Set-Based Model: A New Approach for Information Retrieval

期刊：SIGIR 2002

研究生： 黃士傑

2003/6/30

大綱

- Introduction
- Set-Based Model (SBM) 架構
- SBM核心精神– Closed Termsets
- 範例與演算法流程
- 實驗結果
- Conclusions and future work

Introduction

- 在IR領域,vector space model is popuar,這個model的成功,大部分因為是Salton與他的同事們長期努力的結果.
- VSM中,document與query都由weighted vectors來表示,相似度的算法(ranking) is based on給予document與query中之index terms的weight.
- Term weight的算法有很多種,目前仍是個課題,目前所知求weight最佳的算法,是*tf X idf* scheme
- *tf X idf*考慮兩方面因素來計算index term的weight:

  (1)index term在文件出現次數 (2)整個collection中出現此index term的文件數

Introduction

- 這篇論文利用集合理論,提出新的model來計算index term weights,即*set-based model*.
- 所使用的term weighting scheme是based on data mining中的association rules theory(包含了*tf X idf*的要素, 並提供representative term co-occurrence patterns的量化, 這是有時在*tf X idf*所未present的)
- 最後,將這個新model與vector space model與generalized vector space比較,結果set-based model得到更好的retrieval performance, computing performance is also competitive.

Set-Based Model (SBM) 架構-ranking

$$sim(q, d_j) = \frac{\vec{d_j} \bullet \vec{q}}{|\vec{d_j}| \times |\vec{q}|} = \frac{\sum_{s \in C_q} w_{s,j} \times w_{s,q}}{\sqrt{\sum_{i=1}^{t} w_{i,j}^2} \times \sqrt{\sum_{i=1}^{t} w_{i,q}^2}},$$

C_q: the set of all closed termsets such that all $s \subseteq q$

跟我們實作過的vector model概念一樣，

只是index term換成closed termset

Set-Based Model (SBM) 架構-weight

$$w_{i,j} = sf_{i,j} \times ids_i = sf_{i,j} \times \log \frac{N}{ds_i}$$

$sf_{i,j}$: the number of occurrences of the closed termset i in document j

ids_i: the inverted frequency of occurrence of the closed termset in the collection

N: the number of documents in the collection

SBM核心精神– Closed Termsets

1. Let T={ $k_1, k_2, \ldots, k_t$ }:
   vocabulary of a collection of documents D,
   其中的vocabulary terms具有total ordering關係,
   照字典順序排列,所以$k_i < k_{i+1}$, for $1 \leq i \leq t\text{-}1$
   - 意即整個文件群D共有t個unique terms

2. n-termset s: an ordered set of n unique terms, $s \subseteq T$
   例.T={ a, b, c }, 2-termset 為{a, b}或{b, c}…..

SBM核心精神– Closed Termsets

3. $S = \{s_1, s_2 \ldots\ldots s_{2^t}\}$

S is the vocabulary-set of a collection of D

=> 每個document可能包含好幾個s_i,因為字會重覆算

4. ls_i : for each termset s_i, $1 \le i \le 2^t$, we associate and inverted list,存哪些document出現過此termset

ds_i : frequency of a termset s_i as the number of occurrences of s_i in D (dsi= $|ls_i|$)

A termset s_i is a **frequent** termset if its frquency ds_i is greater than or equal to a given threshold

SBM核心精神– Closed Termsets

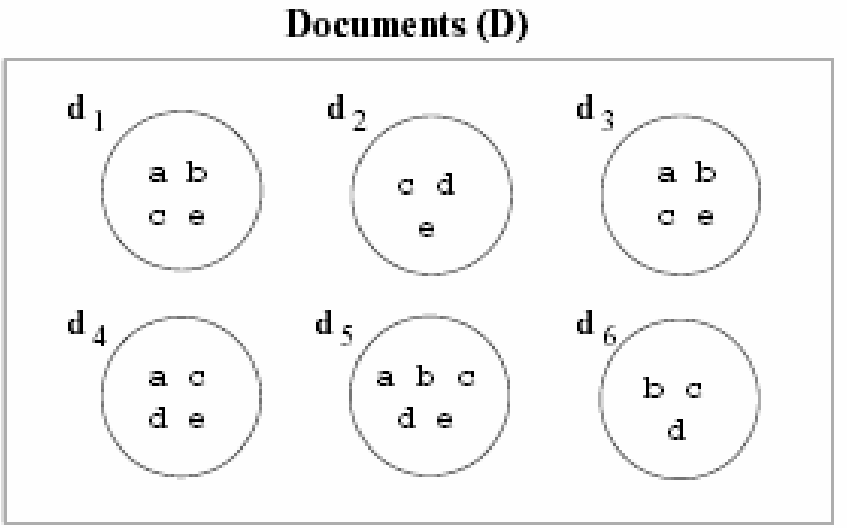
5. A **closed termset cs_i** is a frequent termset that is the largest termset among the termsets that are subsets of cs_i and occur in the same set of documents.
6. A **maximal termset ms_i** is a frequent termset that is not a subset of any other frequent termset.

