中華民國計算語言學學會
通訊 ACLCLP
The Association for Computational Linguistics and Chinese Language Processing
第二十卷第五期
發行人：吳宗憲
主　編：王新民
執行編輯：黃　琪
E-Mail：aclclp@hp.iis.sinica.edu.tw
Website：http://www.aclclp.org.tw
地 址：台北市研究院路二段128號中研院資訊所
電 話：(02)2788-3799 ext:1502
傳 真：(02)2788-1638
劃 撥：19166251

本 期 要 目

2010 International Conference Listings

ACL- 2010

The 48th Annual Meeting of the Association for Computational Linguistics
- Conference Date: July 11-16, 2010
- Submission Deadline: February 8, 2010
- Location: Uppsala, Sweden
- http://acl2010.org/

ACM SIGIR-2010

The 33rd Annual ACM SIGIR Conference
- Conference Date: July 18-23, 2010
- Submission Deadline: January 15, 2010
- Location: Geneva, Switzerland
- http://www.sigir2010.org/

COLING -2010

The 23rd International Conference on Computational Linguistics
- Conference Date: August 23-27, 2010
- Submission Deadline: April 19, 2010
- Location: Beijing, China
- http://www.coling-2010.org/

DiSS-LPSS Joint Workshop-2010

The 5th Workshop on Disfluency in Spontaneous Speech and the 2nd International Symposium on Linguistic Patterns in Spontaneous Speech
- Conference Date: September 25-26, 2010
- Submission Deadline: May 31, 2010
- Location: Tokyo, Japan
- http://cogsci.l.chiba-u.ac.jp/diss-lpss2010/

ICASSP-2010

The 35th International Conference on Acoustics, Speech, and Signal Processing
- Conference Date: March 14-19, 2010
- Submission Deadline: September 14, 2009
- Location: Texas, U.S.A.
- http://www.icassp2010.com/

ICCPOL- 2010

The Joint Conference of 23rd International Conference on the Computer Processing of Oriental Languages-New Generation in Asian Information Processing
- Conference Date: July 1-3, 2010
- Submission Deadline: January 31, 2010
- Location: California, U.S.A.
- http://www.ksi.edu/seke/iccpol10cfp.html

INTERSPEECH- 2010

The 11th Annual Conference of the International Speech Communication Association
- Conference Date: September 26-30, 2010
- Submission Deadline: April 30, 2010
- Location: Makuhari, Japan
- http://www.interspeech2010.org/

ISCSLP -2010

The 7th International Symposium on Chinese Spoken Language Processing
- Conference Date: November 29-December 3, 2010
- Submission Deadline: July 15, 2010
- Location: Tainan, Taiwan
- http://conf.ncku.edu.tw/iscslp2010/

NAACL HLT-2010

The 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics- Human Language Technologies
- Conference Date: June 1-6, 2010
- Submission Deadline: December 1, 2009
- Location: Los Angeles, U.S.A.
- http://naaclhlt2010.isi.edu/

行動資訊檢索暨行動定位服務技術研討會

Workshop on Mobile Information Retrieval and Location Based Service

隨著全球資訊網快速蓬勃的發展，各式各樣的資訊內容和服務不斷地擴增，人群的互動方式逐漸轉移到網路平台，並且伴隨著 WIMAX 及 WIRLESS 快速拓展，人們獲得資訊方式由傳統定點裝置演變到攜帶輕巧的數位行動裝置，以便隨時隨地掌握、接收各式各樣的資訊。因此，於行動裝置有效率的將資料索引(Indexing)、檢索(Retrieval)與呈現(Representation)等課題日趨重要，於是學者提出行動資訊檢索技術(Mobile Information Retrieval)，並且在生活應用的範疇之下，更進一步地結合定位服務技術(Location Based Service, LBS)，以提供因時制地便捷之整合資訊服務，目前有許多的創新研究和重要應用議題已逐漸受到學術界和產業界重視。中央研究院資訊科學研究所、國立台灣大學資訊工程學系、國立成功大學資訊工程學系與中華民國計算語言學學會特別舉辦本次研討會，邀請國內外相關學者專家進行觀念和技術交流。此研討會係繼 2002 年「資訊自動分類技術研討會」、2003 年「資訊檢索與電腦輔助語言教學研討會」、2004 年「文件探勘技術研討會」、2005 年「網路資訊檢索技術與趨勢研討會」、2006 年「網路探勘技術與趨勢研討會」、2007 年「Web 2.0 技術與應用研討會」以及 2008 年「網路社群服務計算暨探勘技術研討會」之後續的年度會議活動。歡迎各界人士踴躍參加。

議程：

時間	講題	主講人
09:00 ~ 09:30	**報到**	
09:30 ~ 10:30	Efficient Support of Location-Based Services on Roads	李旺謙 教授 (Prof. Wang-Chien Lee) Department of Computer Science and Engineering, Pennsylvania State University
10:30 ~ 11:00	**休息**	
11:00 ~ 12:00	Ubiquitous Music Recommendation Using Content- and Context-Aware Information Mining	曾新穆 教授 (Prof. Vincent S. Tseng) Department of Computer Science and Information Engineering, National Cheng Kung University
12:00 ~ 13:30	**午餐**	
13:30 ~ 14:30	CarWeb: Sharing GPS data Points for Traffic Estimation and Trajectory Pattern Mining	彭文志 教授 (Prof. Wen-Chih Peng) Department of Computer Science, National Chiao Tung University
14:30 ~ 15:30	The Multi-Rule Partial Sequenced Route Query	顧維信 教授 (Prof. Wei-Shinn Ku) Department of Computer Science and Software Engineering, Auburn University
15:30 ~ 16:00	**休息**	
16:00 ~ 17:00	車載資通訊前瞻應用技術與發展趨勢	王景弘 博士 (Dr. Ching-Hung Wang) Telecommunication Laboratories, Chunghwa Telecom Co.

