第四章　結果與討論

本研究利用行政院勞工委員會所編製的「電腦軟體應用技能檢定丙級學科」92 年度和 93 年度兩個版本的題庫，經過中文斷詞後利用潛在語意分析所得到的結果如下。

第一節　中文斷詞之分析

中文斷詞有未知詞與歧異性兩種問題，未知詞的問題是指題庫中有辭典未收錄的新詞，歧異性是指一個句子可能會有不同的斷詞組合。本研究先將題庫中 1000 題試題利用反向最大匹配法作斷詞後，研究者再自行判斷斷詞的結果是否正確。

在 1000 題試題中，利用反向最大匹配法斷詞的結果有 17036 個詞彙，其中錯誤的詞彙有 1746 個，正確率為 89.75%。觀察系統判斷的錯誤，屬歧異性的錯誤者，例如「重新　開　機會　造成」，只有 94 個，佔全部錯誤的 5.38%；另 94.62%的錯誤，皆屬於未知詞的錯誤。

表 4-1 歧異性與未知詞問題所佔的比率

錯誤類型	個數	比率
歧異性	94	5.38
未知詞	1652	94.62

分析其錯誤的原因，由於研究者使用「電腦軟體應用技能檢定丙級學科」的題庫，有許多電腦專有名詞，如「軟碟機、另存新檔、偶同位性等」，與一些口語化的詞，如「那一個、那項、某一種等」，這些詞彙並不在研究者所使用的辭典中，因此系統對這些詞彙無法輸出正確的斷詞結果。

由以上的結果，我們可知利用反向最大匹配法來作斷詞，有其一定的效果，顯示長詞優先是中文普遍的現象，但由於此方法只在局部區域內選擇較長的詞，並沒有考慮到整個句子的訊息，所以會有歧異性的問題。而在未知詞的部份，由於本系統並沒有加入辨識未知詞的功能，且題庫中有大量的電腦專有名詞，因而導致系統出現許多未知詞的斷詞錯誤。

第二節　向量空間模型和潛在語意分析之比較

由於行政院勞工委員會所編製的「電腦軟體應用技能檢定丙級學科」的試題，尚無研究對其做相似度的分析，因此無法得知本研究利用潛在語意分析所得到的結果是否較佳，故研究者比較向量空間模型（Vector Space Model，VSM）和潛在語意分析（latent semantic analysis，LSA）兩種資訊檢索技術，以分析何者效果較佳。

以下列出在判斷不同相似程度的試題時，取系統判斷出的前 N 組，比較利用 VSM 和 LSA，何者的精確率與召回率較佳。

表 4-2 使用 VSM 或 LSA 在判斷完全相同試題的情況時其召回率與精確率

系統判斷出的相似試題數量	使用 VSM 的召回率	使用 LSA 的召回率	使用 VSM 的精確率	使用 LSA 的精確率
50	0.1237	0.1253	0.9800	1.0000
100	0.2348	0.2506	0.9300	1.0000
150	0.3359	0.3759	0.8867	1.0000
200	0.4217	0.5013	0.8350	1.0000
250	0.5076	0.6266	0.8040	1.0000
300	0.5530	0.7393	0.7300	0.9833
350	0.6288	0.7519	0.7114	0.8571
400	0.6894	0.7544	0.6825	0.7525
450	0.7374	0.7594	0.6489	0.6733
500	0.7677	0.7644	0.6080	0.6100
550	0.7929	0.8145	0.5709	0.5909
600	0.8308	0.8296	0.5483	0.5517
650	0.8409	0.8296	0.5123	0.5092
700	0.8460	0.8371	0.4786	0.4771
750	0.8535	0.8471	0.4507	0.4507
800	0.8737	0.8672	0.4325	0.4325
850	0.8763	0.8672	0.4082	0.4071
900	0.8763	0.8697	0.3856	0.3856
950	0.8788	0.8697	0.3663	0.3653
1000	0.8864	0.8772	0.3510	0.3500
2000	0.9167	0.9073	0.1815	0.1810
3000	0.9343	0.9298	0.1233	0.1237
4000	0.9394	0.9674	0.0930	0.0965
5000	0.9646	0.9674	0.0764	0.0772
6000	0.9672	0.9699	0.0638	0.0645
7000	0.9672	0.9825	0.0547	0.0560
8000	0.9747	0.9850	0.0483	0.0491
9000	0.9747	0.9850	0.0429	0.0437
10000	0.9823	0.9850	0.0389	0.0393

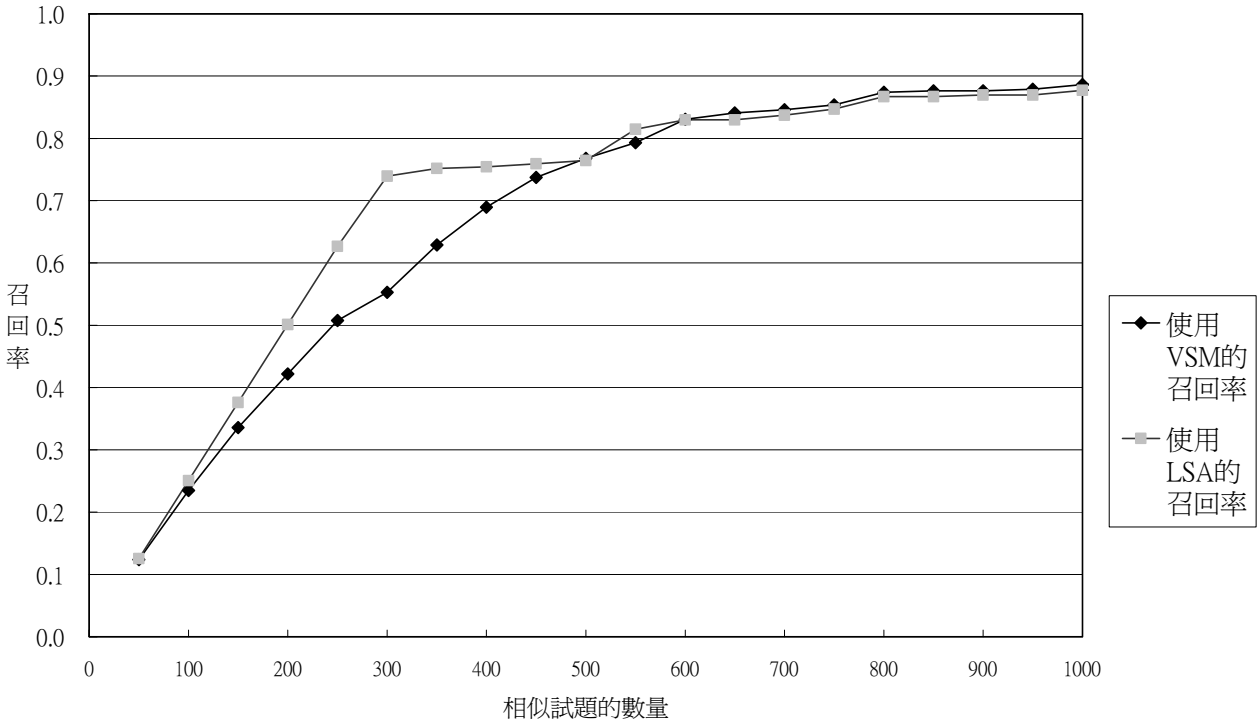

圖 4-1 使用 VSM 或 LSA 在判斷完全相同試題的情況時其召回率之比較

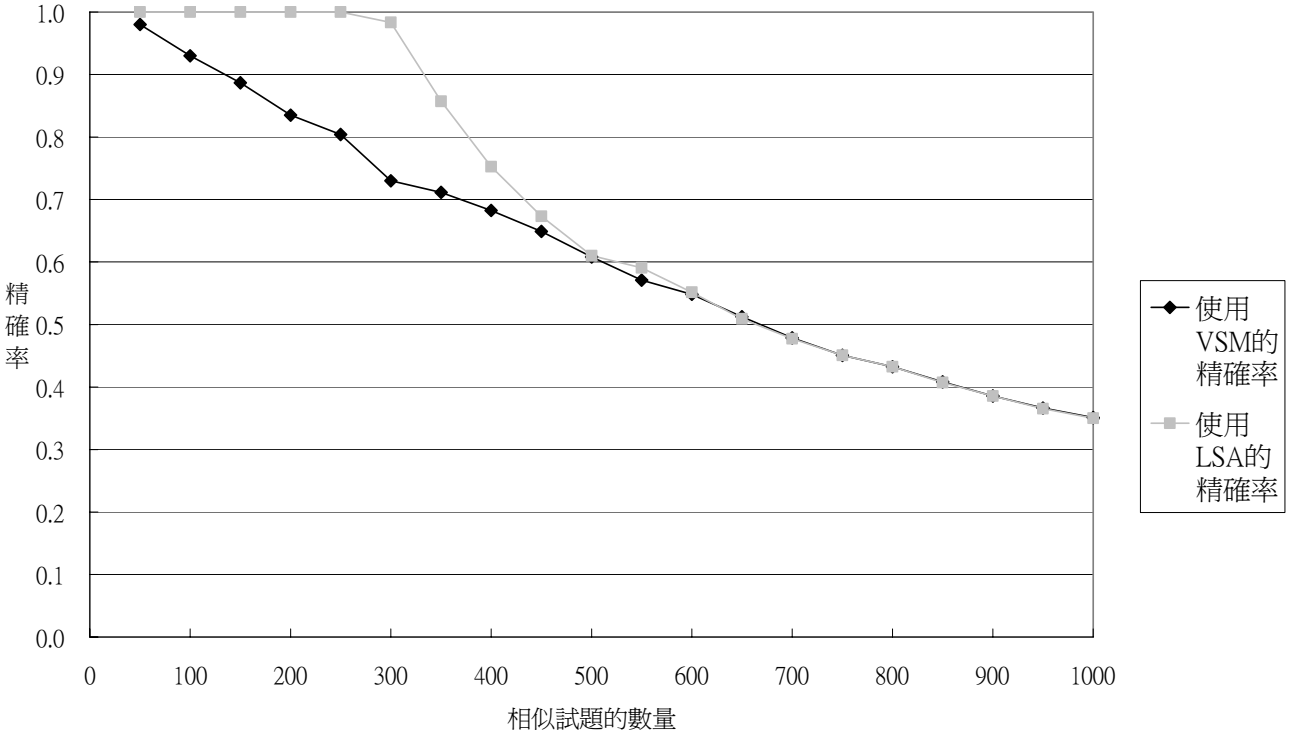

圖 4-2 使用 VSM 或 LSA 在判斷完全相同試題的情況時其精確率之比較

表 4-3 使用 VSM 或 LSA 在判斷非常相似試題的情況時其召回率與精確率

系統判斷出的相似試題數量	使用 VSM 的召回率	使用 LSA 的召回率	使用 VSM 的精確率	使用 LSA 的精確率
50	0.0819	0.0821	0.9800	1.0000
100	0.1639	0.1642	0.9800	1.0000
150	0.2475	0.2463	0.9867	1.0000
200	0.3261	0.3284	0.9750	1.0000
250	0.4080	0.4105	0.9760	1.0000
300	0.4833	0.4877	0.9633	0.9900
350	0.5619	0.5501	0.9600	0.9571
400	0.6271	0.6059	0.9375	0.9225
450	0.6756	0.6634	0.8978	0.8978
500	0.7040	0.6962	0.8420	0.8480
550	0.7324	0.7438	0.7964	0.8236
600	0.7692	0.7635	0.7667	0.7750
650	0.7776	0.7734	0.7154	0.7246
700	0.7826	0.7865	0.6686	0.6843
750	0.7926	0.7980	0.6320	0.6480
800	0.8094	0.8145	0.6050	0.6200
850	0.8144	0.8210	0.5729	0.5882
900	0.8144	0.8309	0.5411	0.5622
950	0.8177	0.8325	0.5147	0.5337
1000	0.8261	0.8374	0.4940	0.5100
2000	0.8662	0.8867	0.2590	0.2700
3000	0.8829	0.9278	0.1760	0.1883
4000	0.8896	0.9524	0.1330	0.1450
5000	0.9130	0.9524	0.1092	0.1160
6000	0.9147	0.9655	0.0912	0.0980
7000	0.9231	0.9803	0.0789	0.0853
8000	0.9298	0.9819	0.0695	0.0748
9000	0.9331	0.9819	0.0620	0.0664
10000	0.9448	0.9819	0.0565	0.0598
15000	0.9532	0.9918	0.0380	0.0403
20000	0.9649	1.0000	0.0289	0.0305

