計畫編號：101-EC-17-A-34-S1-149

財經語料詞義探勘平台建置及其於風險測度與證券異常報酬之應用

新聞量化指標網頁之說明文件

102 年 10 月

銘傳大學財務金融研究中心

111 台北市士林區中山北路五段 250 號　(02)28824564#2514

目錄

表目錄

圖目錄

壹、 新聞量化指標之意涵

一、 公開訊息淨樂觀程度(SR)

公開訊息淨樂觀程度(Sentiment Ratios, SR)，是以本研究構建之人工智慧樂觀詞與悲觀詞為基礎，統計出該篇新聞之情緒詞詞頻數。統計新聞文詞在各情緒詞分類中之詞頻，將樂觀詞詞頻總數減掉悲觀詞詞頻總數後，調整為機率百分比之概念。樂觀詞及悲觀詞分別呈現於表 1 及表 2。

二、 危機事件發生強度之量化指標(ITDC)

危機事件發生強度之量化指標(Intensity of Default-Corpus, ITDC)，每一家樣本公司之財務危機發生率強度，係由該樣本公司之財務危機詞發生頻率與非財危詞發生頻數交互比對而來，因此本研究利用財務危機詞頻率對非財務危機詞頻數之比值，定義出該公司文詞語義中的財務危機發生強度。危機詞及非危機詞分別呈現於表 3 及表 4。

表 1 樂觀詞

上探	不落人後	下單	上攻漲停板	不同凡響
不畏	上漲停	世界第一	上揚激勵	不俗
上揚	不虞匱乏	人氣不墜	亮眼激勵	世界前矛
不遜色	上漲有助於	人氣旺	上揚走高	不看淡
佳績	不遑多讓	供不應求	依舊火熱	上攻
上市	人氣鼎沸	令人驚豔	上揚突破	亮眼
不負眾望	倍增	上漲停板	上漲領頭羊	人氣增溫
上漲逼近	不墜	上看	不錯	亮麗

註：上表僅列出部分樂觀詞供參考。

表 2 悲觀詞

不景氣	不良	不樂觀	低檔	供給不足
低過	不振	倒債	不敵	偏低
不穩	假扣押	不易	假造	不當
倒閉	不穩定	停產	不順	停滯
下跌超過	不好	低於	偽造	不彰
不足	債信危機	下降	不及	債務
不理想	人氣退潮	付諸流水	債信風暴	不佳衝擊
丟棄	下跌損失	侵蝕	令人失望	債務高築

註：上表僅列出部分悲觀詞供參考。

表 3 危機詞

上訴	下台	下挫	下滑	下波修正
不如人意	不振	不明朗	不景氣	不樂觀
不幸	不翼而飛	世事難料	不足	不良
低於	低迷不振	供應吃緊	供給過剩	供給過量
信心不足	倒帳	倒楣	停產	破產
停工	債務償還	債台高築	利空	剝奪
動盪不安	危險	勞資糾紛	失望	失靈
威脅	延滯	成長趨緩	拋售	挫敗

註：上表僅列出部分危機詞供參考。

表 4 非危機詞

上升	上揚	上攻	上看	人潮帶動錢潮
佳音	併購	分紅	創新	創舉
創新高	力挺	功耗低	受惠	可望
合作	可觀	協力	卓越	合資
大買	大量出貨	大躍進	居全球之冠	成長
投資	成立	拉升	成長強勁	擴展
攜手合作	支持	攀升	滿單滿產	滿載
站上	突破	穩定成長	發展	登陸

註：上表僅列出部分非危機詞供參考。

貳、 新聞量化指標之演算概念

一、 公開訊息淨樂觀程度(SR)

其方法主要針對斷詞處理後新聞文本，統計分詞在各情緒詞分類中之詞頻後，將樂觀詞詞頻總數減掉悲觀詞詞頻總數後，除上該篇新聞總詞頻數，最後調整為機率百分比之概念，公式如下：

$$SR_{i,t} = \frac{\sum_{i=1}^{I} ptf_{i,t} - \sum_{i=1}^{I} ntf_{i,t}}{TF_{i,t}} \times 100\%$$

