_{22} + 2c_{23}c_{33}}{\sqrt{c_{11}^2 + (c_{12} + c_{22})^2 + c_{23}^2}\sqrt{(c_{11} + 2c_{31})^2 + c_{12}^2 + 2c_{33}^2}}\end{aligned}$$

$$\vec{S} = \vec{q}GA = \begin{bmatrix} q_{11}\ q_{12}\ \dots\ q_{1n} \\ q_{21}\ q_{22}\ \dots\ q_{2n} \\ .\ .\ .\ . \\ .\ .\ .\ . \\ q_{k1}\ q_{k2}\ \dots\ q_{kn} \end{bmatrix} \begin{bmatrix} \vec{t_1}\vec{t_2}\ \vec{t_1}\vec{t_2}\ \dots\ \vec{t_1}\vec{t_n} \\ \vec{t_2}\vec{t_2}\ \vec{t_2}\vec{t_2}\ \dots\ \vec{t_2}\vec{t_n} \\ .\ .\ .\ . \\ .\ .\ .\ . \\ \vec{t_n}\vec{t_2}\ \vec{t_n}\vec{t_2}\ \dots\ \vec{t_n}\vec{t_n} \end{bmatrix} \begin{bmatrix} a_{11}\ a_{21}\ \dots\ a_{p1} \\ a_{12}\ a_{22}\ \dots\ a_{p2} \\ . \\ . \\ \\ a_{1n}\ a_{2n}\ \dots\ a_{pn} \end{bmatrix}$$