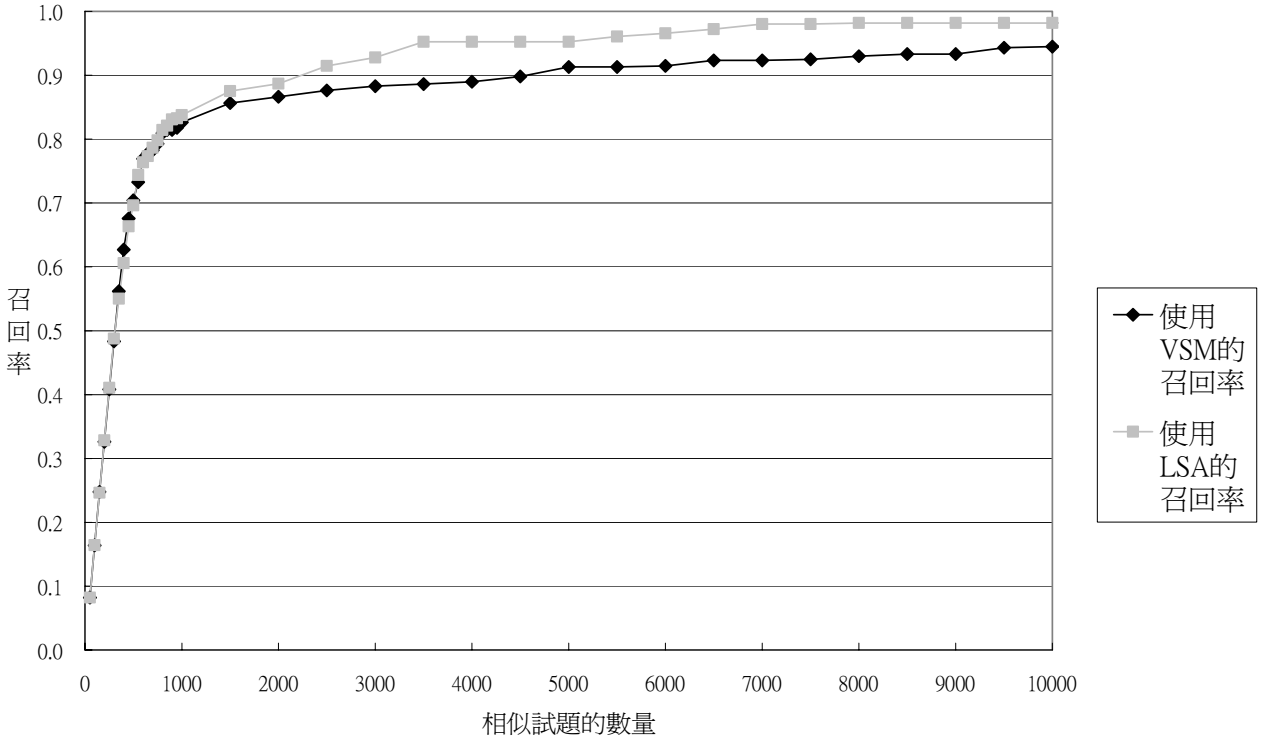

圖 4-3 使用 VSM 或 LSA 在判斷非常相似試題的情況時其召回率之比較

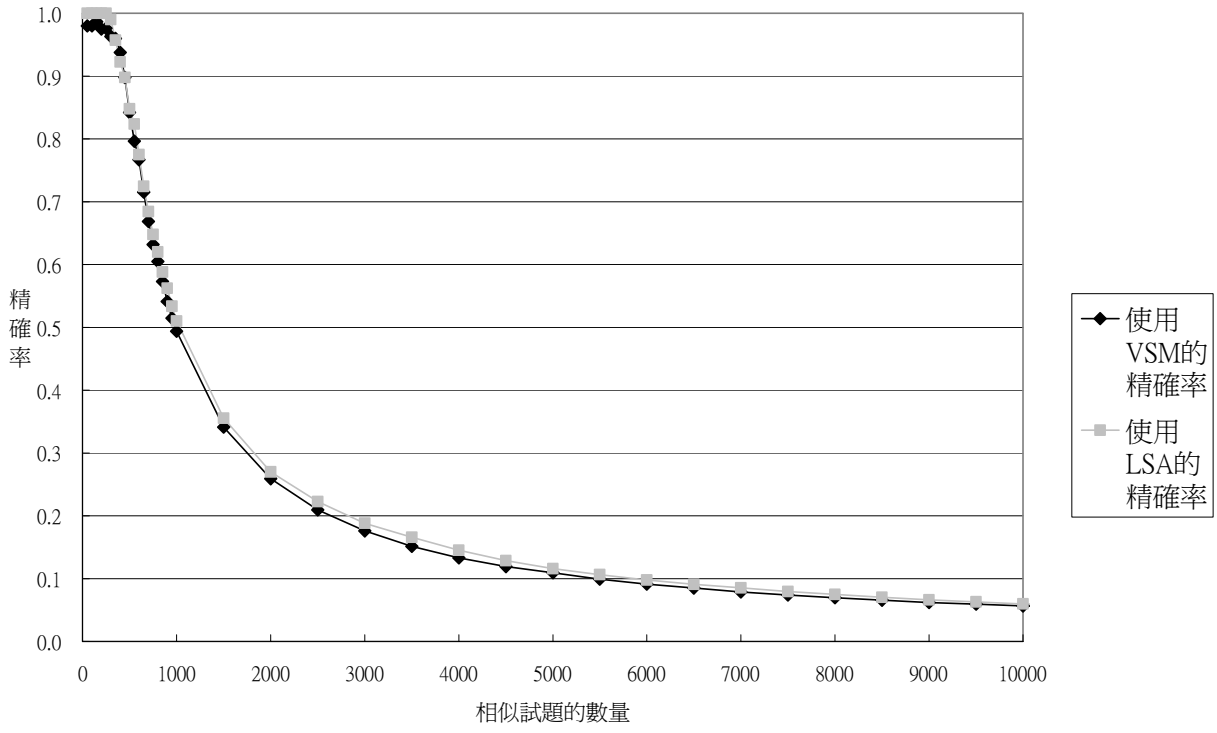

圖 4-4 使用 VSM 或 LSA 在判斷非常相似試題的情況時其精確率之比較

表 4-4 使用 VSM 或 LSA 在判斷部分相似試題的情況時其召回率與精確率

系統判斷出的相似試題數量	使用 VSM 的召回率	使用 LSA 的召回率	使用 VSM 的精確率	使用 LSA 的精確率
50	0.0383	0.0328	0.9800	1.0000
100	0.0767	0.0657	0.9800	1.0000
150	0.1158	0.0985	0.9867	1.0000
200	0.1549	0.1313	0.9900	1.0000
250	0.1941	0.1642	0.9920	1.0000
300	0.2308	0.1950	0.9833	0.9900
350	0.2684	0.2206	0.9800	0.9600
400	0.3067	0.2449	0.9800	0.9325
450	0.3365	0.2725	0.9556	0.9222
500	0.3537	0.2994	0.9040	0.9120
550	0.3693	0.3309	0.8582	0.9164
600	0.3889	0.3559	0.8283	0.9033
650	0.4014	0.3736	0.7892	0.8754
700	0.4069	0.3966	0.7429	0.8629
750	0.4186	0.4163	0.7133	0.8453
800	0.4288	0.4334	0.6850	0.8250
850	0.4351	0.4478	0.6541	0.8024
900	0.4390	0.4629	0.6233	0.7833
950	0.4468	0.4767	0.6011	0.7642
1000	0.4554	0.4865	0.5820	0.7410
2000	0.5313	0.6494	0.3395	0.4945
3000	0.5790	0.7347	0.2467	0.3730
4000	0.6291	0.7938	0.2010	0.3023
5000	0.6698	0.8260	0.1712	0.2516
6000	0.6917	0.8470	0.1473	0.2150
7000	0.7105	0.8739	0.1297	0.1901
8000	0.7285	0.8949	0.1164	0.1704
9000	0.7426	0.9068	0.1054	0.1534
10000	0.7598	0.9219	0.0971	0.1404
15000	0.8083	0.9678	0.0689	0.0983
20000	0.8592	0.9928	0.0549	0.0756

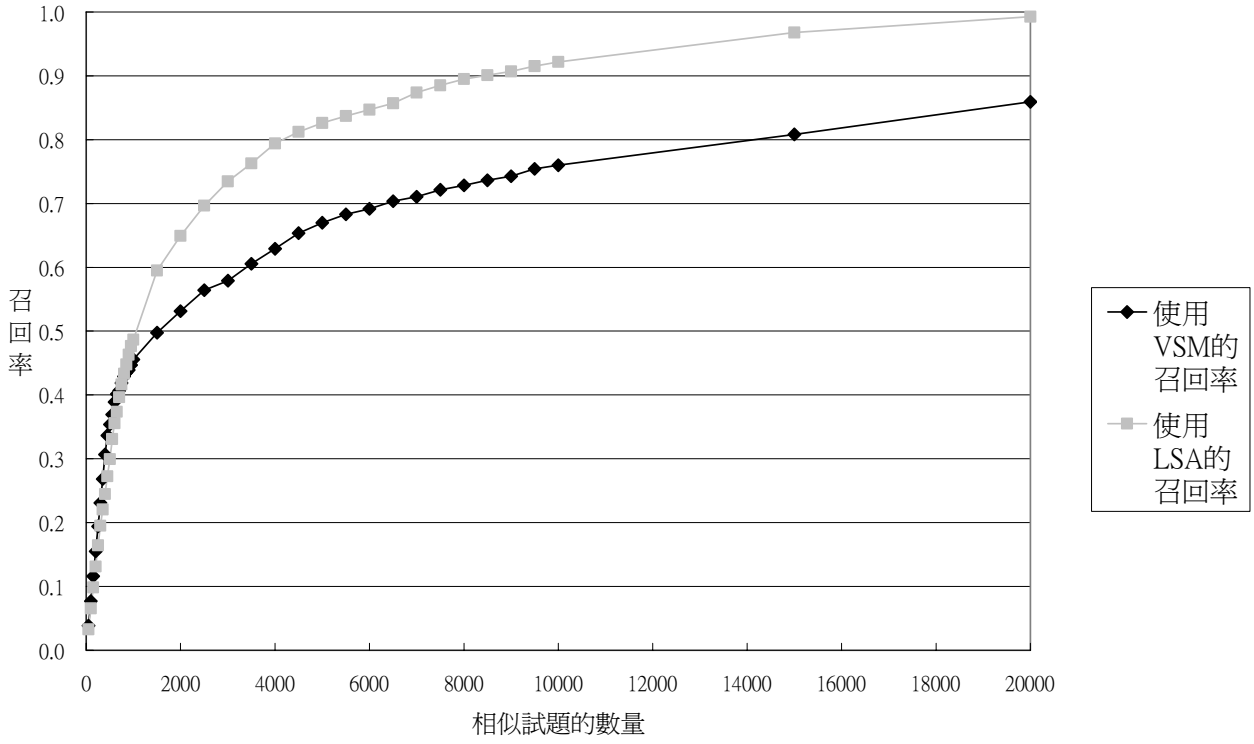

圖 4-5 使用 VSM 或 LSA 在判斷部分相似試題的情況時其召回率之比較

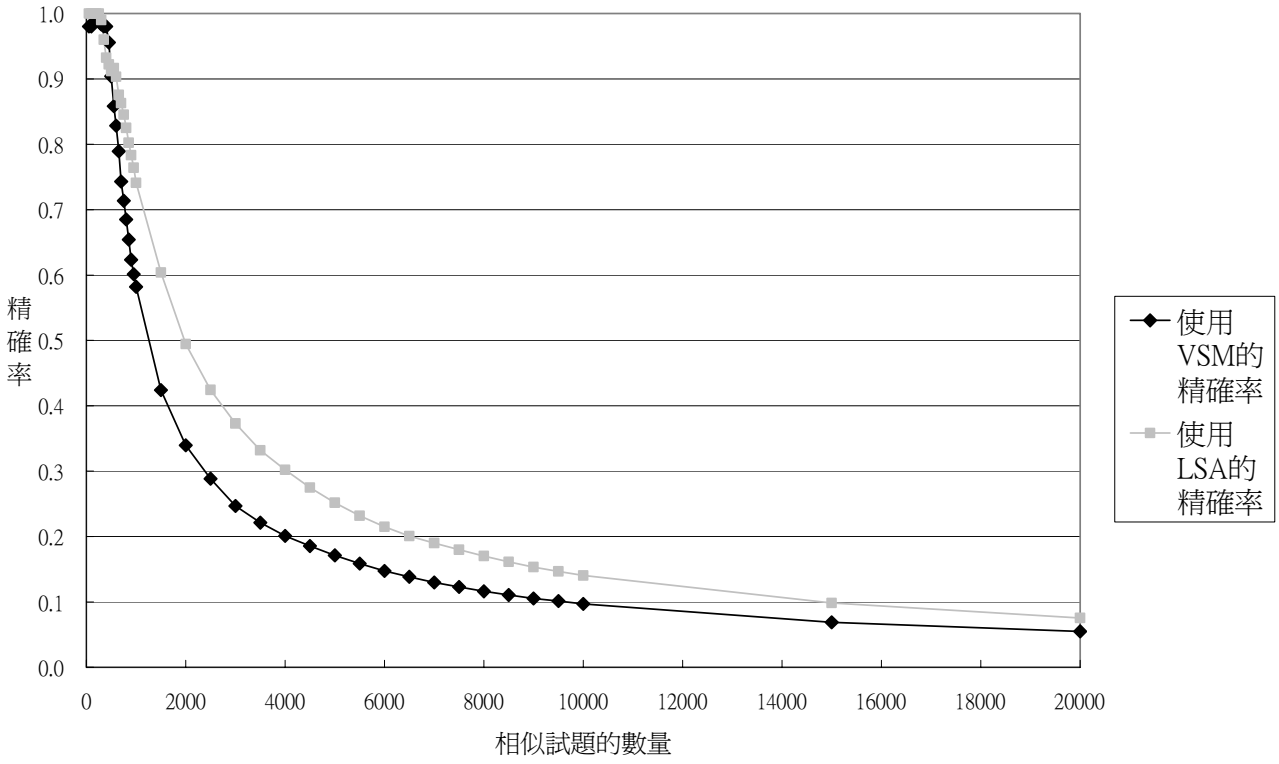

圖 4-6 使用 VSM 或 LSA 在判斷部分相似試題的情況時其精確率之比較

表 4-5 使用 VSM 或 LSA 在判斷些微相似試題的情況時其召回率與精確率

系統判斷出的相似試題數量	使用 VSM 的召回率	使用 LSA 的召回率	使用 VSM 的精確率	使用 LSA 的精確率
50	0.0097	0.0075	0.9800	1.0000
100	0.0194	0.0151	0.9800	1.0000
150	0.0293	0.0226	0.9867	1.0000
200	0.0392	0.0302	0.9900	1.0000
250	0.0491	0.0377	0.9920	1.0000
300	0.0586	0.0452	0.9867	1.0000
350	0.0685	0.0528	0.9886	1.0000
400	0.0784	0.0592	0.9900	0.9825
450	0.0879	0.0668	0.9867	0.9844
500	0.0956	0.0743	0.9660	0.9860
550	0.1033	0.0819	0.9491	0.9873
600	0.1097	0.0888	0.9233	0.9817
650	0.1158	0.0963	0.9000	0.9831
700	0.1207	0.1036	0.8714	0.9814
750	0.1255	0.1107	0.8453	0.9787
800	0.1314	0.1182	0.8300	0.9800
850	0.1340	0.1245	0.7965	0.9718
900	0.1378	0.1315	0.7733	0.9689
950	0.1417	0.1381	0.7537	0.9642
1000	0.1455	0.1441	0.7350	0.9560
2000	0.2114	0.2662	0.5340	0.8830
3000	0.2601	0.3612	0.4380	0.7987
4000	0.3078	0.4438	0.3888	0.7360
5000	0.3466	0.5097	0.3502	0.6762
6000	0.3840	0.5643	0.3233	0.6238
7000	0.4125	0.6196	0.2977	0.5871
8000	0.4365	0.6581	0.2756	0.5456
9000	0.4628	0.6983	0.2598	0.5147
10000	0.4830	0.7348	0.2440	0.4874
15000	0.5855	0.8581	0.1972	0.3795
20000	0.6653	0.9442	0.1681	0.3132

