計畫編號：MOST 103-2622-H-130-001

讓電腦讀懂海量財經訊息-財經文詞解析引擎及企業價值與風險評鑑(2/3)

新聞結構化標記系統

103 年 8 月

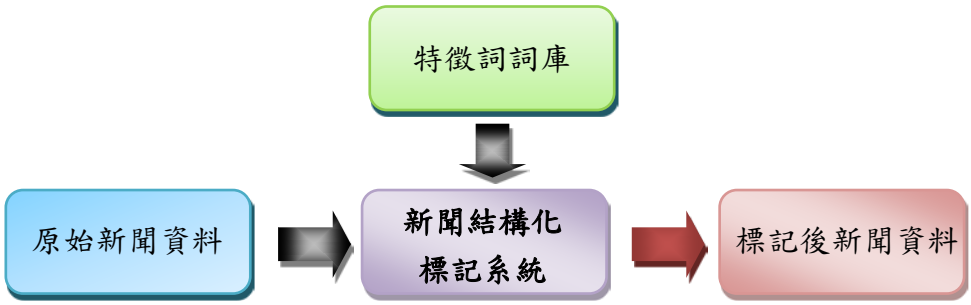
銘傳大學財務金融中心

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

一、 結構化標記系統介紹

每日龐大的海量資訊琳瑯滿目，如何將這些充滿雜訊且非結構化的新聞資料發揮其最大效益是非常重要。透過結構化標記系統可使這些即時且多樣化的海量新聞資料系統化處理，讓使用者能一目了然其標記結果，有助於提升語料後續應用的分類處理效率。

新聞結構化標記系統可將原始語料根據不同詞彙之分類給予不同的標記記號。如圖一新聞結構化標記流程圖所述，首先透過特徵詞篩選模組所建構之特徵詞詞庫，並依據系統格式將不同的特徵詞給予分類，本研究團隊將特徵詞分為「樂觀詞」、「悲觀詞」、「危機詞」、「非危機詞」四大類，再將原始新聞資料依據系統格式匯入新聞結構化標記系統中，即可得到標記後之新聞資料。除新聞情緒特徵詞之分類標記外，結構化標記亦包括原始新聞之見報日、引題、主標、副標、內文、版別等分類標記，新聞格式之標記有助於模組之建構，透過原始新聞篩選條件之設定，快速查詢特定新聞標記結果。



■ 圖 1、新聞結構化標記流程圖

二、 範例說明

透過新聞結構化標記系統，可將原始新聞之見報日、引題、主標、副標、內文、版別等標示清楚，如下圖 3，以利後續原始新聞匯入資料庫能符合欄位格式，避免造成欄位錯誤顯示或新聞資料無法匯入之情形。

其標記欄位說明如下：

(1) 見報日：新聞發佈之時間

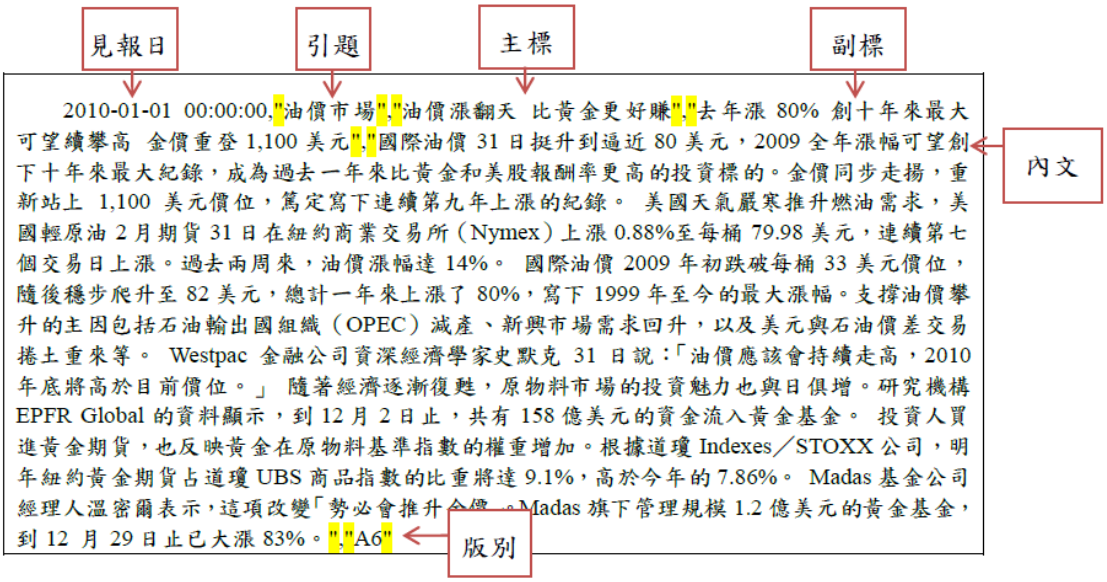
(2) 引題：在主標前，具引導的效用，讓讀者在閱讀新聞前了解情境

(3) 主標：新聞主要之大標題

(4) 副標：新聞之副標題

(5) 內文：新聞之文章內容

(6) 版別：位於報紙之版面位置



■ 圖 3、新聞格式與限制範例

經過新聞結構化標記系統，得到標記後新聞資料，如下圖 4，在文章中標記特徵詞詞彙，有助於讀者快速且正確的辨別該篇新聞情緒。

標記系統處理之特徵詞標記共包含以下之類別：

1. 「**樂觀詞**」

   (1) 定義: 對未來之市場有正面性的看法及憧憬，即對未來市場充滿信心之詞彙

   (2) 標記記號: 文字顯示為紅色

2. 「**悲觀詞**」

   (1) 定義:對於未來之市場具較反面性的看法，即對未來市場不具信心之詞彙

   (2) 標記記號: 文字顯示為綠色

3. 「**危機詞**」

   (1) 定義: 表示公司未來之發展，可能存在負面衝擊，如發生「倒閉、重整、跳票擠兌、抒困求援、淨值為負、全額交割或下市」等之負面事件之相關詞彙

   (2) 標記記號: 整段文字之背景顯示為黃色

4. 「**非危機詞**」

   (1) 定義: 此詞彙對於公司未來之發展，有正面之意涵

   (2) 標記記號: 整段文字之背景顯示為藍色

■ 圖 4、完成新聞結構化標記之新聞語料