其中 $ptf_{i,t}$ 為第 i 間公司在第 t 期新聞報導之樂觀詞詞頻數；$ntf_{i,t}$ 為第 i 間公司在第 t 期新聞報導之悲觀詞詞頻數；$TF_{i,t}$ 為該篇新聞斷詞後之總詞頻。

二、 危機事件發生強度之量化指標 (ITDC)

其方法係由樣本公司之財務危機詞發生權值與非財務危機詞發生權值交互比對而來，定義出該公司文詞語意中的財務危機發生強度，計算方法如下：

$$ITDC_i = \frac{\sum_j tf_{ij}^{D} W_{ij}^{D}}{\sum_j tf_{ij}^{ND} W_{ij}^{ND}}$$

其中，D 為危機樣本；ND 為非危機樣本；tf_{ij} 為第 i 家公司在第 j 個特徵詞上的詞頻，W_{ij} 為第 i 家公司在第 j 個特徵詞上的權重。

參、　新聞量化指標運算架構及資料庫之串聯

時報資訊每日提供研究團隊中國時報、工商時報與中時電子報當日之新聞語料，進行新聞量化指標之運算。研究團隊欲將處理每日時報資訊所提供之新聞語料，建立時報資訊資料庫儲存量化指標運算過程所需之表單。研究團隊取得時報資訊之語料授權，授權期間為：2001 年至今，並儲存至資料庫表單：**OriginalNews**，紀錄原始新聞語料上之股票代碼、新聞發布日期、原始新聞檔案中之新聞序號及報別等欄位，已呈現原始新聞之原貌。

進入量公開訊息淨樂觀程度(SR)與公開訊息之危機發生強度指標(ITDC)前，須透過斷詞系統處理，將原始新聞處理成多個分詞構成之文本，請參閱**表 5 斷詞系統處理後之文本結構**。將斷詞系統處理後之各分詞，透過與樂、悲觀詞及危機、非危機詞之比對，判斷出各分詞是屬於樂、悲觀詞及危機、非危機詞之屬性，並統計該則新聞內屬於樂、悲觀詞及危機、非危機詞之分詞出現之頻率，即詞頻統計。並將詞頻統計之結果，儲存至資料庫表單：**TermFrequency**。

表 5 斷詞系統處理後之文本結構

Panel A 未經斷詞系統處理之文本

國泰大樹計畫 億元助學
2013-10-23
【國泰世華銀行基金會】
【記者張中昌／台北報導】
國泰世華銀行基金會 102 學年度第一學期「大樹計畫」助學金捐贈典禮昨（22）日舉行，由國泰世華銀行董事長暨基金會執行董事陳祖培主持，頒發助學金給宜蘭、雲林、彰化、嘉義、南投五縣。本學期共捐贈 500 萬元助學金，總計有 125 所國中、小學，共 4,313 名學童受惠，用於支付學童的代收代辦費等。
陳祖培表示，為協助弱勢學童能平等受教育，國泰世華銀行基金會自 93 年開始推動「大樹計畫」，十年來已捐出逾 1 億元，共幫助逾 63,000 名清寒學童繳交學雜費。基金會同時也著重心靈層面照顧，辦理「童書募集活動」，至今已達 72,817 冊，鼓勵孩子藉由閱讀充實自我，創造更優質未來。….. (略)