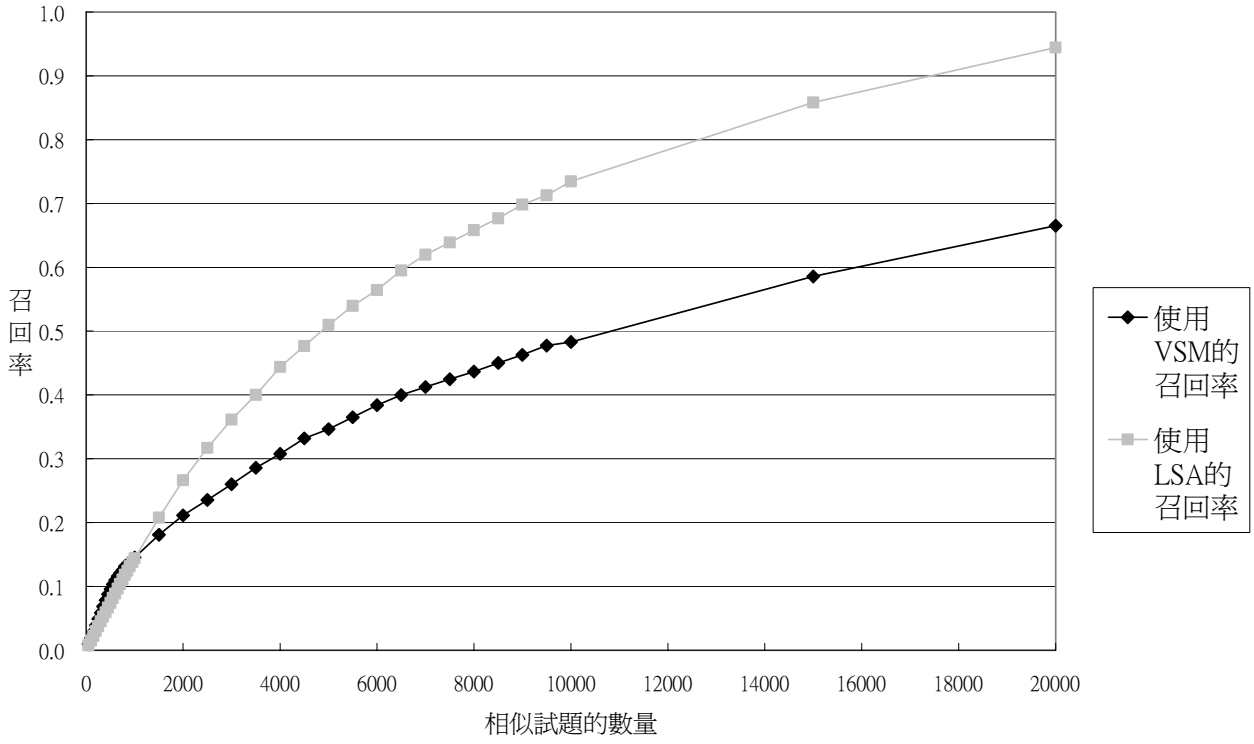

圖 4-7 使用 VSM 或 LSA 在判斷些微相似試題的情況時其召回率之比較

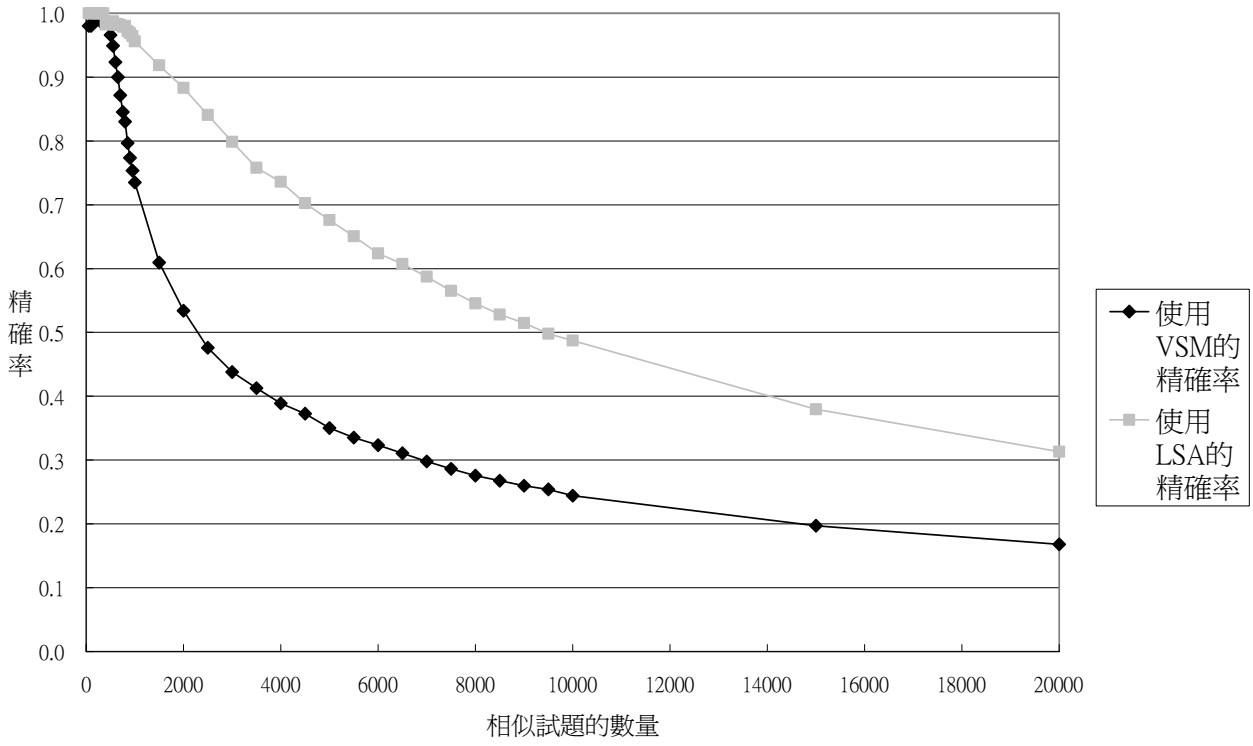

圖 4-8 使用 VSM 或 LSA 在判斷些微相似試題的情況時其精確率之比較

因為在相似度最大的前 1000 個試題組合中，精確率與召回率的變化程度較大，因此我們以 50 個相似試題為單位，在相似度第 1001 名至第 10000 名，以 1000 個相似試題為單位，而在相似度第 10000 名至第 20000 名，以 5000 個相似試題為單位。

觀察表 4-2、表 4-3 與圖 4-1~圖 4-4，可知在判斷兩試題是否完全相同、非常相似時，利用 LSA 的召回率和精確率優於 VSM，但其差異並不明顯；觀察表 4-4、圖 4-5 與圖 4-6，可知在判斷兩試題是否部份相似時，利用 LSA 的召回率和精確率優於 VSM，且相似試題的數量愈多，利用 LSA 判斷之召回率愈佳，但在相似試題數量大於 3000 時，召回率之差異則趨於穩定，約介於 0.15~0.17 之間，在精確率的部份，相似試題的數量愈多，利用 LSA 判斷之精確率愈佳，但相似試題的數量大於 1500 題時，其效果會趨緩；觀察表 4-5、圖 4-7 與圖 4-8，可知在判斷兩試題是否些微相似時，利用 LSA 的召回率和精確率優於 VSM，而相似試題的數量愈多，利用 LSA 判斷之召回率愈佳，在精確率的部份，相似試題的數量愈多，利用 LSA 判斷之精確率愈佳，但相似試題的數量大於 3000 題時，其效果會趨緩。

因此可知，利用 VSM 和 LSA 判斷兩試題是否相同或相似時，依相似程度之不同，其效果有差異，但整體來說利用 LSA 之效果優於 VSM。

第三節　冗詞去除之分析

由於中文目前尚未有較標準的冗詞索引表，因此本研究參考英文冗詞索引表的內容，與中文詞彙在本題庫中出現的次數及意義（如介系詞或語助詞），建立中文冗詞索引表，如表 4-6，另外除了表中所列的詞彙外，標點符號與數字，在本研究中也視為冗詞。

在表 4-6 中，「何者、中、在、下列、的」等詞彙，在題庫中出現的次數為前五名，這些詞彙常出現，但對於試題的意義影響不大，因此本研究將其視為冗詞，另外也有些詞彙出現頻率很高，如電腦（出現 209 次）、檔案（出現 145 次），但其為名詞且對於試題的意義有一定程度的影響，因此本研究不將其列為冗詞。

表 4-6 中文冗詞索引表

詞彙	出現次數	詞彙	出現次數	詞彙	出現次數	詞彙	出現次數	詞彙	出現次數	詞彙	出現次數	詞彙	出現次數	詞彙	出現次數
的	787	如果	74	有	37	我們	22	指	16	多	13	性	10	向	7
下列	500	上	69	爲何	34	有關	20	做	16	和	13	者	10	其中	7
在	392	應	67	此	32	並	20	欲	16	於	13	第	10	所謂	7
中	363	不	66	會	32	後	20	想	16	目前	12	另	9	藉由	7
何者	324	可	66	其	31	該	20	對於	16	爲了	12	至	9	小	6
那	266	何種	65	關於	28	主	19	各種	15	等	12	他	8	共	6
之	211	則	61	來	27	應該	19	何處	15	須	12	任何	8	因	6
一	209	與	59	或	27	已	18	某	15	對	12	先	8	那些	6
是	195	及	53	哪	24	用	18	最	15	從	11	如何	8	係由	6
一個	164	所	53	組	24	設	18	什麼	14	一下	10	把	8	卻	6
爲	139	屬於	52	人	23	部	18	而	14	了	10	是否	8	且	5
時	99	若	44	能	23	使	17	兩	14	入	10	假設	8	如	5
種	91	將	40	當	23	到	17	幾	14	件	10	這	8	我	5
要	87	個	39	內	22	由	16	發生	14	何	10	需	8		
項	86	被	38	多少	22	再	16	請問	14	但	10	需要	8		
以下	81	以	37	利用	22	係	16	讓	14	每	10	必	7		

利用表 4-6 的結果，我們比較有去除冗詞與無去除冗詞，在「完全相同」、「非常相

似」、「部份相似」與「些微相似」四組試題，取系統判斷出的相似度最大的前 N 組，以分析其精確率與召回率的關係，得到的結果如下：

表 4-7 冗詞去除與否在判斷完全相同試題的情況時其召回率與精確率

系統判斷出的相似試題數量	有去除冗詞的召回率	無去除冗詞的召回率	有去除冗詞的精確率	無去除冗詞的精確率
50	0.1253	0.1302	1.0000	1.0000
100	0.2506	0.1979	1.0000	0.7600
150	0.3759	0.2760	1.0000	0.7067
200	0.5013	0.3542	1.0000	0.6800
250	0.6266	0.4375	1.0000	0.6720
300	0.7393	0.4844	0.9833	0.6200
350	0.7519	0.5286	0.8571	0.5800
400	0.7544	0.5651	0.7525	0.5425
450	0.7644	0.5938	0.6778	0.5067
500	0.7744	0.6146	0.6180	0.4720
550	0.7970	0.6224	0.5782	0.4345
600	0.8045	0.6250	0.5350	0.4000
650	0.8120	0.6406	0.4985	0.3785
700	0.8221	0.6510	0.4686	0.3571
750	0.8346	0.6563	0.4440	0.3360
800	0.8346	0.6641	0.4163	0.3188
850	0.8396	0.6667	0.3941	0.3012
900	0.8421	0.6719	0.3733	0.2867
950	0.8446	0.6797	0.3547	0.2747
1000	0.8521	0.6849	0.3400	0.2630
1500	0.8872	0.7266	0.2360	0.1860
2000	0.8897	0.7448	0.1775	0.1430
2500	0.8947	0.7526	0.1428	0.1156
3000	0.9298	0.7760	0.1237	0.0993
3500	0.9323	0.7839	0.1063	0.0860
4000	0.9399	0.7969	0.0938	0.0765
4500	0.9499	0.8073	0.0842	0.0689

5000	0.9549	0.8099	0.0762	0.0622
5500	0.9549	0.8281	0.0693	0.0578
6000	0.9574	0.8307	0.0637	0.0532
6500	0.9574	0.8333	0.0588	0.0492
7000	0.9574	0.8385	0.0546	0.0460
7500	0.9624	0.8411	0.0512	0.0431
8000	0.9724	0.8411	0.0485	0.0404
8500	0.9724	0.8516	0.0456	0.0385
9000	0.9724	0.8594	0.0431	0.0367
9500	0.9724	0.8724	0.0408	0.0353
10000	0.9850	0.8802	0.0393	0.0338
15000	0.9900	0.9089	0.0263	0.0233
20000	1.0000	0.9271	0.0200	0.0178
25000	1.0000	0.9349	0.0160	0.0144
30000	1.0000	0.9557	0.0133	0.0122
35000	1.0000	0.9635	0.0114	0.0106
40000	1.0000	0.9740	0.0100	0.0094