行動資訊檢索暨行動定位服務技術研討會

報　名　表

編號：

會議時間：98 年 12 月 21 日(星期一)

會議地點：台北市中央研究院資訊所一樓 106 演講廳

主辦單位：中央研究院資訊所、台灣大學資訊工程學系
成功大學資訊工程學系、中華民國計算語言學學會

姓　　名		單　位	
電　　話		E-mail	
聯絡地址			
收據抬頭			
報名費	一般人士：　□ 會員 400 元　　□ 非會員 600 元 學生：　□ 會員 200 元　　□ 非會員 300 元 (報名費含講義、午餐及茶點)		

1. 學生非會員請傳真學生證。
2. 報名及繳費截止日：**12/14** 日(現場報名加收 200 元)。
3. 繳費方式：
   □ 郵政劃撥：帳號 19166251，戶名：中華民國計算語言學學會
   □ 信用卡：

   信用卡持有人：＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿(請以正楷書寫）

   卡號：＿＿＿＿＿＿-＿＿＿＿＿＿-＿＿＿＿＿＿-＿＿＿＿＿＿

   卡片有效期：＿＿＿＿卡片末三碼：＿＿＿＿發卡銀行：＿＿＿＿＿＿

   持卡人簽字：＿＿＿＿＿＿＿＿＿＿＿＿＿＿(簽名方式請與信用卡相同)
4. □ 午餐請備素食。
5. 報名表請傳真或 E-mail 至：
   Fax. No.: (02)2788-1638, E-mail: jessie@hp.iis.sinica.edu.tw。
6. 聯絡電話：(02)27883799 分機 1502，聯絡人：黃琪 小姐。

CALL FOR PAPERS
2010 International Symposium on Chinese Spoken Language Processing (ISCSLP 2010)

November 29 – December 3, 2010 - Tainan and Sun Moon Lake, Taiwan
http://conf.ncku.edu.tw/iscslp2010/

ISCSLP is the flagship conference of ISCA SIG-CSLP (International Speech Communication Association, Special Interest Group on Chinese Spoken Language Processing).

ISCSLP2010 will be held during November 29 – December 3, 2010 in Tainan and Sun Moon Lake, Taiwan, and hosted by National Cheng Kung University.

Tainan, located in south of Taiwan, is the cultural place of origin. There are a lot of historical sites and antique places. Therefore, Tainan has another name called the ancient capital. In addition to these historical sites, Tainan is a modern city with various shopping centers, department stores, and night markets. You can taste all of Taiwanese traditional food, e.g. shrimp roll, tofu pudding, coffin, oyster omelet, rice tube pudding, Taiwanese meatballs in Tainan . Don't miss this chance to experience these toothsome snacks when you visit Tainan .

Sun Moon Lake, situated in Nantou County's Yuchih Township, in the center of Taiwan, is the island's largest lake. It is a beautiful alpine lake, divided by the tiny Lalu Island; the eastern part of the lake is round like the sun and the western side is shaped like a crescent moon, hence the name "Sun Moon Lake". Its crystalline, emerald green waters reflect the hills and mountains which rise on all sides. Natural beauty is enhanced by numerous cultural and historical sites. Well-known both at home and abroad, the Sun Moon Lake Scenic Area has exceptional potential for further growth and recognition as a prime tourism destination. When you come to Sun Moon Lake, slow down your hurried travels, stay, and relax for a few days. Here, the beauty of nature will make you want to keep coming back!

We invite your participation in this premier conference, where the language from ancient civilizations embraces modern computing technology. ISCSLP2010 will feature world-renowned plenary speakers, tutorials, exhibits, and a number of lecture and poster sessions on the following topics:

- Speech Production and Perception
- Phonetics and Phonology
- Speech Analysis
- Speech Coding
- Speech Enhancement
- Speech Recognition
- Speech Synthesis
- Language Modeling and Spoken Language Understanding
- Spoken Dialog Systems
- Spoken Language Translation
- Speaker and Language Recognition
- Computer-Assisted Language Learning
- Indexing, Retrieval and Authoring of Speech Signals
- Multi-Modal Interface including Spoken Language Processing
- Spoken Language Resources and Technology Evaluation
- Applications of Spoken Language Processing Technology

Important Dates

- Full paper submission by July 15, 2010
- Notification of acceptance by Aug. 30, 2010
- Camera ready papers by Sep. 13, 2010
- Registration to cover an accepted paper by Oct.13, 2010