Panel B 斷詞系統處理後之文本

註：上述舉例之語料來源取自於 2013 年 10 月 23 日之中時電子報，其中 Panel A 為原始新聞，Panel B 截取部分文本之斷詞結果。

資料來源網址：http://money.chinatimes.com/express/express-content.aspx?id=21403&cid=1。

接著進行量化指標之運算，透過**貳、新聞量化指標之演算概念**將詞頻統計之結果帶入，即可運算出公開訊息淨樂觀程度(SR)與公開訊息之危機發生強度指標(ITDC)，將其結果分別儲存至資料庫表單：**SRData** 與 **ITDCData**。進一步進一步將量化指標運算結果，透過補日曆日概念之處理，利於量化指標運算結果於「新聞量化指標」網頁上呈現之完整性，並將其結果儲存至資料庫表單：**InfotimesSR** 與 **InfotimesITDC**。其新聞量化指標運算流程，請參閱圖 1。

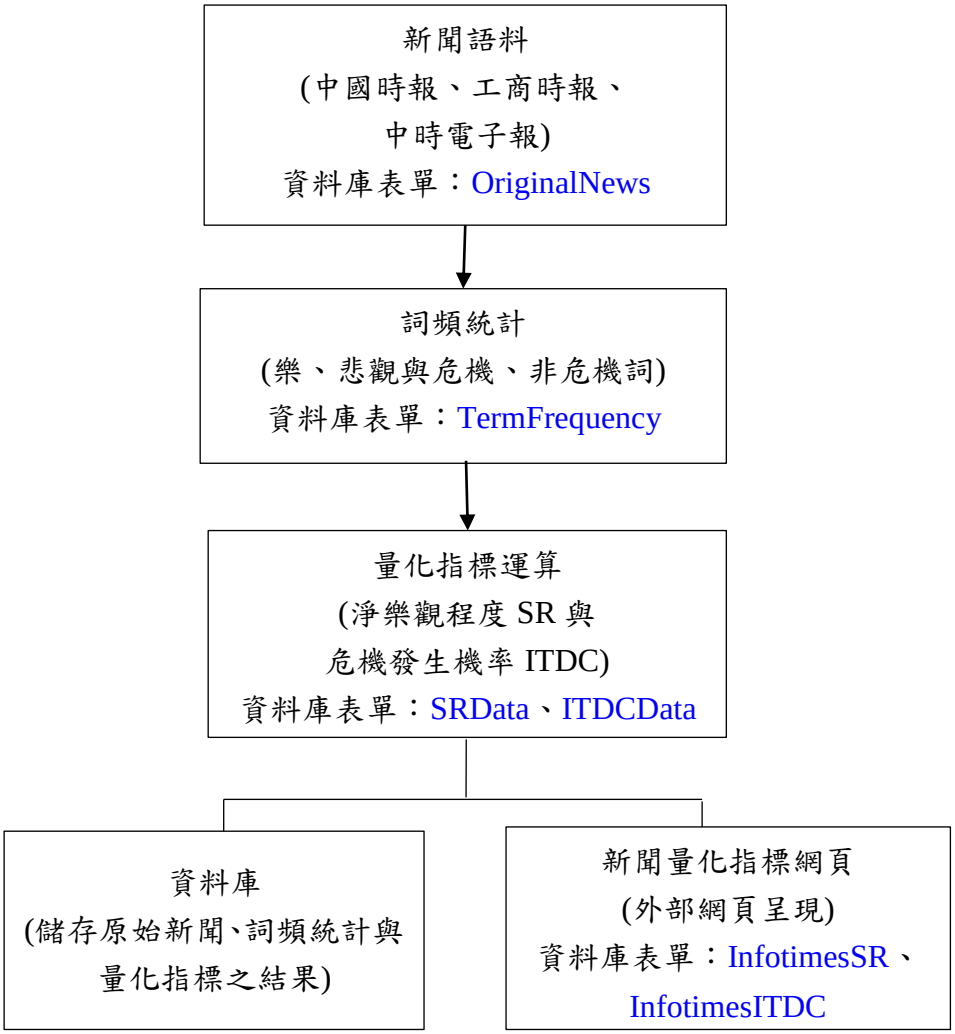

圖 1　新聞量化指標運算流程

註：資料庫表單之欄位說明，請參閱錯誤! 找不到參照來源。錯誤! 找不到參照來源。。

肆、 新聞量化指標網頁

時報資訊每日提供中國時報、工商時報與中時電子報當日之新聞語料，研究團隊將其進行新聞量化指標之運算。目前研究團隊已達到每日將新聞語料透過工作排程器設定進行新聞量化指標之運算，並呈現於「新聞量化指標」網頁。