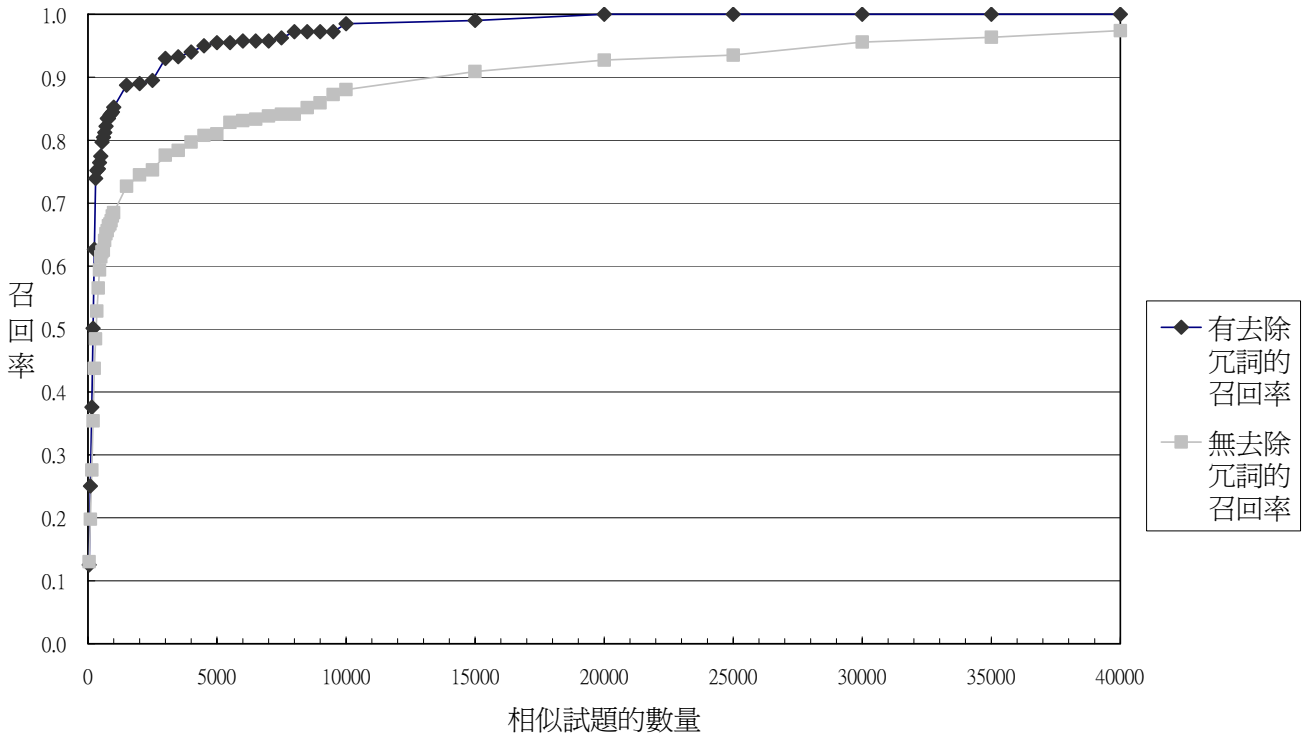

圖 4-9 冗詞去除與否在判斷完全相同試題的情況時其召回率之比較

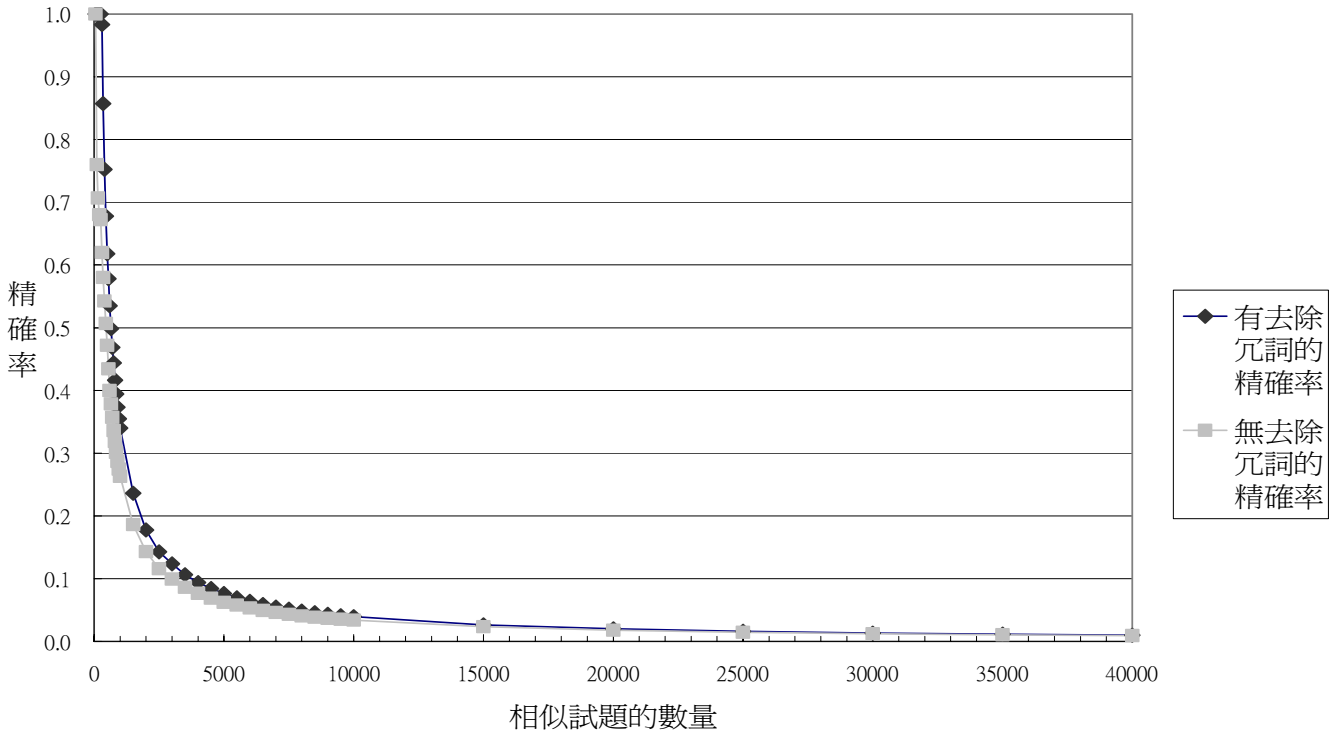

圖 4-10 冗詞去除與否在判斷完全相同試題的情況時其精確率之比較

表 4-8 冗詞去除與否在判斷非常相似試題的情況時其召回率與精確率

系統判斷出的相似試題數量	有去除冗詞的召回率	無去除冗詞的召回率	有去除冗詞的精確率	無去除冗詞的精確率
50	0.0821	0.0000	1.0000	1.0000
100	0.1642	0.0750	1.0000	0.6250
150	0.2463	0.1350	1.0000	0.6136
200	0.3284	0.1750	1.0000	0.5469
250	0.4105	0.2400	1.0000	0.5854
300	0.4877	0.3250	0.9900	0.5702
350	0.5402	0.3750	0.9400	0.5102
400	0.5846	0.4150	0.8900	0.4536
450	0.6568	0.4450	0.8889	0.4009
500	0.7094	0.4600	0.8640	0.3485
550	0.7307	0.4900	0.8091	0.3151
600	0.7389	0.5050	0.7500	0.2806
650	0.7504	0.5200	0.7031	0.2574
700	0.7668	0.5250	0.6671	0.2333
750	0.7849	0.5350	0.6373	0.2149
800	0.7849	0.5450	0.5975	0.2000
850	0.7915	0.5550	0.5671	0.1869
900	0.7997	0.5550	0.5411	0.1729
950	0.8030	0.5550	0.5147	0.1611
1000	0.8145	0.5600	0.4960	0.1520
1500	0.8604	0.6000	0.3493	0.0983
2000	0.8719	0.6150	0.2655	0.0718
2500	0.8818	0.6250	0.2148	0.0565
3000	0.9130	0.6350	0.1853	0.0470
3500	0.9146	0.6600	0.1591	0.0413
4000	0.9376	0.6850	0.1428	0.0371
4500	0.9442	0.6950	0.1278	0.0332
5000	0.9475	0.7050	0.1154	0.0301
5500	0.9475	0.7200	0.1049	0.0278
6000	0.9524	0.7300	0.0967	0.0257
6500	0.9573	0.7350	0.0897	0.0238
7000	0.9573	0.7350	0.0833	0.0220
7500	0.9639	0.7350	0.0783	0.0205

8000	0.9704	0.7350	0.0739	0.0191
8500	0.9704	0.7500	0.0695	0.0184
9000	0.9704	0.7650	0.0657	0.0176
9500	0.9704	0.7700	0.0622	0.0168
10000	0.9787	0.7750	0.0596	0.0160
15000	0.9852	0.8400	0.0400	0.0115
20000	0.9984	0.8850	0.0304	0.0090
25000	0.9984	0.9100	0.0243	0.0074
30000	1.0000	0.9200	0.0203	0.0062
35000	1.0000	0.9400	0.0174	0.0054
40000	1.0000	0.9550	0.0152	0.0048

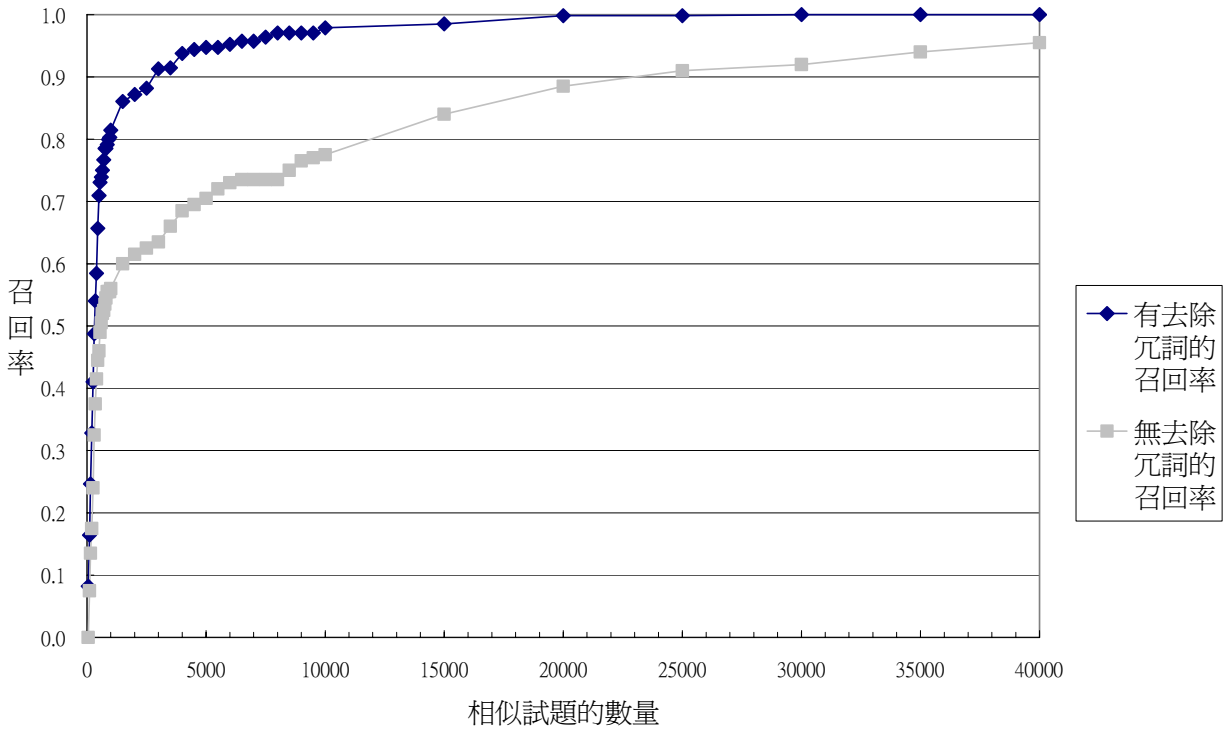

圖 4-11 冗詞去除與否在判斷非常相似試題的情況時其召回率之比較

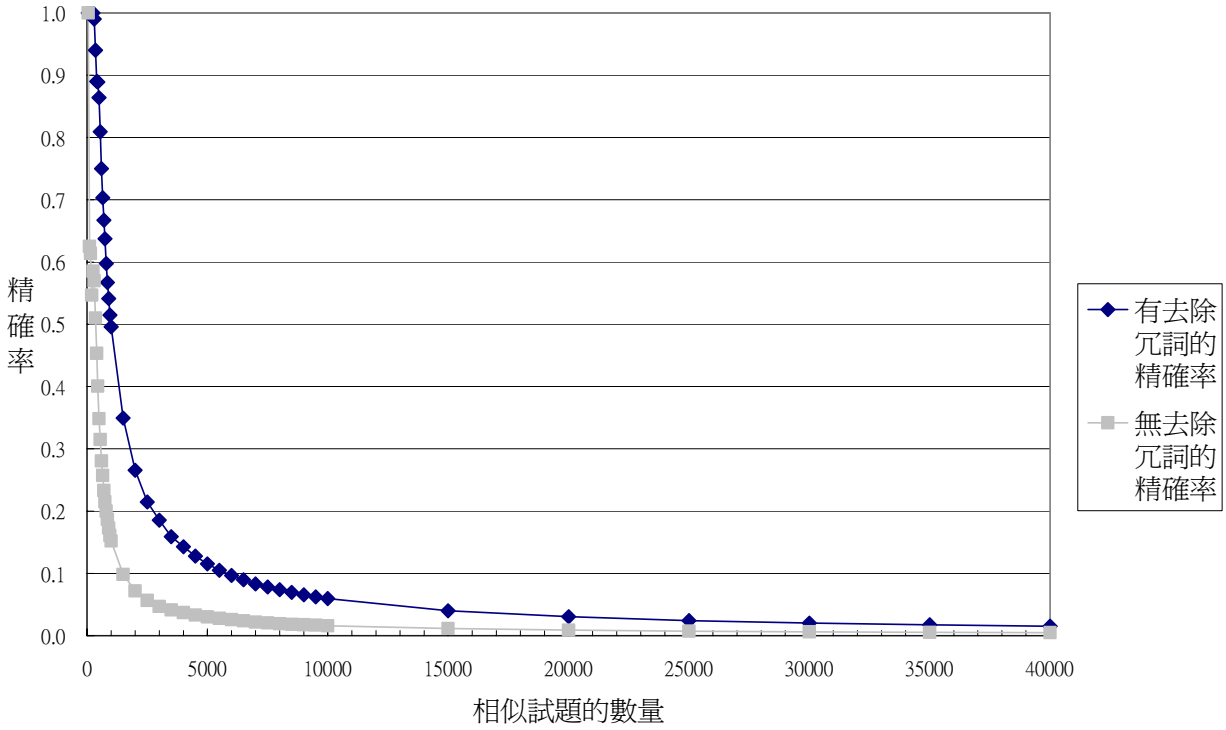

圖 4-12 冗詞去除與否在判斷非常相似試題的情況時其精確率之比較

表 4-9 冗詞去除與否在判斷部分相似試題的情況時其召回率與精確率

系統判斷出的相似試題數量	有去除冗詞的召回率	無去除冗詞的召回率	有去除冗詞的精確率	無去除冗詞的精確率
50	0.0323	0.0000	1.0000	1.0000
100	0.0646	0.0082	1.0000	0.6667
150	0.0969	0.0178	1.0000	0.7647
200	0.1292	0.0274	1.0000	0.6897
250	0.1615	0.0287	1.0000	0.6176
300	0.1919	0.0356	0.9900	0.5306
350	0.2138	0.0424	0.9457	0.4306
400	0.2339	0.0451	0.9050	0.3300
450	0.2649	0.0588	0.9111	0.3233
500	0.2946	0.0657	0.9120	0.2791
550	0.3127	0.0739	0.8800	0.2535
600	0.3353	0.0876	0.8650	0.2471
650	0.3514	0.0944	0.8369	0.2300
700	0.3753	0.1012	0.8300	0.2145
750	0.4018	0.1081	0.8293	0.2020
800	0.4205	0.1122	0.8138	0.1881
850	0.4341	0.1218	0.7906	0.1843
900	0.4477	0.1300	0.7700	0.1789
950	0.4567	0.1354	0.7442	0.1713
1000	0.4683	0.1409	0.7250	0.1648
1500	0.5646	0.2066	0.5827	0.1371
2000	0.6395	0.2394	0.4950	0.1100
2500	0.6796	0.2668	0.4208	0.0935
3000	0.7235	0.2886	0.3733	0.0819
3500	0.7364	0.3146	0.3257	0.0750
4000	0.7578	0.3393	0.2933	0.0697
4500	0.7726	0.3653	0.2658	0.0659
5000	0.7920	0.3871	0.2452	0.0622
5500	0.8068	0.4063	0.2271	0.0590
6000	0.8172	0.4213	0.2108	0.0556
6500	0.8301	0.4364	0.1977	0.0529
7000	0.8430	0.4487	0.1864	0.0502
7500	0.8572	0.4596	0.1769	0.0478

8000	0.8682	0.4692	0.1680	0.0456
8500	0.8721	0.4788	0.1588	0.0436
9000	0.8798	0.4870	0.1513	0.0418
9500	0.8857	0.4979	0.1443	0.0404
10000	0.8947	0.5034	0.1385	0.0387
15000	0.9322	0.5978	0.0962	0.0302
20000	0.9587	0.6840	0.0742	0.0257
25000	0.9774	0.7538	0.0605	0.0225
30000	0.9903	0.8071	0.0511	0.0200
35000	0.9955	0.8495	0.0440	0.0180
40000	0.9987	0.8714	0.0387	0.0162