中文意見分析之概況、技術與應用

古倫維　陳信希

國立台灣大學資訊工程系

意見分析簡介

意見分析主要的目的在於利用電腦自動分析人所發表的意見。在資訊爆炸的時代，爲了解大眾的意見，有太多需要閱讀的文件，要一篇篇把它們看完，需耗費相當多的時間，這樣的環境，正突顯了意見分析技術的重要性。意見分析的首要任務是探勘意見，將之擷取出來，並加以整理與應用，讓使用者最終能獲得有用的資訊。

要分析意見，首先得先定義出何謂意見。一個意見是一個能表達主觀立場的文本，其中主要的元素包括意見的極性，也就是正面、中立、或負面；意見的強度；意見持有者，也就是表達意見的人或組織；以及意見的目標，亦即被評論的對象。意見分析是整個研究範疇的統稱，但其中針對這些不同的元素的研究，產生了許多子範疇，例如意見擷取 (將包含意見的文本找出來)、意見的極性判定 (自動判定目前意見所持立場)、意見持有者辨識 (找出評論者)、及意見目標辨識 (找出被評論的對象) 等，在這些子範疇中，分別發展出了各種不同的技術。

意見分析經常與情緒分析一同被討論，意見分析有時也被包含在情緒分析之中，但意見與情緒並不完全相同，且兩者不必然相關。意見，如前所提，用來表達立場，而情緒則是一種心理狀態。情緒的議題一般在心理學上已發展出一些分析方式，例如以激動程度爲一軸，情緒狀態 (如快樂程度) 爲另一軸畫出一情緒分析平面，或者是將各類情緒 (快樂、悲傷、憤怒、驚訝等) 加以分類。無論如何，發表意見立場時不必然帶有情緒，特定情緒也不必然連結到特定立場，例如「開心地同意」、「悲傷地同意」、或「堅定地同意」都可以表達支持立場。

帶有意見的文本可以是詞彙、句子、或是文章。詞彙可以表達一個完整的意義，因此可以帶有意見；句子則是一般用來表達立場最自然的單位；而文章則通常含有數個意見，但整體仍能表達特定立場。對於詞彙，如果我們能夠知道它的意義，也就能知道它是否帶有意見，或者目前也有意見詞典可供查詢；對於文章，如果我們能夠正確判斷文中所提大部份的意見，即使少數意見無法正確判別，仍然能夠決定整體的立場；但在句子中，不同的構句方式，將詞彙以不同的方式組合起來，即使用詞很相似，仍能表達各式各樣的意見。因此一般咸認爲句子層次的意見分析較爲困難。

意見分析這個議題，之所以從自然語言相關議題中被區分出來，還有另一個與此議題的本質有關的原因：意見與人類的認知有關，因此本身就帶有不一致性。因其成長背景、知識與立場的不同，同樣的一句話由不同的人讀來，可能會有不同的感受。因此在產生意見分析的標準答案時，並不如其他議題，只要根據專業的知識給予正確的標記即可，而是需要參考多數人的想法，產生一個代表群體想法的可靠標記結果。

但意見分析並不因此與其他自然語言的技術相分離。爲了要能夠自動分析大量文章內的意見，除了與意見分析本身相關的技術，也牽涉到其他自然語言處理或資訊檢索的議題。例如，在實用上通常會先收集與特定主題相關的意見，再加以分析，這時就需要檢索相關句的技術；如果希望能夠分析不同國家，不同文化的人的立場，就需要機器翻譯的技術；如果目標是收集一些特定人物的意見並加以比較，那麼擷取姓名實體與建立同指涉鏈的技術更是不可或缺。因此我們可以說，意見分析的技術要能夠被應用，其他技術的配合亦不可少。

接下來我們將介紹意見分析的相關研究與此領域目前的研究近況，並說明我們所提出的中文意見分析技術，最後展示數個此技術可能的應用。

相關研究

意見分析這個議題，在 2001 年就已被提出來討論 (Wiebe *et al.*, 2001)，到了 2002 年，Pang 及他的研究團隊發表了以機器學習的方法來分類意見的論文 (Pang *et al.*, 2002)，此議題便開始廣泛地受到研究者的注意與重視，之後相關的研究也如雨後春筍般出現：有研究者專注於處理意見詞彙 (Kim and Hovy, 2004)，也有研究者分析句子與文章 (Riloff and Wiebe, 2003)。與意見分析本身相關的技術，大抵分成以語言學知識爲本與以機器學習爲本兩類。使用語言學的知識，無論是找尋特定的語言結構或資訊，都需要專業；而使用機器學習的方式雖然比較上不那麼依賴語言學的專業，但目前研究者提出的實驗結果，效果卻不像應用在其他自然語言領域那麼好。因此在技術方面，還有研究改進的空間。