已經有人證明, the set of maximal termsets associated with a codument collection are the minimum amount of information necessary to derive all frequent termsets associated with a colletion

範例

T={a,b,c,d,e}

threshold=50%



Frequency (ds)	Frequent Termsets	Closed Termsets	Maximal Termsets
100% (6)	c	c	
83% (5)	e,ce	ce	
67% (4)	a,ac,ae,ace	ace	
67% (4)	cd	cd	
67% (4)	b,bc	bc	
67% (4)	d,cd	cd	
50% (3)	ab,abc,abe be,bce,abce	acbe	abce
50% (3)	de,cde	cde	cde

演算法流程–determine closed termsets

1.1-termsets is above a given threshold?

若是,將此termset設為closed,並進入2

2.A new n+1-termset s_{new} is determined by $s_i \cup s_j$ (s_i, s_j both n-termset, have the same first n-1 terms)而產生

$l_{new} = l_i \cap l_j$

3.檢查 s_{new}是否frequent(Apriori algorithm)

原則: n-termset may be frequent only if all of its n−1-termsets are also frequent

4.若$s_{new=}$frequent,檢查是否最大,是則將較小的取消closed並將s_{new}設為closed,否則s_{new} is discared

實驗結果
三種collection的main features

Table 2: Characteristics of the reference collections

Characteristics	Collection		
	CFC	WSJ	TReC-3
Number of Documents	1,240	173,252	1,078,166
Number of Distinct Terms	2,105	230,902	1,016,709
Number of Queries	100	300	300
Average Terms per Query	3.82	18.88	22.43
Average Relevants per Query	29.04	235.99	286.89
Size (MB)	1.9	509	3,225