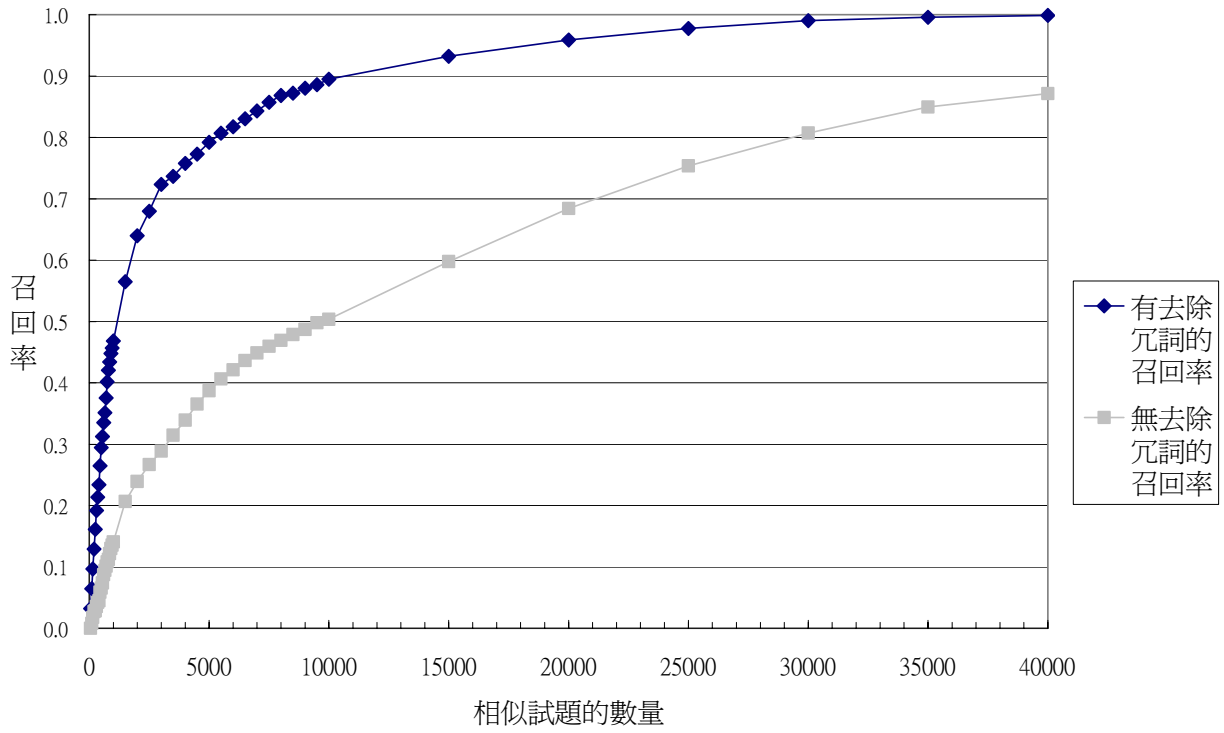

圖 4-13 冗詞去除與否在判斷部分相似試題的情況時其召回率之比較

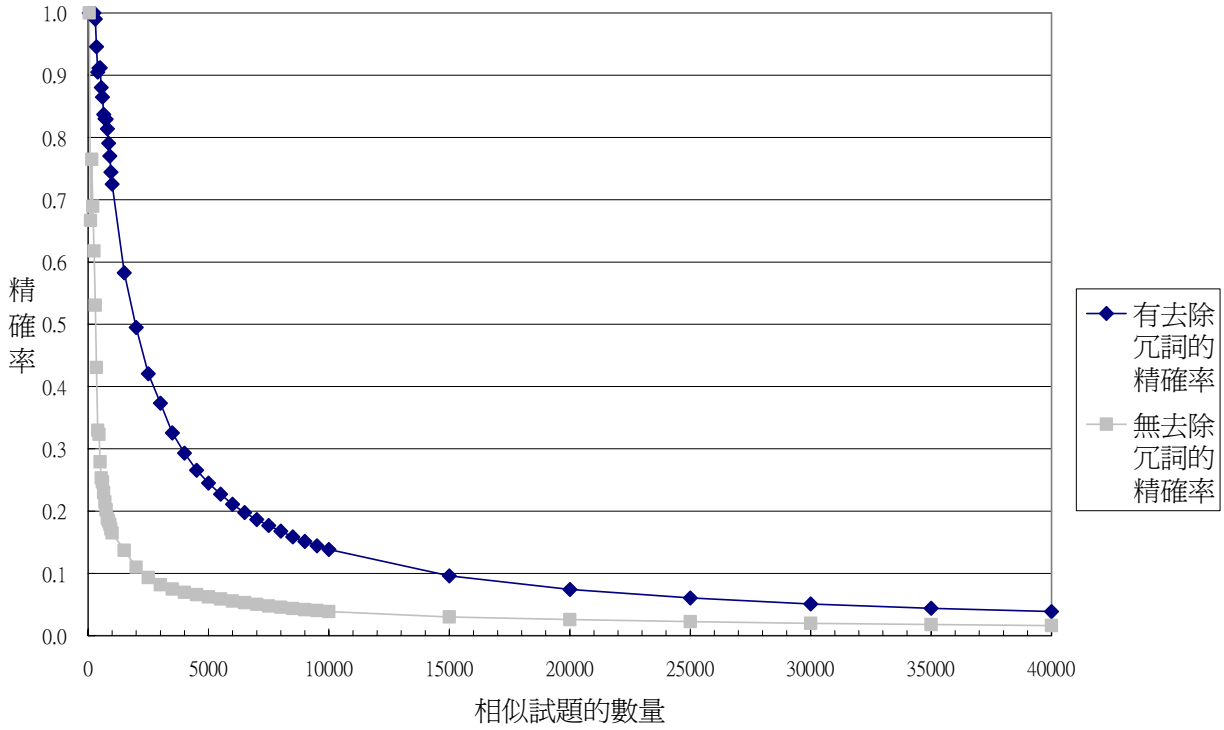

圖 4-14 冗詞去除與否在判斷部分相似試題的情況時其精確率之比較

表 4-10 冗詞去除與否在判斷些微相似試題的情況時其召回率與精確率

系統判斷出的相似試題數量	有去除冗詞的召回率	無去除冗詞的召回率	有去除冗詞的精確率	無去除冗詞的精確率
50	0.0063	0.0000	1.0000	1.0000
100	0.0127	0.0000	1.0000	1.0000
150	0.0190	0.0002	1.0000	0.6667
200	0.0254	0.0015	1.0000	0.6667
250	0.0317	0.0020	1.0000	0.6154
300	0.0381	0.0040	1.0000	0.6957
350	0.0438	0.0065	0.9857	0.6341
400	0.0499	0.0084	0.9825	0.5075
450	0.0562	0.0107	0.9844	0.4778
500	0.0626	0.0127	0.9860	0.4113
550	0.0689	0.0149	0.9873	0.3774
600	0.0752	0.0161	0.9883	0.3333
650	0.0816	0.0189	0.9892	0.3290
700	0.0874	0.0203	0.9843	0.3026
750	0.0935	0.0228	0.9827	0.2949
800	0.0986	0.0256	0.9713	0.2910
850	0.1039	0.0278	0.9635	0.2843
900	0.1103	0.0295	0.9656	0.2729
950	0.1165	0.0313	0.9663	0.2630
1000	0.1221	0.0352	0.9620	0.2720
1500	0.1729	0.0511	0.9087	0.2168
2000	0.2217	0.0732	0.8735	0.2083
2500	0.2589	0.0958	0.8160	0.2041
3000	0.2970	0.1199	0.7803	0.2043
3500	0.3233	0.1380	0.7280	0.1960
4000	0.3531	0.1543	0.6958	0.1880
4500	0.3807	0.1702	0.6667	0.1813
5000	0.4109	0.1873	0.6476	0.1770
5500	0.4352	0.2000	0.6236	0.1700
6000	0.4581	0.2146	0.6017	0.1655
6500	0.4815	0.2270	0.5838	0.1601
7000	0.5040	0.2412	0.5674	0.1567
7500	0.5279	0.2553	0.5547	0.1537

8000	0.5471	0.2658	0.5390	0.1490
8500	0.5634	0.2759	0.5224	0.1449
9000	0.5806	0.2903	0.5084	0.1434
9500	0.5975	0.3020	0.4957	0.1407
10000	0.6098	0.3132	0.4806	0.1381
15000	0.7165	0.4149	0.3765	0.1190
20000	0.7894	0.4983	0.3111	0.1059
25000	0.8461	0.5727	0.2667	0.0965
30000	0.8957	0.6409	0.2353	0.0895
35000	0.9401	0.7017	0.2117	0.0836
40000	0.9727	0.7573	0.1917	0.0787

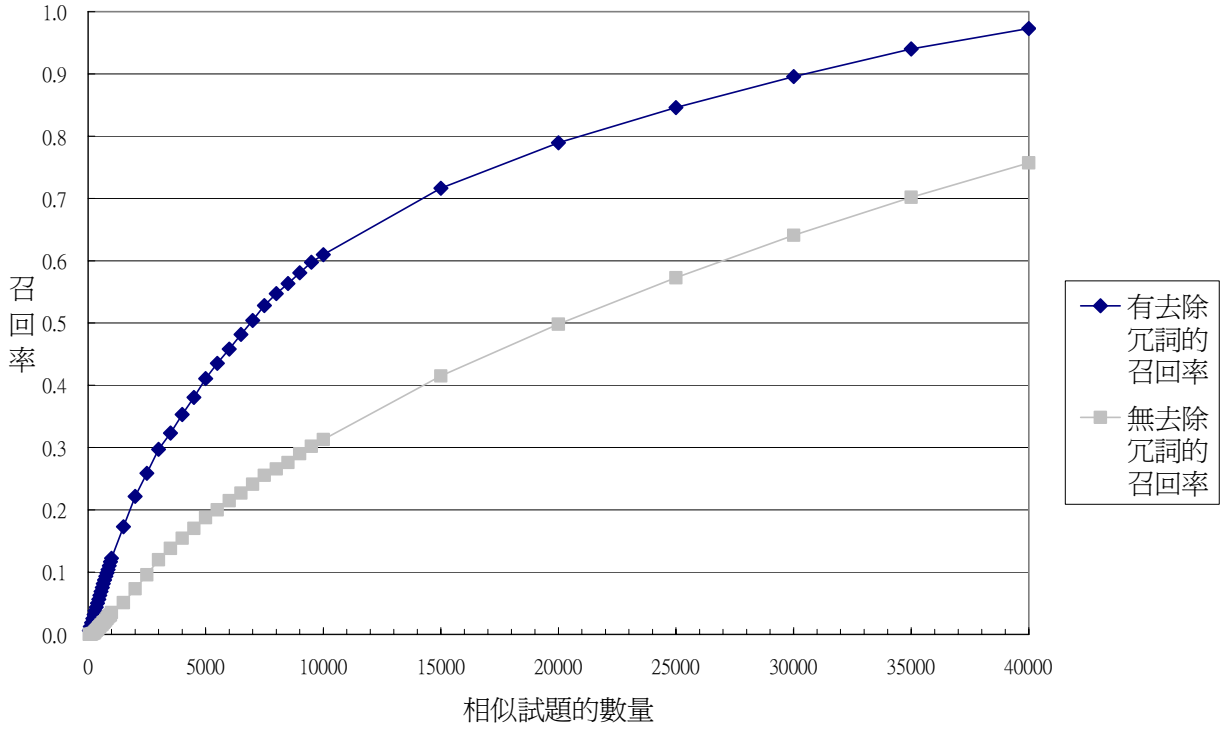

圖 4-15 冗詞去除與否在判斷些微相似試題的情況時其召回率之比較

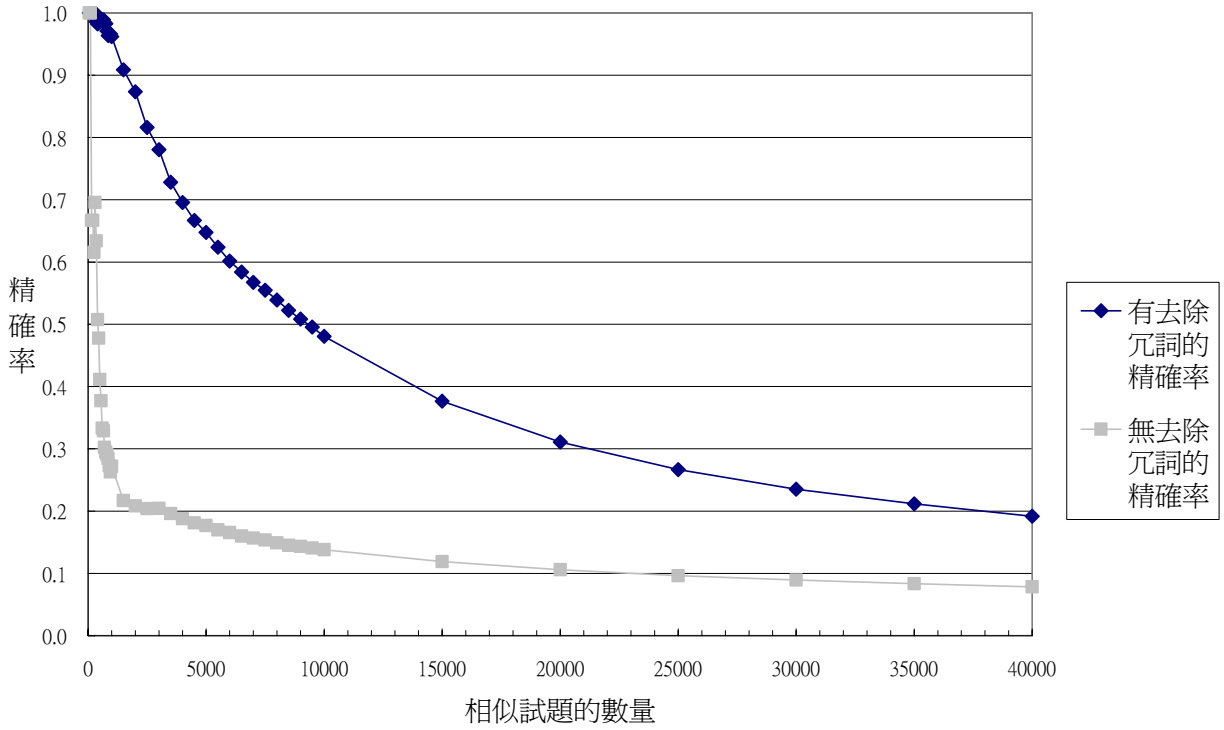

圖 4-16 冗詞去除與否在判斷些微相似試題的情況時其精確率之比較

因為在相似度最大的前 1000 個試題組合中，精確率與召回率的變化程度較大，因此我們以 100 個相似試題為單位，在相似度第 1001 名至第 10000 名，以 500 個相似試題為單位，而在相似度第 10000 名至第 40000 名，以 5000 個相似試題為單位，觀察表 4-7~表 4-10，可知精確率與召回率是成反比的關係，即取的相似試題愈多，精確率會遞減，召回率則遞增。