在各種文體的文件上，都需要意見分析的技術。目前常做爲意見分析文章來源的包括評論網站的文章、新聞文件、以及部落格文件。因爲評論網站的文章格式較一致，且通常主題明確，因此最常被用來找尋意見，常見的此類文本像是 3C 產品或電影的評論 (Bai et al., 2005; Ghose et al., 2007)。與其他文本相比，從評論裡擷取意見並加以分析時，必須先找出產品的評價特徵 (Hu and Liu, 2004)，例如對相機來說，評價特徵包括鏡頭、電池壽命、外觀、使用方式、色彩飽和度等等；而對電影來說，評價特徵則包括劇情、演員演技、音樂、化妝等等。對不同的產品就需要找出不同的評價特徵是從評論裡分析意見的一個挑戰。另一方面，如果我們想從新聞或部落格的文章中擷取出意見，由於主題通常埋藏在文章之中，主題與相關度偵測就成爲處理這類文本的挑戰 (Ku *et al.*, 2005; Branavan *et al.*, 2008)。在新聞文件中，除了內容之外，文中所提到的具名實體也很重要，通常可以做爲擷取意見持有者，或是評論對象的分析資料，並以此爲依據整合各方意見，目前意見持有者及評論對象的擷取，尤其是評論對象的擷取，因爲牽涉到概念的擷取 (有時被評論的是一個概念)，在意見分析領域中，仍屬於較新的議題 (Choi *et al.*, 2005; Kim and Hovy, 2005; Breck *et al.*, 2007)。

意見分析技術的應用並非僅限於資訊領域。其他領域也有研究者從事意見分析的相關研究，例如政治科學 (Martin and Vanberg, 2008)、市場行銷 (Chen and Xie, 2008)、廣告 (Jin et al., 2007)、政府政策 (Cardie et al., 2006) 等等。實際上，知道其他人的意見，

在很多領域都是相當實用的，只是過去皆由人工調查的方式得知，有了意見分析的技術，將可從更大量的資料中整理出更可靠的資訊供參考與應用，解決因爲依賴人工而造成的成本過高與效率不彰等問題。

中文意見分析技術

爲了分析意見，首先我們必須設計用來探勘意見的演算法。爲了開發演算法，必須找尋語料供學習之用，並準備一個良好的評估環境，以便在開發過程中適當地改進演算法。以下我們將介紹開發意見分析技術的資源、意見分析實驗的語料、意見分析的演算法、以及意見分析實驗的評估方式。

意見分析相關資源

前面提到，意見分析最小的單位是詞彙，詞彙已能夠表達一個完整的意涵。因此在看到某一個詞彙的時候，我們希望能夠知道這個詞彙是否帶有意見，又詞彙的意見是屬於正面、中立、或負面，最簡單容易的方法就是使用詞典查詢。

目前在各語言中，還是以英文的資源最多。在英文的部份，目前與意見詞彙相關的資源，較廣爲人知的包括有最早期由心理學家所制訂的詞典 General Inquire[1]，還有根據 Wordnet[2]的詞彙語義架構所產生的 SentiWordnet[3]等等。

在中文的部份，我們也建立了一部意見詞典 NTUSD[4]，供研究者使用。這部詞典主要收集了 General Inquire 的中文翻譯詞彙，並收集了中文的網路意見用語。目前詞典中收錄約一萬個中文詞彙，並將其區分成正面及負面兩類中文詞。之後在準備語料與實驗集的過程中，我們更將 NTUSD 擴展至四萬中文詞，並加入中立意見詞 (仍然帶有意見，只是其意見的立場爲中立，與非意見詞並不相同)。

意見分析的實驗語料

除了自行標記的方式外，要獲得意見分析的實驗語料主要有三個管道，一是從網路上自行下載資料，一是使用其他研究者自行產生並分享出來的語料，另一則是參加技術評比會議，取得評比語料。通常實驗的語料必須包括答案，才能在實驗之後評估效能，因此能從網路下載供意見分析的語料，多屬產品評論類。英文部份，有少數由研究者整理標記後分享給研究社群的語料，例如 MPQA[5]。這類語料可提供研究者一個起始的研究基礎，但缺點是它們通常數量不大。

技術評比會議通常能夠提供研究者一定量的語料，其數量較私人提供的語料爲大，評估方式也公開，並提供免費的評估工具，因此可在同一個基準上比較不同的技術，並

[1] http://www.wjh.harvard.edu/~inquirer/
[2] http://wordnet.princeton.edu/
[3] http://sentiwordnet.isti.cnr.it/
[4] http://nlg18.csie.ntu.edu.tw:8080/opinion/index.html
[5] http://www.cs.pitt.edu/mpqa/

得知它們的優劣。目前在英文部份，包含意見分析評比項目的評比會議有 TREC[6] (TREC-BLOG[7])及 NTCIR[8] (NTCIR MOAT[9])。它們兩者在項目工作目標的定義上不盡相同，主要差別在於，TREC 是以部落格爲標記單位，而 NTCIR 則是以句子爲標記單位；在文類上兩者亦不相同，TREC 使用的是部落格語料，而 NTCIR 使用的是新聞語料。在中文部份，則只有 NTCIR 包含意見分析評比項目 (同時包含正體與簡體中文)。一般來說，欲取得評比會議所提供的實驗語料，通常也必須成爲該評比項目的參加者。

前面提過，意見分析的答案無法只靠專業知識給定。目前較常使用的方式包括利用三個以上的標記者標記同一份資料，以取得標記者之間共同的標記，做爲最後的標準答案；不同的標記者所標記出的答案，也會計算 *kappa* 值並加以分析，做爲評斷標記出的語料庫是否可靠的參考 (Ku *et al.*, 2007)。