Retrieval Performance- CVC collection

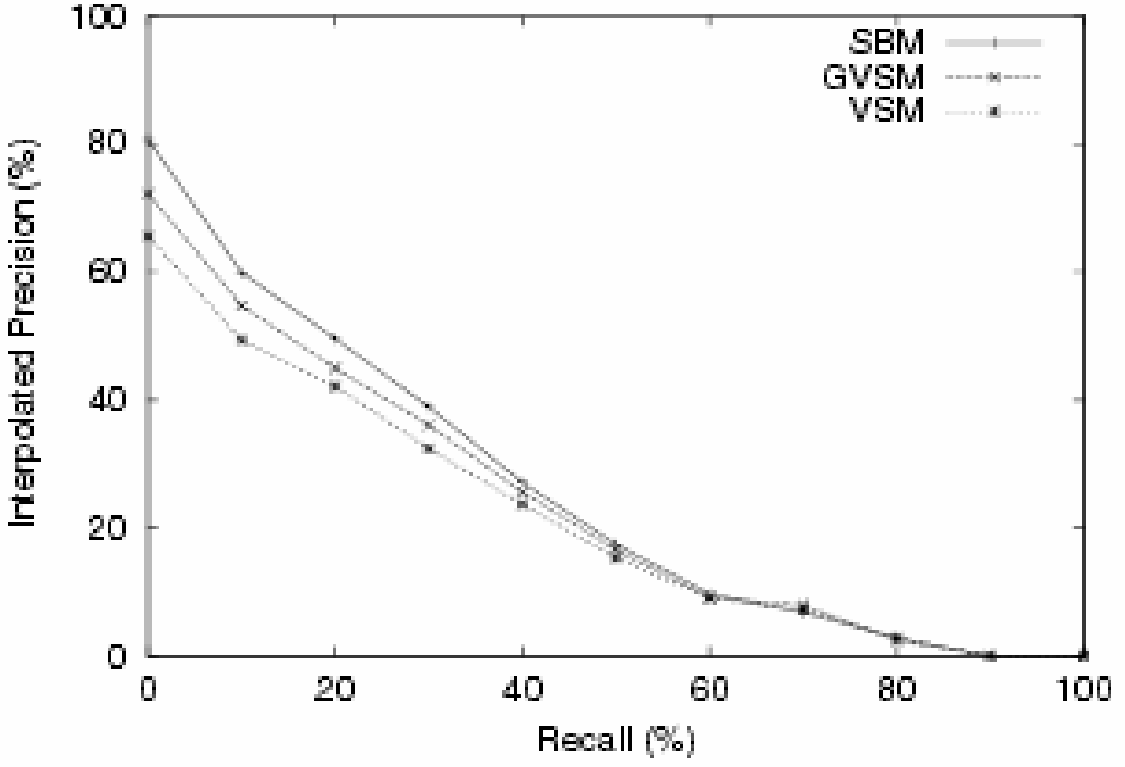
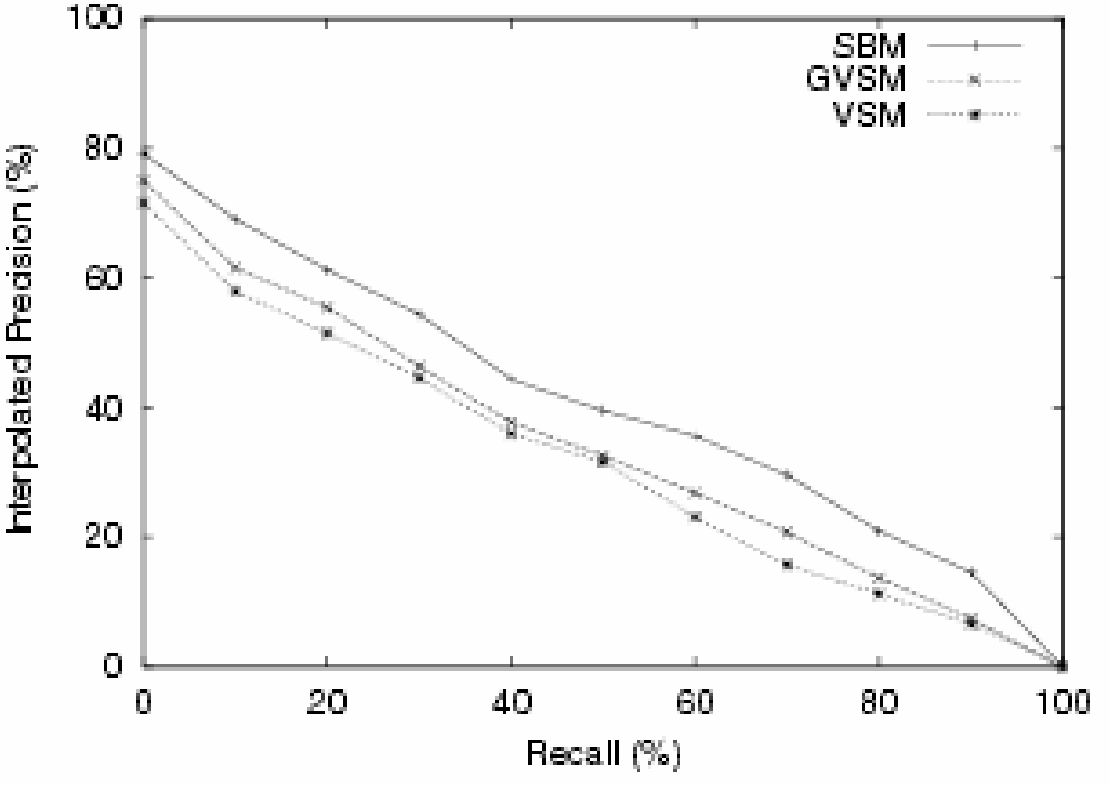