觀察圖 4-9～圖 4-16，可發現在完全相同、非常相似、部份相似與些微相似四類試題中，有去除冗詞的精確率與召回率，都優於沒有去除冗詞者，但其差異會隨著相似試題數量的增加而遞減。

第四節　不同權重對試題相似度之影響

建立詞彙和試題的關係矩陣時，如果只統計每個詞彙在試題中出現的次數，則會有部份詞彙雖然經常出現，但其重要性並不高，如「敘述、使用、功能」等詞彙，或者雖然出現次數不多，但對試題占有有重有影響者，如「剪貼簿、試算表、列印」等詞彙，因此在做奇異值分解前，須先調整矩陣中每個值的權重。

調整 local 權重的方法有 binary、term frequency(tf)、log 等三種，調整 global 權重的方法有 normal、inverse document frequency(idf)、idf squared(idf2)、entropy 等四種，因此共有 12 種調整權重的組合，如表 4-11，本節將比較各種 local 權重與 global 權重的組合，以分析何種組合的效果最佳。

表 4-11　local 權重和 global 權重的組合

編號	local 權重	global 權重
1	binary	normal
2	binary	idf
3	binary	idf2
4	binary	entropy
5	tf	normal
6	tf	idf
7	tf	idf2
8	tf	entropy
9	log	normal
10	log	idf
11	log	idf2
12	log	entropy

表 4-12　12 種調整權重的方式在不同召回率下之精確率(在判斷完全相同試題的情況時)

調整權重的方式 / 召回率	binary-normal	binary-idf	binary-idf2	binary-entropy	tf-normal	tf-idf	tf-idf2	tf-entropy	log-normal	log-idf	log-idf2	log-entropy
0.10	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
0.15	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
0.20	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
0.25	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
0.30	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
0.35	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
0.40	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
0.45	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
0.50	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
0.55	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
0.60	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
0.65	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
0.70	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
0.75	1.0000	0.9836	0.9740	0.9091	0.8409	0.9868	0.9675	0.9522	0.8655	0.9835	0.9524	0.9036
0.80	1.0000	0.5170	0.5333	0.5872	0.6639	0.5445	0.6151	0.5963	0.6436	0.6092	0.5433	0.5993
0.85	0.7792	0.1810	0.3782	0.3393	0.0752	0.2513	0.4112	0.3919	0.0779	0.2582	0.3403	0.3279
0.90	0.1696	0.0872	0.1254	0.1990	0.0065	0.0842	0.1441	0.1530	0.0065	0.0965	0.1319	0.1947
0.95	0.0463	0.0389	0.0692	0.0708	0.0024	0.0436	0.0721	0.0454	0.0023	0.0362	0.0824	0.0561
1.00	0.0064	0.0124	0.0136	0.0150	0.0014	0.0130	0.0216	0.0112	0.0014	0.0154	0.0259	0.0098

表 4-13　12 種調整權重的方式取前 350 名相似試題之精確率(在判斷完全相同試題的情況時)

調整權重的方式	平均精確率
binary-entropy	0.9352
log-entropy	0.9345
tf-idf2	0.9345
tf-entropy	0.9345
log-idf2	0.9344
binary-idf2	0.9335
tf-idf	0.9280
binary-normal	0.9259
log-idf	0.9249
binary-idf	0.9210
log-normal	0.9127
tf-normal	0.9114

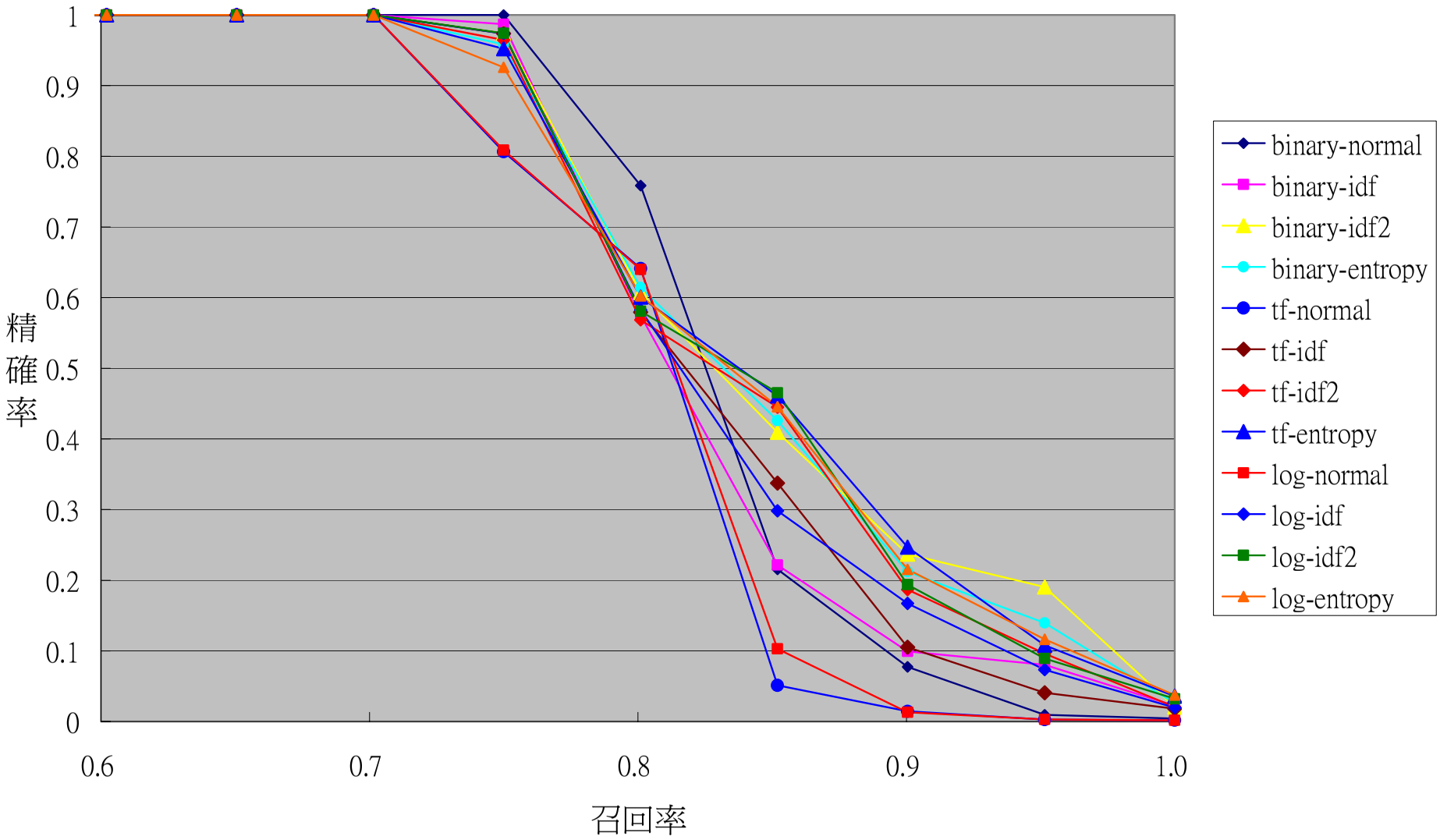

圖 4--17　12 種調整權重的方式在不同召回率下之精確率(在判斷完全相同試題的情況時)

表 4-14　12 種調整權重的方式在不同召回率下之精確率(在判斷非常相似試題的情況時)

調整權重的方式 / 召回率	binary-normal	binary-idf	binary-idf2	binary-entropy	tf-normal	tf-idf	tf-idf2	tf-entropy	log-normal	log-idf	log-idf2	log-entropy
0.10	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
0.15	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
0.20	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
0.25	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
0.30	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
0.35	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
0.40	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
0.45	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
0.50	1.0000	0.9903	0.9903	0.9903	1.0000	0.9901	0.9902	0.9902	1.0000	0.9902	0.9744	0.9903
0.55	1.0000	0.8679	0.9437	0.9911	0.9910	0.8616	0.9356	0.9356	0.9910	0.8579	0.9410	0.9767
0.60	0.9487	0.6214	0.9242	0.9606	0.9918	0.7073	0.9032	0.9239	0.9918	0.6935	0.8884	0.9409
0.65	0.8870	0.3101	0.9083	0.9384	0.9849	0.6321	0.8955	0.9184	0.9849	0.4617	0.8899	0.9252
0.70	0.7510	0.2190	0.8896	0.8914	0.8938	0.5198	0.8781	0.8910	0.9073	0.4247	0.8714	0.8714
0.75	0.4941	0.1856	0.7254	0.8004	0.5568	0.4514	0.7546	0.8096	0.6701	0.3477	0.7031	0.7879
0.80	0.2861	0.1760	0.6100	0.5668	0.0932	0.2701	0.6163	0.6802	0.0992	0.2655	0.5253	0.6032
0.85	0.1383	0.1351	0.3863	0.4390	0.0415	0.1902	0.3529	0.3445	0.0519	0.1686	0.4111	0.3481
0.90	0.0515	0.0797	0.1669	0.2294	0.0068	0.1174	0.2023	0.2028	0.0092	0.0994	0.2008	0.2186
0.95	0.0161	0.0551	0.1002	0.0836	0.0036	0.0657	0.0995	0.0630	0.0035	0.0549	0.0981	0.0847
1.00	0.0091	0.0189	0.0208	0.0229	0.0022	0.0196	0.0167	0.0164	0.0021	0.0221	0.0243	0.0149

表 4-15　12 種調整權重的方式取前 500 名相似試題之精確率(在判斷非常相似試題的情況時)

調整權重的方式	平均精確率
log-entropy	0.9472
tf-entropy	0.9451
binary-entropy	0.9443
log-idf2	0.9376
tf-idf2	0.9361
binary-idf2	0.9339
log-normal	0.9170
tf-normal	0.9137
tf-idf	0.8445
binary-normal	0.8278
log-idf	0.8020
binary-idf	0.7717

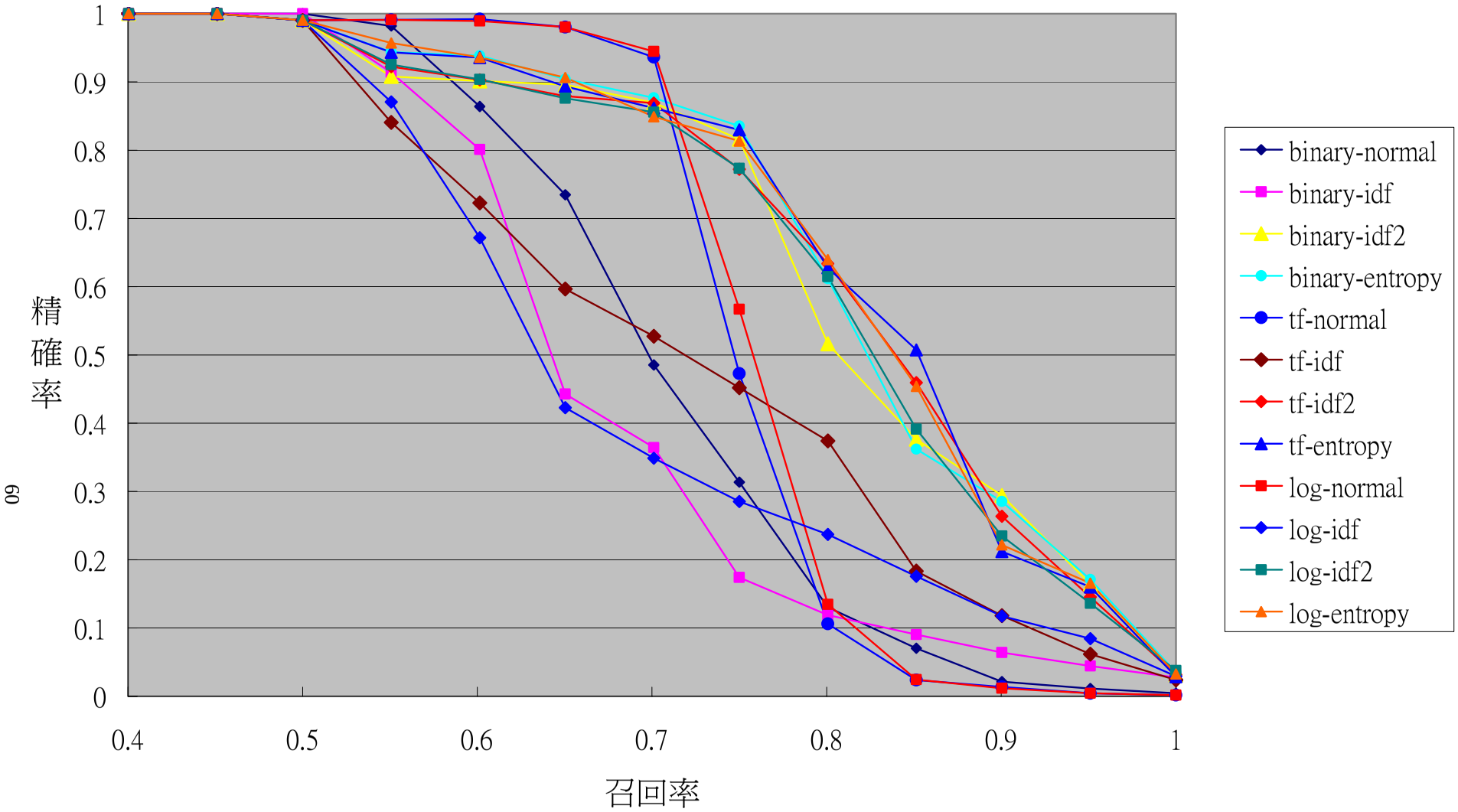

圖 4-18　12 種調整權重的方式在不同召回率下之精確率(在判斷非常相似試題的情況時)

表 4-16　12 種調整權重的方式在不同召回率下之精確率(在判斷部分相似試題的情況時)