意見分析的演算法

前面我們介紹過，意見分析演算法的設計主要分成語言學知識爲本及機器學習方法爲本兩類，兩種概念各自都有許多研究者採用。

目前在中文意見分析機器學習方法爲本的演算法中，效能較好的主要採用多個分類器分類後，再同時考量這些結果，產生最後的答案這樣的概念，來發展技術 (Liu *et al.*, 2008)。

我們在中文上所提出的演算法，是一語言學知識爲本的演算法。我們從兩個不同的角度切入：微觀與巨觀。演算法將會在詞彙、句子、與文章三個不同的層次尋找意見。在巨觀的演算法中，我們關心的是每一個組成成份的意見傾向，我們相信由這些成份的意見傾向可以決定整體的意見傾向，例如詞彙的意見傾向是由字所決定，句子的意見傾向是由詞彙所決定，而文章的意見傾向則是由句子所決定，以此類推。在微觀的演算法中，我們從一個完整的訊息切入，關心它的成份中互相之間的關連與影響，從而決定整個訊息的意見傾向。考慮到各成份之間可能會互相影響，我們引入了構詞學與句子結構等資訊。我們期待微觀的演算法能補足巨觀演算法的不足。

所提出的演算法以詞彙本身所包含的字來幫助判定詞彙是否含有意見，並得知它們的意見傾向。這個概念之所以可行，原因是除了極少數的情況外，中文詞彙的意義，與該詞中所包含的字之字義相關。再加上，中文的詞彙並非是固定的集合，新詞彙隨著時代改變不斷增加，但中文字卻可視爲固定的集合，因此利用中文字的資訊來決定詞彙的意見，可以減少新詞所造成的問題 (Ku and Chen, 2007)。

爲了判定字彙的意見傾向，並給定每個字彙一個意見分數，我們利用上面所提到的意見詞典 NTUSD 一萬字精簡版。我們認爲，較常出現在正面詞的字，本身應帶有正面的意涵，反之亦然，若一字常出現在正面詞彙也常出現在負面詞彙，表示該字對詞彙的

[6] http://trec.nist.gov/
[7] http://ir.dcs.gla.ac.uk/wiki/TREC-BLOG
[8] http://research.nii.ac.jp/ntcir/
[9] http://www2.kde.ics.tut.ac.jp/moat/index.php/MOAT

意見傾向並不帶來影響，或者我們無法根據它的出現來幫助判定詞彙的意見傾向，又或者它本身帶有的是中立的意見。因此我們利用每個字出現在正面詞彙及負面詞彙的機率，訂出該字的意見分數，公式如下：

$$P_{c_i} = \frac{fp_{c_i} / \sum_{j=1}^{n} fp_{c_j}}{fp_{c_i} / \sum_{j=1}^{n} fp_{c_j} + fn_{c_i} / \sum_{j=1}^{m} fn_{c_j}} \qquad (1)$$

$$N_{c_i} = \frac{fn_{c_i} / \sum_{j=1}^{m} fn_{c_j}}{fp_{c_i} / \sum_{j=1}^{n} fp_{c_j} + fn_{c_i} / \sum_{j=1}^{m} fn_{c_j}}, \qquad (2)$$

$$S_{c_i} = (P_{c_i} - N_{c_i}) \qquad (3)$$

公式中 fp_{c_i} 代表出現字彙 *Ci* 的正面詞彙數， fn_{c_i} 代表出現字彙 *Ci* 的負面詞彙數; *n* 是 NTUSD 中所有詞彙中出現過的相異字數， P_{c_i} 與 N_{c_i} 分別代表一個字彙的正面與負面意見分數， S_{c_i} 則是字彙 *Ci* 最後的意見分數。

在巨觀的演算法中，我們並不關心這些字互相之間的影響，只單純利用這些字的意見組合出詞彙的意見，公式如下：

$$S_w = \frac{1}{p} \times \sum_{j=1}^{p} S_{c_j} \qquad (4)$$

S_w 是詞彙的意見分數，*p* 則是該詞中所含有的字數，而其中每個字的意見分數 *Sc* 則由前面的公式 (4) 算出，公式將所有字的意見分數平均後，得到詞彙的意見分數。

但我們由語言學的知識得知，字在組合成詞彙的時候，的確會互相影響。它們之間的關係也由語言學家清楚地定義在構詞規則之中。因此，在微觀的演算法中，我們依據五種不同的構詞規則，計算出詞彙的意見，又由於大部份的中文詞爲二字詞，且三字以上的詞若要分析，又牽涉到切分詞素的問題，所以我們在這裡僅特別將二字詞區分出來，公式如下 (以下公式中，C_1 爲二字詞中的第一個字/詞素，C_2 則爲二字詞中的第二個字/詞素， $S(C_1C_2)$ 爲該詞最後的意見分數)：