銘傳大學與時報資訊共同開發之「新聞量化指標」網頁(網址：http://nqi.chinatimes.com/default.aspx)下建立「量化指標排名」、「量化指標查詢」、「量化指標試算」等之網頁查詢功能，請參閱圖 2 新聞量化指標網頁。

圖 2 新聞量化指標網頁

新聞量化指標網頁功能說明如下：

一、　量化指標排名

使用者可於「量化指標排名」網頁上自行選擇查詢參數進行量化指標排名之查詢。其參數包括欲查詢量化指標之期間、查詢公開訊息淨樂觀程度(SR)或危機事件發生強度之量化指標(ITDC)、排名最高或最低，及排名「前 10 名」、「前 20 名」、「前 30 名」等。使用者可透過上述參數設定並查詢，其查詢結果呈現包括股票代碼、公司名稱與量化指標之結果。

如圖 3 為例，使用者欲查詢所有上市櫃公司於 2013/10/29 之前半年公開訊息淨樂觀程度(SR)排名最高前 10 名的公司。

圖 3 量化指標排名

二、　量化指標查詢

使用者可於「量化指標查詢」網頁上自行選擇查詢參數進行量化指標之查詢。其參數包括檢索期間，即設定西元年月日為查詢基準點向前追朔「一周」、「一個月」、「一季」或「一年」；選公司，即使用者可設定欲查詢之上市櫃公司，新聞

來源，即時報資訊之報系包括中國時報、工商時報、中時電子報，呈現方式，即量化指標之呈現包含以 SR、ITDC 與 SR&ITDC 等方式。

如圖 4 為例，使用者欲查詢台泥(1101)於 2013/10/29 向前追朔一年之量化指標，其資料來源設定來自於中國時報、工商時報與中時電子報，量化指標結果以 SR&ITDC 之方式呈現，請參閱**圖 5 量化指標查詢之結果**。

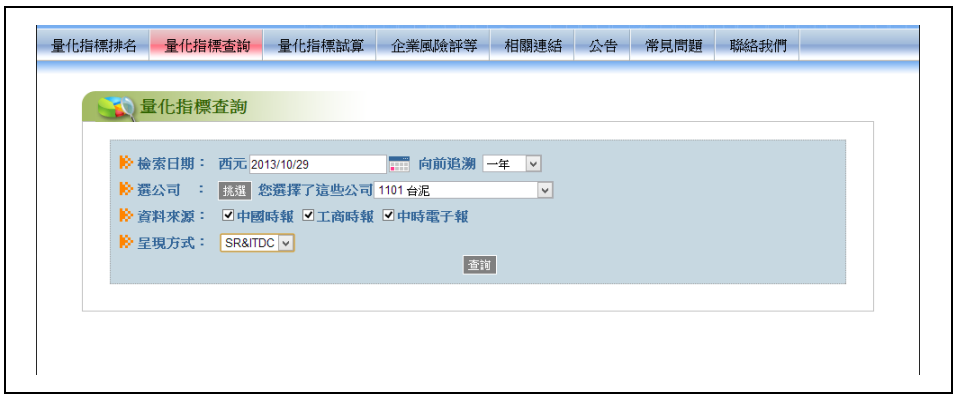

圖 4 量化指標查詢之參數設定

股票代碼	公司名稱	日期	淨樂觀程度(SR)	危機發生指標(ITDC)
1101	台泥	2013/10/06	59.5745	0.0742
1101	台泥	2013/10/05	50.4092	0.0951
1101	台泥	2013/10/04	59.5745	0.0951
1101	台泥	2013/10/03	50.4092	0.0989
1101	台泥	2013/10/02	-18.0925	4.1912
1101	台泥	2013/10/01	7.4481	3.5139
1101	台泥	2013/09/30	-39.9368	4.5667