調整權重的方式 / 召回率	binary-normal	binary-idf	binary-idf2	binary-entropy	tf-normal	tf-idf	tf-idf2	tf-entropy	log-normal	log-idf	log-idf2	log-entropy
0.10	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
0.15	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
0.20	1.0000	1.0000	0.9903	0.9904	1.0000	1.0000	0.9903	0.9903	1.0000	1.0000	0.9748	0.9904
0.25	1.0000	0.8678	0.9093	0.9529	0.9920	0.9010	0.9207	0.9389	0.9920	0.8919	0.9063	0.9440
0.30	1.0000	0.7364	0.8606	0.9022	0.8782	0.7548	0.9198	0.9200	0.8702	0.7328	0.8891	0.9137
0.35	0.9451	0.4956	0.7873	0.7959	0.4995	0.6764	0.8576	0.9086	0.4742	0.5804	0.8364	0.8803
0.40	0.7829	0.4418	0.6767	0.7135	0.2430	0.5479	0.8304	0.8399	0.2379	0.5345	0.8289	0.8381
0.45	0.4234	0.3751	0.5398	0.6245	0.1326	0.4874	0.7707	0.7977	0.1348	0.4599	0.7609	0.7898
0.50	0.2443	0.3545	0.4700	0.5267	0.0950	0.3937	0.6855	0.7270	0.0938	0.4209	0.6499	0.7638
0.55	0.1408	0.3135	0.3929	0.4617	0.0718	0.3141	0.5745	0.6768	0.0716	0.3622	0.6112	0.7426
0.60	0.0822	0.2812	0.3320	0.4014	0.0507	0.2571	0.4989	0.6093	0.0507	0.2537	0.5689	0.7131
0.65	0.0484	0.2131	0.2677	0.2952	0.0335	0.2059	0.4215	0.5058	0.0348	0.1940	0.4641	0.6171
0.70	0.0390	0.1591	0.2177	0.2245	0.0248	0.1407	0.3783	0.4222	0.0251	0.1492	0.3999	0.4919
0.75	0.0285	0.1240	0.1775	0.1818	0.0204	0.1223	0.3327	0.3494	0.0204	0.1205	0.3169	0.3387
0.80	0.0259	0.1058	0.1530	0.1564	0.0156	0.0831	0.2644	0.2784	0.0160	0.1026	0.2343	0.2458
0.85	0.0238	0.0809	0.1130	0.1202	0.0126	0.0710	0.1953	0.1946	0.0128	0.0804	0.1816	0.1930
0.90	0.0198	0.0696	0.0831	0.0892	0.0099	0.0623	0.1184	0.1258	0.0097	0.0655	0.1280	0.1300
0.95	0.0172	0.0546	0.0608	0.0624	0.0076	0.0537	0.0639	0.0652	0.0075	0.0510	0.0772	0.0835
1.00	0.0155	0.0381	0.0374	0.0335	0.0052	0.0334	0.0334	0.0319	0.0051	0.0374	0.0371	0.0324

表 4-17　12 種調整權重的方式取前 1000 名相似試題之精確率(在判斷部分相似試題的情況時)

調整權重的方式	平均精確率
log-entropy	0.8522
tf-entropy	0.8418
log-idf2	0.8333
tf-idf2	0.8164
binary-entropy	0.8013
binary-idf2	0.7783
tf-idf	0.6583
log-idf	0.6240
binary-idf	0.5855
log-normal	0.5784
tf-normal	0.5756
binary-normal	0.5458

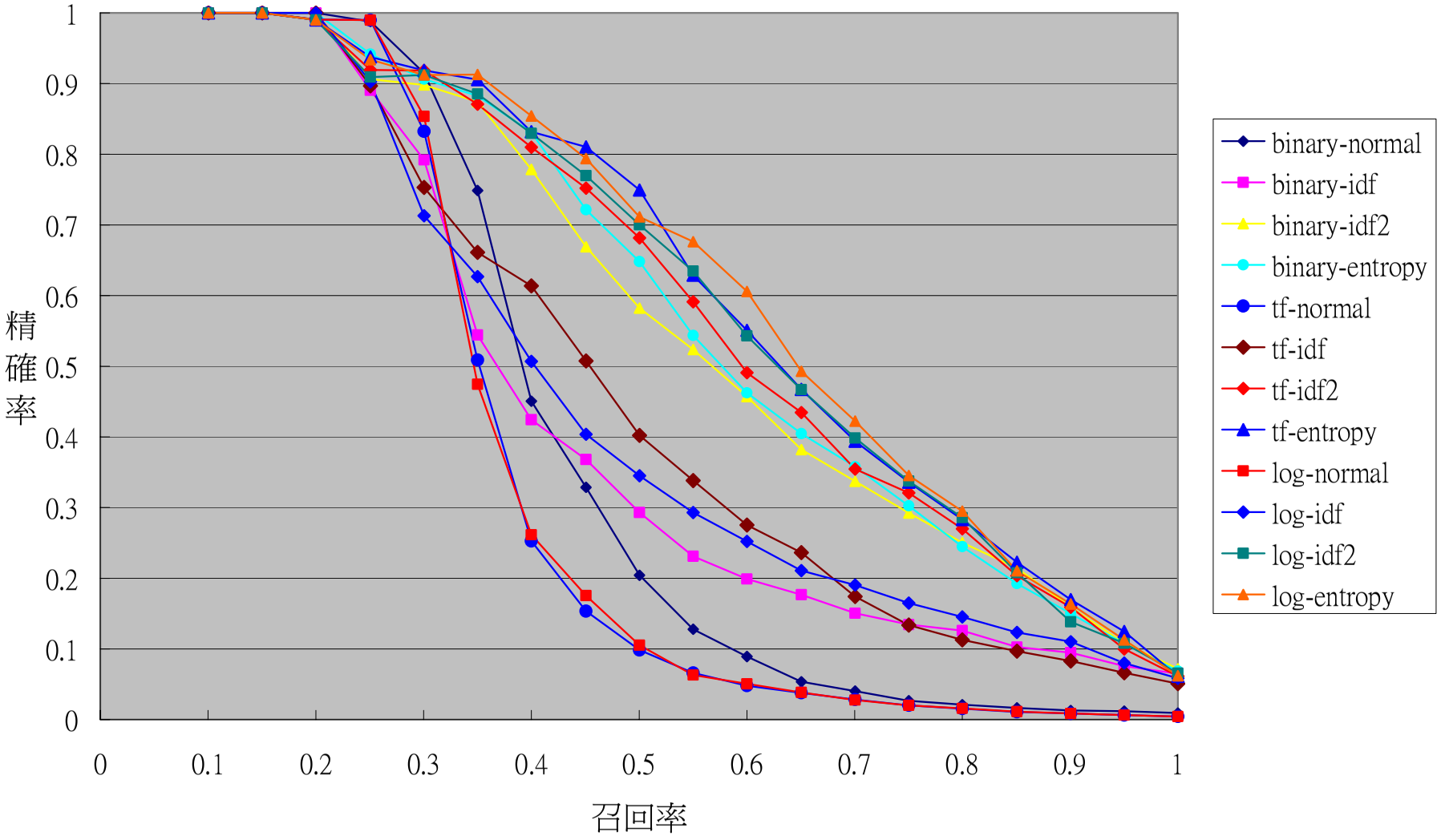

圖 4--19　12 種調整權重的方式在不同召回率下之精確率(在判斷部分相似試題的情況時)

表 4-18　12 種調整權重的方式在不同召回率下之精確率(在判斷些微相似試題的情況時)

調整權重的方式 / 召回率	binary-normal	binary-idf	binary-idf2	binary-entropy	tf-normal	tf-idf	tf-idf2	tf-entropy	log-normal	log-idf	log-idf2	log-entropy
0.10	1.0000	0.9022	0.9348	0.9260	0.4857	0.9322	0.9679	0.9311	0.4496	0.9377	0.9717	0.9179
0.15	0.6181	0.8052	0.8624	0.8753	0.2402	0.8952	0.9248	0.8351	0.2310	0.8860	0.9374	0.8523
0.20	0.3488	0.7521	0.8279	0.8090	0.1769	0.8321	0.8538	0.7297	0.1625	0.8198	0.8875	0.8169
0.25	0.1959	0.6864	0.7677	0.7463	0.1406	0.7651	0.7690	0.7009	0.1332	0.7663	0.8247	0.7479
0.30	0.1500	0.6215	0.7322	0.6961	0.1150	0.6676	0.7048	0.6468	0.1101	0.7164	0.7706	0.7030
0.35	0.1121	0.5408	0.6738	0.6598	0.0932	0.5305	0.6504	0.5905	0.0918	0.5809	0.6988	0.6422
0.40	0.0930	0.4510	0.6178	0.6013	0.0787	0.4625	0.5993	0.5479	0.0818	0.4324	0.6531	0.5912
0.45	0.0836	0.3529	0.5705	0.5598	0.0694	0.3706	0.5556	0.4990	0.0703	0.3522	0.6090	0.5522
0.50	0.0709	0.3011	0.5355	0.5260	0.0611	0.3196	0.5056	0.4574	0.0648	0.3085	0.5710	0.5166
0.55	0.0629	0.2825	0.5038	0.4874	0.0550	0.2663	0.4540	0.4190	0.0591	0.2801	0.5363	0.4556
0.60	0.0618	0.2583	0.4629	0.4472	0.0507	0.2379	0.4100	0.3860	0.0540	0.2578	0.4935	0.3869
0.65	0.0602	0.2383	0.4265	0.4122	0.0474	0.2244	0.3732	0.3429	0.0487	0.2460	0.4403	0.3416
0.70	0.0591	0.2298	0.3814	0.3720	0.0445	0.2146	0.3340	0.3060	0.0448	0.2380	0.3902	0.3026
0.75	0.0566	0.2216	0.3407	0.3329	0.0396	0.2086	0.2960	0.2711	0.0411	0.2300	0.3448	0.2739
0.80	0.0523	0.2118	0.3006	0.2988	0.0366	0.2008	0.2661	0.2403	0.0375	0.2208	0.3014	0.2445
0.85	0.0502	0.2006	0.2643	0.2606	0.0336	0.1926	0.2430	0.2190	0.0342	0.2081	0.2650	0.2202
0.90	0.0485	0.1926	0.2320	0.2311	0.0302	0.1731	0.2204	0.1968	0.0311	0.1967	0.2338	0.1982
0.95	0.0475	0.1777	0.2039	0.1963	0.0275	0.1619	0.1943	0.1783	0.0275	0.1772	0.2072	0.1806
1.00	0.0461	0.1626	0.1762	0.1645	0.0247	0.1434	0.1638	0.1543	0.0244	0.1553	0.1763	0.1681

表 4-19　12 種調整權重的方式取前 5000 名相似試題之精確率(在判斷些微相似試題的情況時)

調整權重的方式	平均精確率
log-entropy	0.7734
log-idf2	0.7596
binary-entropy	0.7444
tf-idf2	0.7318
tf-entropy	0.7177
binary-idf2	0.6978
tf-idf	0.5516
log-idf	0.5314
binary-idf	0.4802
log-normal	0.2702
tf-normal	0.2665
binary-normal	0.2029

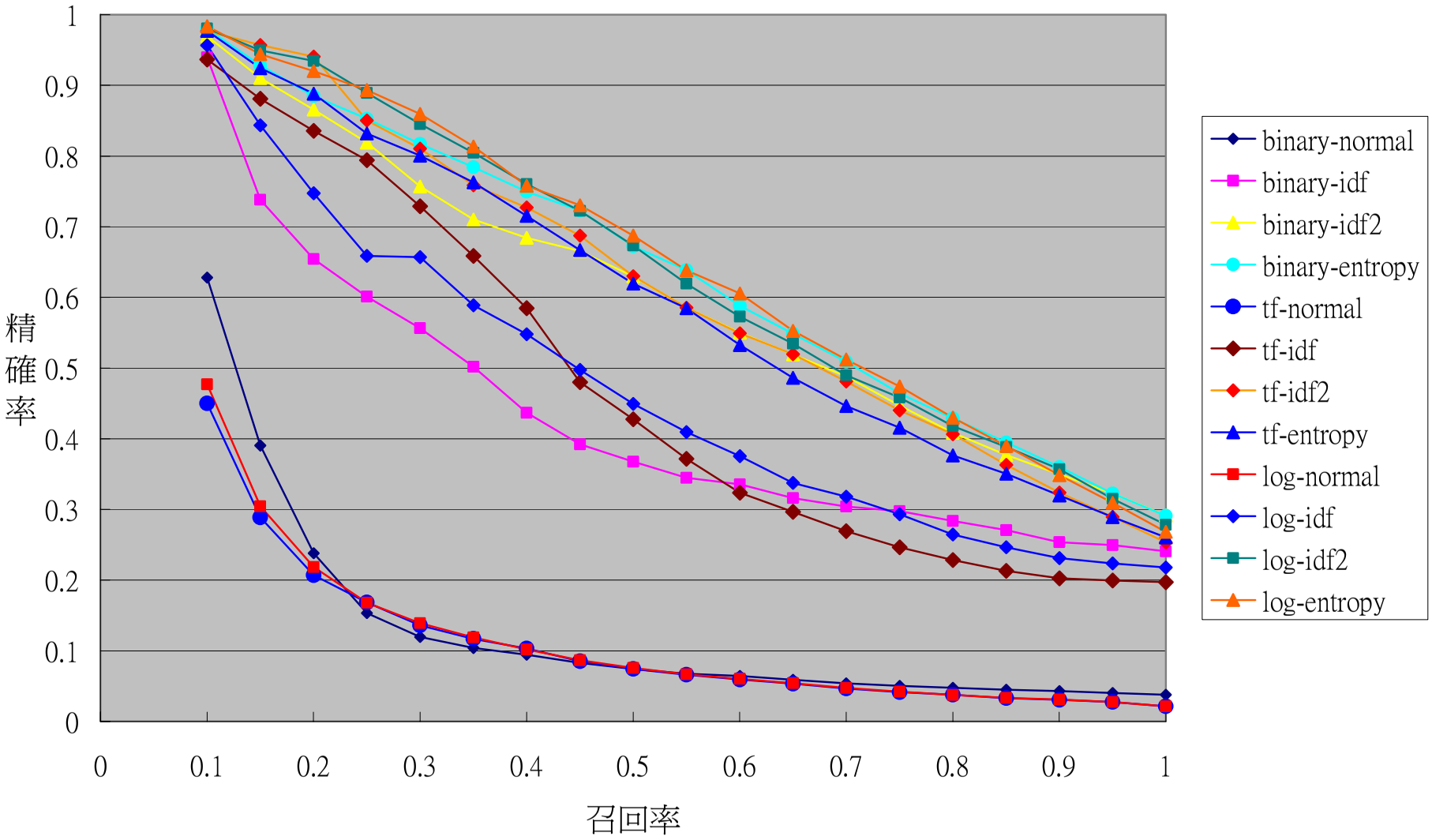

圖 4-20　12 種調整權重的方式在不同召回率下之精確率(在判斷些微相似試題的情況時)

觀察表 4-12、表 4-13 與圖 4-17，可知利用潛在語意分析判斷兩試題是否完全相同，各種調整權重的方式其效果之差異不大，而其中以使用 binary-entropy 調整權重時的效果略佳。

觀察表 4-14、表 4-15 與圖 4-18，可知利用潛在語意分析判斷兩試題是否非常相似，使用 binary-idf2、tf-idf2、log-idf2、inary-entropy、tf-entropy 與 log-entropy 的效果較好，其中以 log-entropy 效果最佳，因此可知在判斷兩試題是否非常相似時，global 的權重為 idf2 與 entropy 者較佳，而 local 的權重則較無影響。