並列關係：詞彙中的兩個字處於對等的地位，因此詞彙本身的意見即爲組成詞的意見分數之平均值。這樣的詞例如「疲乏」。它由「疲」與「乏」組成，因此它的意見分數是此二字之意見分數的平均。

$$S(C_1C_2) = \frac{S(C_1) + S(C_2)}{2} \qquad (5)$$

修飾關係：在這個關係中，第一個詞素修飾第二個詞素，因此意見強度應來自第一個詞素的意見分數絕對值；而意見傾向則取決於是否有負面詞素出現，若有則應爲負面，

若兩詞素其中有正面詞素但無負面詞素，則該詞之意見傾向應爲正面。例如詞彙「痛哭」是由詞素「痛」與「哭」組成，其中兩個詞素皆帶有負面意見，因此該詞彙爲負面；而它的意見強度則由第一個詞素「痛」決定，因爲它修飾後面「哭」的動作，說明了該詞彙是哪一種程度的哭泣。又例如在詞彙「浩劫」中，詞素「浩」表明了該詞指的是「很大」的「劫」難；辭彙「高強」中的「高」也說明該詞指的並不是普通的「強」而已，這些都是由不同詞素組成的修飾類詞彙。

$$\begin{aligned}&\text{if } (S(C_1)\neq 0 \text{ and } S(C_2)\neq 0)\\&\quad \text{if } (S(C_1)>0 \text{ and } S(C_2)>0)\\&\qquad S(C_1C_2)=S(C_1)\\&\quad \text{else } S(C_1C_2)=\text{-}1\times\left|S(C_1)\right|\\&\text{else } S(C_1C_2)=S(C_1)+S(C_2)\end{aligned} \tag{6}$$

主謂關係：在主謂詞中，第一個詞素扮演主詞的角色，而第二個詞素則是它所做的動作，因此這類詞的意見通常取決於第二個詞素。如果代表動作的第二個詞素並不帶有意見，或者它是中性的詞素，那麼該詞的意見就由主詞決定。例如在「山崩」這個詞中，因爲「崩」是負面的動作，而這個詞的意見由它決定，而第一個詞素「山」不帶有意見，因此可決定「山崩」是一個負面詞。

$$\text{if } (S(C_2)\neq 0) \text{ then } S(C_1C_2)=S(C_2) \text{ else } S(C_1C_2)=S(C_1) \tag{7}$$

動受關係：在動受詞中，第一個詞素扮演動詞的角色，而第二個詞素則是它的受詞。整個詞的意見不僅只由動作決定，也需要同時考慮受詞。意見的強度由動作決定，但意見傾向是由兩個詞素共同決定，沒有負面詞素時是正面詞，有一個負面詞素時是負面詞，兩個都是負面詞素時，因爲它們有動詞與受詞的關係，因此會產生負負得正的效果，此時意見傾向設定爲正面詞。例如在「避暑」一詞中，「避」與「暑」兩者都是負面詞素. 它的意見強度取決於「避」但意見傾向同時取決於「避」與「暑」，文章中提到「避開」的時候，大部份都是負面的文意，但在這個例子中，避開「暑氣」這種不好的東西，卻變成是正面的了。類似的例子還有「抗病」。

$$\begin{aligned}&\text{if } (S(C_1)\neq 0 \text{ and } S(C_2)\neq 0)\\&\quad \text{then } S(C_1C_2)=\left|S(C_1)\right|\times SIGN(S(C_1))\times SIGN(S(C_2))\\&\quad \text{else } S(C_1C_2)=S(C_1)+S(C_2)\end{aligned} \tag{8}$$

動補關係：這類詞彙的計分方式與主謂詞相同。扮演補語角色的第二個詞素決定了這類詞的意見分數。例如，「提高」是由「提」與「高」所組成。補語的詞素「高」描述了動作語素「提」的最後狀態，由此可見動補詞整體詞彙的意見強度與意見傾向皆取決於第二個補語詞素。

以上五種是語言學家所定義的構詞分類，除了這五種分類之外，我們特別爲意見分析增加了兩類：否定關係詞與肯定關係詞。判斷此二類詞所需的否定詞素與肯定詞素皆可由查表得知。

否定關係：詞中的否定詞素會將另一詞素之意見反轉。例如「不悅」一詞是由否定詞素「不」與另一詞素「悅」組成，該詞的意見強度由非否定詞素的「悅」所決定，但詞中「悅」的正面傾向被「不」所反轉，因此該詞爲一負面意見詞。

$$\text{if } (C_1 \in NC) \text{ then } S(C_1C_2) = (-1)\times S(C_2) \text{ else } S(C_1C_2) = (-1)\times S(C_1) \qquad (9)$$

肯定關係：詞中的肯定詞素用來確認另一詞素之意見。肯定詞素除表示確定語氣外，對該詞之意見並無影響。例如「有利」一詞是由肯定詞素「有」確認另一詞素「利」之存在。該詞的意見強度與意見傾向皆由「利」所決定。

$$\text{if } (C_1 \in PC) \text{ then } S(C_1C_2) = S(C_2) \text{ else } S(C_1C_2) = S(C_1) \qquad (10)$$

其他關係：在中文裡，雖然大部份的字都可作爲詞素，但也有一些詞的組成字本身並非詞素，整個詞才是最小的詞素，例如「蝴蝶」，如果詞內的字並非詞素，我們就無法探討它們之間的關係，因爲它們各自無法表達一個完整的概念。另外，有些詞裡的詞素包含前綴詞與後綴詞，例如「阿媽」、「蝦子」；或者疊字如「姑姑」，又或者譯名如「披薩」等，我們都無法以構詞關係來分析它們，因此我們將他們歸爲其他關係類，在不考慮構詞關係的情況下，以巨觀的演算法來計算這類詞彙的意見。最後，我們總共將詞彙分成八類。