圖 5 量化指標查詢之結果

三、　　量化指標試算

使用者可於「量化指標試算」網頁上進行量化指標之試算。使用者除了可透過「量化指標查詢」網頁上查詢時報資訊所提供之新聞量化指標外，也可至「量化指標試算」之線上試算平台中放入一則新聞進行量化指標之試算，其試算結果之呈現，包括公開訊息淨樂觀程度(SR)或公危機事件發生強度之量化指標(ITDC)之量化指標運算結果，以及該則新聞中出現之樂、悲觀詞與危機、非危機詞及詞頻統計之結果，請參閱圖 6 至圖 7。

圖 6 量化指標試算

圖 7 量化指標試算之結果

附件 一、中研院平衡語料庫詞類標記集

簡化標記	對應的CKIP詞類標記	
A	A	/*非謂形容詞*/
Caa	Caa	/*對等連接詞，如：和、跟*/
Cab	Cab	/*連接詞，如：等等*/
Cba	Cbab	/*連接詞，如：的話*/
Cbb	Cbaa, Cbba, Cbbb, Cbca, Cbcb	/*關聯連接詞*/
Da	*Daa*	/*數量副詞*/
Dfa	Dfa	/*動詞前程度副詞*/
Dfb	Dfb	/*動詞後程度副詞*/
Di	Di	/*時態標記*/
Dk	Dk	/*句副詞*/
D	*Dab*, Dbaa, Dbab, Dbb, Dbc, Dc, Dd, Dg, Dh, Dj	/*副詞*/
Na	Naa, Nab, Nac, Nad, Naea, Naeb	/*普通名詞*/
Nb	Nba, Nbc	/*專有名稱*/
Nc	Nca, Ncb, Ncc, Nce	/*地方詞*/
Ncd	Ncda, Ncdb	/*位置詞*/
Nd	Ndaa, Ndab, Ndc, Ndd	/*時間詞*/
Neu	*Neu*	/*數詞定詞*/.
Nes	*Nes*	/*特指定詞*/
Nep	*Nep*	/*指代定詞*/
Neqa	*Neqa*	/*數量定詞*/
Neqb	*Neqb*	/*後置數量定詞*/
Nf	Nfa, Nfb, Nfc, Nfd, Nfe, Nfg, Nfh, Nfi	/*量詞*/
Ng	Ng	/*後置詞*/
Nh	Nhaa, Nhab, Nhac, Nhb, Nhc	/*代名詞*/
I	I	/*感嘆詞*/
P	P*	/*介詞*/
T	Ta, Tb, Tc, Td	/*語助詞*/
VA	VA11,12,13,VA3,VA4	/*動作不及物動詞*/
VAC	VA2	/*動作使動動詞*/
VB	VB11,12,VB2	/*動作類及物動詞*/
VC	VC2, VC31,32,33	/*動作及物動詞*/
VCL	VC1	/*動作接地方賓語動詞*/
VD	VD1, VD2	/*雙賓動詞*/
VE	VE11, VE12, VE2	/*動作句賓動詞*/
VF	VF1, VF2	/*動作謂賓動詞*/
VG	VG1, VG2	/*分類動詞*/
VH	VH11,12,13,14,15,17,VH21	/*狀態不及物動詞*/
VHC	VH16, VH22	/*狀態使動動詞/
VI	VI1,2,3	/*狀態類及物動詞*/

VJ	VJ1,2,3	/*狀態及物動詞*/
VK	VK1,2	/*狀態句賓動詞*/
VL	VL1,2,3,4	/*狀態謂賓動詞*/
V_2	V_2	/*有*/
DE	/*的，之，得，地*/	
SHI	/*是*/	
FW	/*外文標記*/	