觀察表 4-16、表 4-17 與圖 4-19，可知利用潛在語意分析判斷兩試題是否部分相似時，使用 tf-idf2、log-idf2、tf-entropy 與 log-entropy 調整權重時的效果較好，其中以 log-entropy 效果最佳，因此可知在判斷兩試題是否部分相似時，local 權重為 tf 與 log，global 權重為 idf2 與 entropy 者，效果較佳。

觀察表 4-18、表 4-19 與圖 4-20，使用 binary-idf2、tf-idf2、log-idf2、inary-entropy、tf-entropy 與 log-entropy 的效果較好，其中以 log-entropy 效果最佳，因此可知在判斷兩試題是否些微相似時，global 的權重為 idf2 與 entropy 者較佳，而 local 的權重則較無影響。

整體來看，判斷兩試題是否完全相同時，使用 binary-entropy 調整權重效果較佳，判斷兩試題是否非常相似、部分相似與些微相似時，使用 log-entropy 調整權重的效果較佳。

第五節　不同約化維度對試題相似度之影響

維度約化即為將詞彙和試題的關係矩陣 X，利用奇異值分解得到的三個矩陣 T_r、S_r、D_r 後，保留其 k 個維度相乘以得到新的矩陣 X’，如此可消除語意空間中的雜訊。但由於要保留多少個維度 k，並沒有理論上的最佳值，需經由實驗的方式找出不同文件集最佳的 k 值，本節將比較不同的 k 值以找出成效最佳者，而依據第三節與第四節的結果，在建立詞彙和試題的關係矩陣 X 時，研究者事先去除冗詞，判斷完全相同的試題使用 binary-entropy 調整權重，判斷非常相似、部分相似與些微相似使用 log-entropy 調整權重，所得到的結果如表 4-20、表 4-21、表 4-22、表 4-23 與圖 4-21。

觀察表 4-20，可知在判斷兩試題是否完全相同時，保留的維度愈高，精確率愈佳，但保留的維度大於 5 後，精確率的增加會趨緩，觀察表 4-21，可知在判斷兩試題是否非常相似時，保留維度為 30 時，精確率最佳，觀察表 4-22，可知在判斷兩試題是否部分相似時，保留維度為 15 時，精確率最佳，觀察表 4-23，可知在判斷兩試題是否些微相似時，保留維度為 14 時精確率最佳；而分析圖 4-21，可知保留的維度到達某個臨界值後，精確率的增加會趨緩甚至降低。

表 4-20　約化維度之差異在判斷完全相同試題情況時之精確率

保留的維度	精確率
110	0.9577
118	0.9577
100	0.9570
80	0.9569
90	0.9563
70	0.9556
60	0.9503
50	0.9469
40	0.9412
30	0.9374
20	0.9355
15	0.9344
19	0.9343
16	0.9340
18	0.9339
17	0.9332
14	0.9324
13	0.9318
12	0.9296
11	0.9288
10	0.9267
9	0.9253
8	0.9231
7	0.9196
6	0.9175
5	0.9162
4	0.9014
3	0.8795
2	0.8569

表 4-21　約化維度之差異在判斷非常相似試題情況時之精確率

保留的維度	精確率
30	0.9413
25	0.9396
40	0.9385
15	0.9376
19	0.9370
18	0.9367
20	0.9366
13	0.9360
14	0.9355
16	0.9354
17	0.9345
11	0.9342
12	0.9337
8	0.9320
6	0.9315
50	0.9311
9	0.9303
10	0.9302
7	0.9297
115	0.9259
60	0.9258
118	0.9258
110	0.9251
100	0.9227
5	0.9220
90	0.9188
70	0.9137
80	0.9123
4	0.8979
3	0.8300
2	0.6422

表 4-22　約化維度之差異在判斷部分相似試題情況時之精確率

保留的維度	精確率
15	0.8333
14	0.8330
10	0.8308
17	0.8306
9	0.8275
13	0.8271
16	0.8265
18	0.8263
19	0.8256
11	0.8243
12	0.8242
20	0.8228
8	0.8149
7	0.8101
6	0.8043
25	0.7940
30	0.7811
40	0.7695
50	0.7568
60	0.7483
5	0.7460
115	0.7393
100	0.7373
110	0.7373
90	0.7372
118	0.7367
70	0.7347
80	0.7298
4	0.6592
3	0.4990
2	0.3546

表 4-23　約化維度之差異在判斷些微相似試題情況時之精確率

保留的維度	精確率
14	0.7610
15	0.7596
13	0.7593
12	0.7536
10	0.7518
11	0.7516
16	0.7496
19	0.7373
9	0.7365
18	0.7362
17	0.7330
20	0.7299
8	0.6934
25	0.6788
7	0.6707
30	0.6575
40	0.6322
50	0.6290
70	0.6224
60	0.6207
80	0.6180
90	0.6154
100	0.6075
6	0.6074
110	0.5977
115	0.5954
118	0.5926
5	0.4545
4	0.3509
3	0.2257
2	0.1456

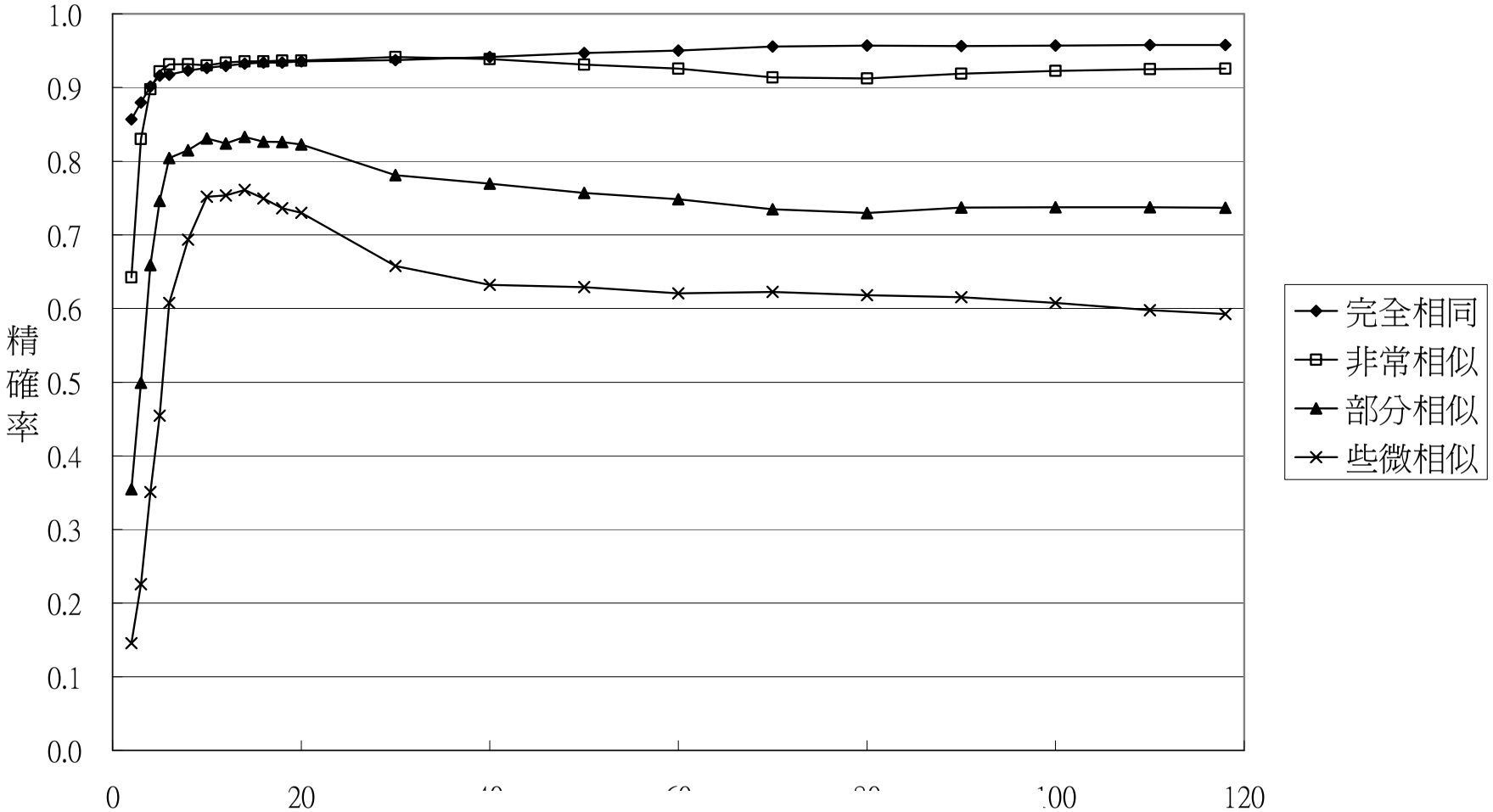

圖 4-21　約化維度之差異在不同試題相似程度下之精確率

第六節　試題相似度之分析

分析行政院勞工委員會所編製的「電腦軟體應用技能檢定丙級學科」92 年度與 93 年度試題共 1000 題選擇題，其試題組合共有 C(1000,2)= 499500 組，經研究者判斷後，其中有 399 組相同的試題組合，7880 組相似的試題組合，分析利用潛在語意分析是否能判斷出相同與相似的試題，得到的結果如下：

（一）相同的試題

92 年度和 93 年度「電腦軟體應用技能檢定丙級學科」共有 399 組完全相同的試題組合，利用潛在語意分析判斷的相似度皆大於 0.9，其中有 295 組系統判斷之相似度為 1，由表 4-24 可知，取相似度最大的前 500 組試題，召回率以可達 0.7744，表示 399 組完全的相同試題組已可找出 309 組。

表 4-24 92 年度與 93 年度相似試題的數量精確率、召回率的關係(在判斷相同試題的情況時)

系統判斷出的相似試題數量	召回率	精確率
50	0.1253	1.0000
100	0.2506	1.0000
150	0.3759	1.0000
200	0.5013	1.0000
250	0.6266	1.0000
300	0.7393	0.9833
350	0.7519	0.8571
400	0.7544	0.7525
450	0.7644	0.6778
500	0.7744	0.6180

550	0.7970	0.5782
600	0.8045	0.5350
650	0.8120	0.4985
700	0.8221	0.4686
750	0.8346	0.4440
800	0.8346	0.4163
850	0.8396	0.3941
900	0.8421	0.3733
950	0.8446	0.3547
1000	0.8521	0.3400
1500	0.8872	0.2360
2000	0.8897	0.1775
2500	0.8947	0.1428
3000	0.9298	0.1237
3500	0.9323	0.1063
4000	0.9399	0.0938
4500	0.9499	0.0842
5000	0.9549	0.0762
5500	0.9549	0.0693
6000	0.9574	0.0637
6500	0.9574	0.0588
7000	0.9574	0.0546
7500	0.9624	0.0512
8000	0.9724	0.0485
8500	0.9724	0.0456
9000	0.9724	0.0431
9500	0.9724	0.0408
10000	0.9850	0.0393
15000	0.9900	0.0263
20000	1.0000	0.0200
25000	1.0000	0.0160
30000	1.0000	0.0133
35000	1.0000	0.0114
40000	1.0000	0.0100

（二）部份辭彙不同的相似試題

而在相似的試題組合方面，經過改版後的試題，若只將 Windows98 修改為 WindowsXP 者，如

92 年度第 732 題：

「使用 Windows98 的網路上的芳鄰，無法分享下列那一項？」

93 年度第 731 題：

「使用 WindowsXP 的「網路上的芳鄰」，無法分享下列那一項」，

或是

92 年度第 733 題：

「以下那一個不是 Windows98 提供的通訊協定？」

93 年度第 732 題：

「以下那一個不是 WindowsXP 提供的通訊協定」

相似度皆在 0.9999 以上，表示經過改版後的試題，若只是修改軟體的名稱，系統可將這類的試題組合找出。

（三）敘述方式不同的相同試題

題意相同，但敘述方式不同的試題，如

92 年度第 591 題：

「電腦執行數值運算的速度受到下列何者影響？」

93 年度第 590 題：

「下列何者會影響電腦執行數值運算的速度？」

或

92 年度第 634 題：

「一般編寫程式的流程爲」

93 年度第 633 題：

「編寫程式的一般流程爲何？」

相似度皆為 1，表示系統可判別題意相同，但敘述方式不同的試題。

（四）辭彙不同，但意義相同的相同試題，如

93 年度第 895 題：

「關於「電腦病毒」的敘述中，下列何者有誤？　」

93 年度第 898 題：

「關於「電腦病毒」的敘述中，下列何者不正確？」

其相似度為 0.9959

92 年度第 876 題：

「預防電腦病毒，下列敘述何者有誤？」

92 年度第 940 題：

「避免電腦中毒的方法，下列何者不正確？」

其相似度為 0.9817

92 年度第 576 題：

「二進制 1011，1001，1100，0011 以十六進制表示爲」

92 年度第 615 題：

「二進制數值 1101001 轉換爲十六進制時，其值爲」

這兩題的相似度為 0.9941。

由以上的例子可知，系統能判別試題內辭彙不同、但意義相同的試題。

兩年度的試題，由研究者判斷其相似的試題組合共有 7880 組，計算其召回率與精確率，得到表 4-25，由表 4-25 可知，取前 10000 組相似試題，召回率為 0.6098，表示已有 4806 組相似試題被找出。

表 4-25 92 年度與 93 年度相似試題的數量與精確率、召回率的關係(在判斷相似試題的情況時)

系統判斷出的相似試題數量	召回率	精確率
50	0.0063	1.0000
100	0.0127	1.0000
150	0.0190	1.0000
200	0.0254	1.0000
250	0.0317	1.0000
300	0.0381	1.0000
350	0.0438	0.9857
400	0.0499	0.9825
450	0.0562	0.9844
500	0.0626	0.9860
550	0.0689	0.9873
600	0.0752	0.9883
650	0.0816	0.9892
700	0.0874	0.9843
750	0.0935	0.9827
800	0.0986	0.9713
850	0.1039	0.9635
900	0.1103	0.9656
950	0.1165	0.9663
1000	0.1221	0.9620
1500	0.1729	0.9087
2000	0.2217	0.8735
2500	0.2589	0.8160
3000	0.2970	0.7803
3500	0.3233	0.7280
4000	0.3531	0.6958
4500	0.3807	0.6667
5000	0.4109	0.6476
5500	0.4352	0.6236
6000	0.4581	0.6017
6500	0.4815	0.5838

7000	0.5040	0.5674
7500	0.5279	0.5547
8000	0.5471	0.5390
8500	0.5634	0.5224
9000	0.5806	0.5084
9500	0.5975	0.4957
10000	0.6098	0.4806
15000	0.7165	0.3765
20000	0.7894	0.3111
25000	0.8461	0.2667
30000	0.8957	0.2353
35000	0.9401	0.2117
40000	0.9727	0.1917