得出詞彙的意見傾向後，進一步自然是希望能藉由它們得出句子的意見傾向。從巨觀的角度看這個問題，我們仍然不關心詞彙之間可能互相影響的變數，但在此我們考量了對意見分析很重要的否定詞。在沒有句子剖析樹的狀況下，我們假設否定詞會修飾最接近它的意見詞，而將該詞的意見傾向反轉。演算法如下示：

$$S_p = S_{opinion-holder} \times \sum_{j=1}^{n} S_{w_j} \qquad (11)$$

,

演算法：意見句擷取

1. 對每個句子 p
2. 對句中的每個意見詞 w
3. 如果有否定詞在附近出現且尚未將任何意見詞的意見傾向反轉，就將目前的意見詞 w 之意見傾向反轉
4. 利用意見詞及意見持有者的公式 (公式 11) 決定句子 p 的意見傾向

公式中 S_p 是句子之意見分數。若不考慮構詞結構的資訊，公式中的 S_w 可由公式 (5) 得知；若考慮構詞結構的資訊，公式中的 S_w 則由公式 (6) 到 (10) 決定，取決於 S_w 的類別。如果有判定意見持有者重要性的技術，可算出代表每個持有者的分數，那麼公式中的 $S_{opinion-holder}$ 就可以代入這個分數；目前我們的討論仍專注於意見本身，因此該分數暫時皆代入 1，也就是將所有意見持有者的重要性視爲相同。

接著我們再從微觀的角度來研究詞彙之間的影響。同樣地，我們依照語言學家之定義，將詞與詞之間的關係分類，並針對各類別計算出詞組的總意見，這些詞組關係的舉例如下：

並列關係：兩個詞彙在句子中扮演對等的關係。例如，在「美麗而聰慧」詞組中，「美麗」和「聰慧」就扮演對等的角色。

修飾關係：前面的詞彙修飾後面的詞彙。例如，在「淒涼地笑著」詞組中，「淒涼地」修飾「笑著」。

主謂關係：通常前面是被描述的主詞，而後面是謂語，主要表達判斷、描寫或說明性的意涵。例如「討論熱烈」。

動受關係：前面的詞彙通常是動詞，而後面的詞彙則是它的受詞。.例如「恢復疲勞」。

動補關係：前面的詞彙通常爲動詞，但有時是形容詞，而後面的詞彙則從某個角度補充說明前面動詞或形容詞的程度。例如「收拾乾淨」。

因爲這些詞組關係與構詞關係的分類概念是相同的，所以計算詞組意見的方式，就比照處理詞彙時，依照不同構詞關係所設計的公式。在微觀的演算法中，否定詞的影響也同樣必須考慮。

從詞彙與句子的說明中，我們可以類推，文章的意見傾向，也是由文章所包含的句子來決定的。有時候文章中雖然同時有正面與負面的意見，但最後經常是以支持意見較多的立場爲文章的最後立場。因此一個簡單的方法，即是將文章中所有句子的意見加總起來，做爲文章最後的總意見。

$$S_d = \sum_{j=1}^{m} S_p \quad , \qquad (12)$$

S_d 是文章的意見分數，m 是該文章的總句數，而每句的意見分數 S_p 可由公式 (11) 算出。

在文章的層次中，如果我們也要從微觀的角度去研究，就必須考慮句子之間的關係。主要有兩個因素需要加以考量：句子上下文的關係，及句子本身的重要性。例如句意的

轉折及連接詞的使用等，是考慮上下文時可以參考的線索；而句子在文章中的位置及它們可能代表的意義，例如首句可能破題，或結尾句通常爲結論，在考量文章意見的時候，這些句子也許比其他句子更爲重要，這是屬於文章寫作結構討論的範疇；另外意見是由哪些人所表達，可能也會影響它的重要性，意見領袖或專家所表達的意見經常比其他人更重要，這就又牽涉到意見持有者的代表性及社群的問題，並且需要配合意見持有者擷取的技術才能解決。這些都是未來技術發展可能的方向。

實驗結果

從巨觀的角度，如果我們使用 NTCIR-6 的語料，在句子的層次我們可以達到 f-score 0.761 的效能。在 NTCIR-7 的語料上，我們同時考慮利用結構資訊以及未利用結構資訊的演算法，實驗結果顯示，利用結構資訊的演算法，效能是 f-score 0.6717，比未利用結構資訊的演算法之效能 (f-score 0.6635) 只提升了約一個百分點，效能增進並不多。但若改用 Opinion Treebank (一個在中文 Treebank 5.1[10]上標記意見的語料) 做爲實驗語料，則效能可從 f-score 0.72 增加到 f-score 0.80，增加了八個百分點。與 NTCIR-7 的參加隊伍相比較，我們的意見分析演算法效能不論在意見擷取或是意見傾向判斷都在前三名之列。