Experimental

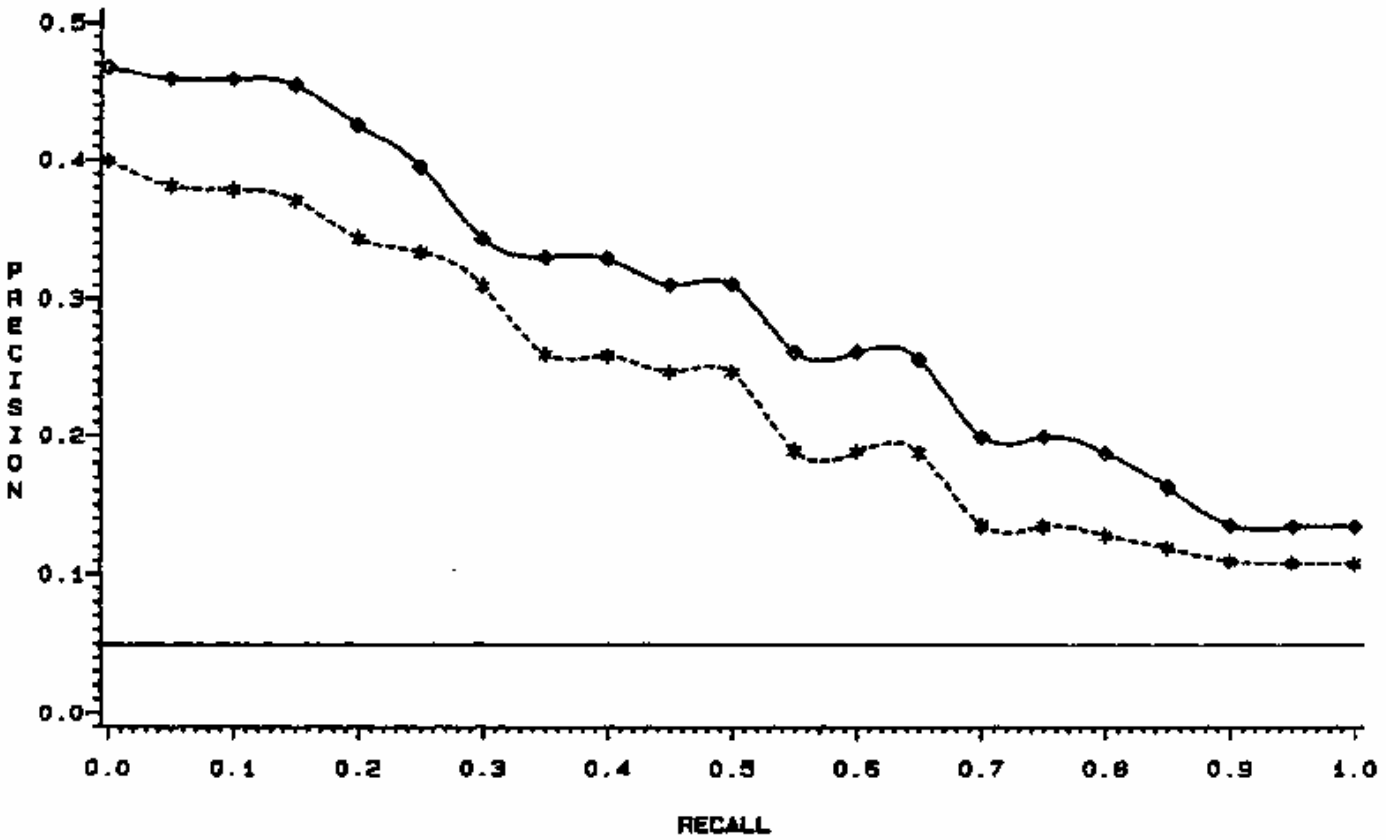
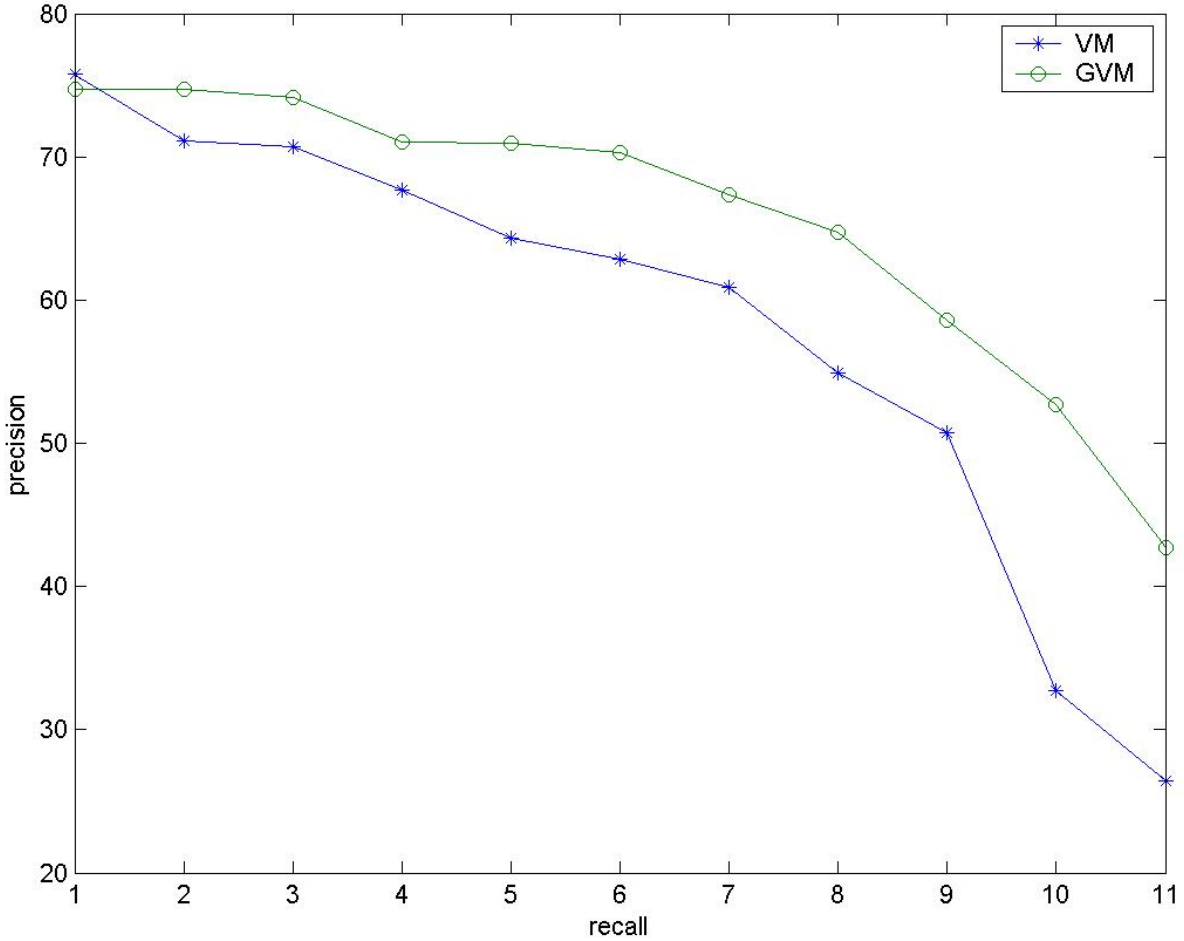


Figure 4. Comparsion of recall-precision between GVSM (solid curve) and VSM (dotted curve) in ADINUL.

Experimental



VM : map = 0.572479 **doc : teacher**

GVM : map = 0.639798 **query : jones**

Thanks