Figure 2: Recall-precision for CFC

Retrieval Performance-WSJ colletion



Retrieval Performance-TREC-3 collection

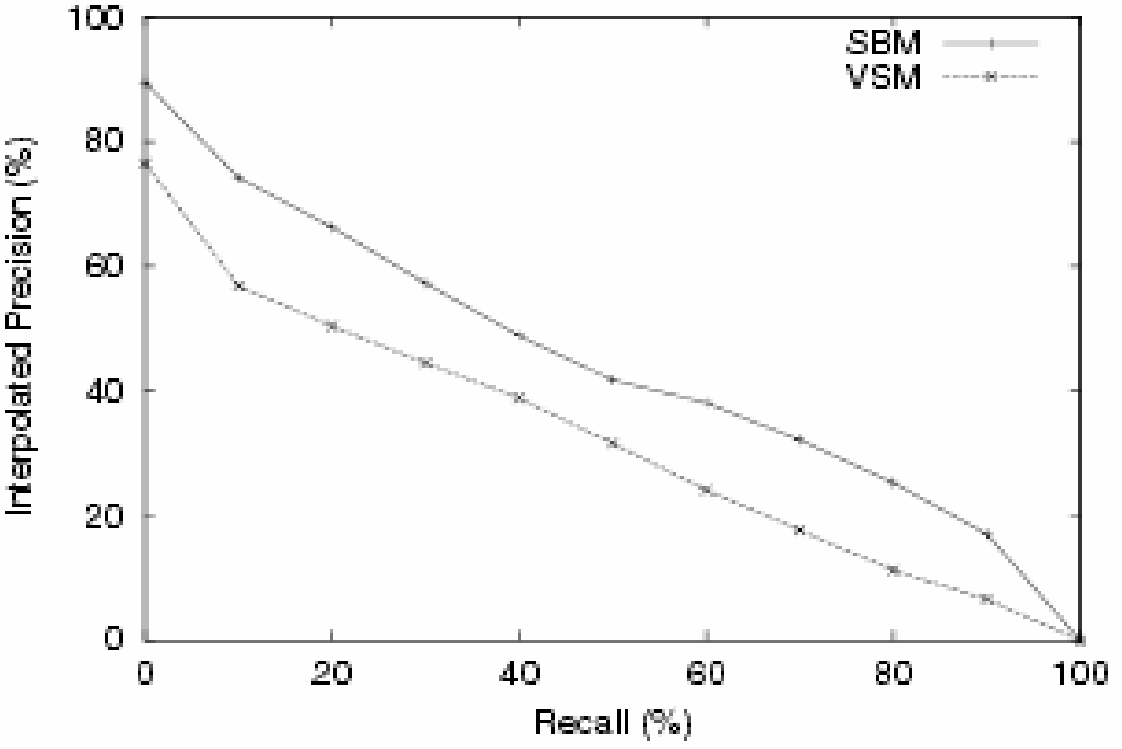


Figure 4: Recall-precision for TReC-3

Overall average precision

Table 3: Average precision curves and gain provided by the SBM model

Collection	Average Precision (%)			SBM Gain(%)	
	VSM	*GVSM*	*SBM*	*VSM*	*GVSM*
CFC	22.42	24.47	26.56	18.47	8.54
WSJ	31.76	34.27	41.78	31.55	21.91
TReC-3	32.58	*	44.59	36.86	*

* The GVSM could not be evaluated for the TReC-3 collection due to the exponential cost of the min-term build phase

Average precision of top 10 documents

Table 4: Average precision curves for top 10 documents and gains provided by the SBM model

Collection	Average Precision at 10 (%)			SBM Gain(%)	
	VSM	*GVSM*	*SBM*	*VSM*	*GVSM*
CFC	10.97	12.93	16.02	46.03	23.90
WSJ	12.71	16.58	19.17	50.82	15.62
TReC-3	13.66	*	21.42	56.80	*

* The GVSM could not be evaluated for the TReC-3 collection due to the exponential cost of the min-term build phase

Average number of closed termsets and the average list sizes while using SBM

Table 6: Average number of closed termsets and inverted list size

Collection	Closed Termsets	Inverted List Size
CFC	3.14	7.22
WSJ	3,217.28	140.10
TReC-3	4,081.25	162.06

Response time

Table 7: Average response time and response time increase

Collection	Average Response Time (s)			Increase(%)	
	VSM	*GVSM*	*SBM*	*GVSM*	*SBM*
CFC	0.0023	0.0056	0.0025	243.5	8.7
WSJ	0.4286	2.0143	0.6296	469.9	46.9
TReC-3	1.2732	*	2.2930	*	80.1

Conclusions and future work

- SBM improve retrieval effectiveness
- The computation of frequent termsets enumerated by an algorithm to generate association rules lead to a direct extension of the vector space model
- For future work we will extend SBM to account for the proximity information about query terms in documents