若是退回詞彙層次來探討，我們會發現，考慮構詞結構來決定詞彙的意見傾向時，效能可達到 f-score 0.62，比原先的 f-score 0.57 要高了五個百分點，效能上的增益較明顯。再進一步探究，我們更發現，在語言學家定義的五類構詞方式之中，考慮結構對於第二類及第四類的辭彙在決定意見傾向時較有助益。原因很明顯：這兩類都是具有類似於「修飾關係」的詞彙，只是在修飾關係詞中，前面的詞素會修飾後面的詞素，但在動補關係詞中，卻是後面的詞素修飾前面的詞素：不管在哪一種情況下，考慮修飾詞素再決定最後的意見，都是有幫助的。

分析詞彙層次的結果後，我們回頭來看句子層次的問題：重點是需找到互相修飾的詞彙。在句子中，一是結構較爲複雜，要找到修飾的詞組 (f-score 0.52) 比區分詞彙的構詞結構 (accuracy 0.70) 不易；另一是當詞彙的意見無法百分之百被正確判定時，它的錯誤也會影響到句子層次的判斷，尤其若是依照不同的詞組關係來計算句子的意見，這些錯誤的影響會更明顯。在句子的層次該如何利用詞組的關係來增進效能，仍是一個值得研究的課題。

前面曾經提過，意見分析牽涉到的技術相當多，包括相關性、意見持有者及意見目標的擷取等等。但這些技術與意見分析的技術可以分開來探討，因此我們在這裡不多贅述，僅提供參考資料：在使用 NTCIR-7 語料做爲實驗材料的狀況下，相關句判別效能可達 f-score 0.93，意見持有者標記可達 f-score 0.82，意見目標標記可達 f-score 0.63 (Seki *et al.*, 2007)。

[10] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2005T01U01

意見分析技術之相關應用

在發展意見探勘的演算法，進而能夠找到意見資訊之後，我們提出了數個可能的應用，包括意見摘要、意見追蹤、意見問答及利用意見進行的關聯探索，意見摘要與意見追蹤提供了整理意見，並將意見呈現給使用者的不同方式。意見摘要提供一個以文字及量化數據表示的整體意見，而意見追蹤則提供正負面意見評論隨時間消長的趨勢圖。意見摘要的文字範例如下表：

意見傾向	內容
正面	上述建議來自四名科學家所組成的專家小組，該小組於一月應英國政府之邀成立，就複製所衍生的法律與倫理問題提出相關建議。
負面	在複製羊成功的消息宣布之後，美國總統柯林頓及「生物倫理顧問委員會」斥複製人不道德，柯林頓禁止使用聯邦經費從事複製人類的實驗，並要求民間自我克制不作這種研究。

表 1: NTCIR 語料中關於動物複製的意見摘要範例

意見傾向	內容
正面	而複製技術如果成熟，它將會是一種強大有用的工具，任何工具都可能被善用或誤用，評價一個工具不能只拿它被誤用的情境去批評它，因而禁制了它被善用的原始目的與機會，妥善的立法規範管理似乎才是較理性的作爲。
負面	有人反對複製人，因爲違反了上帝的旨意。

表 2: BLOG 語料中關於動物複製的意見摘要範例

意見趨勢圖的範例如下圖：

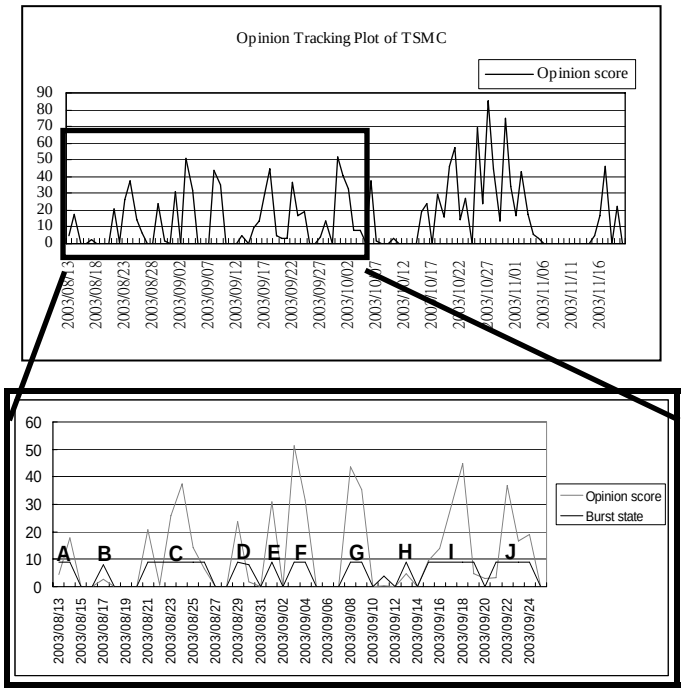

圖1: 台積電之正面意見趨勢圖/意見爆量偵測

區間	從	至	摘要
A	2003/8/13	2003/8/15	晶圓雙雄第四季營收不看淡
B	2003/8/18	2003/8/18	透過摩台指套利避險最近半年外資現貨期貨兩面賺
C	2003/8/20	2003/8/26	那斯達克創 16 個月新高，產業景氣有利中多格局
D	2003/8/29	2003/8/30	IDM 廠產能預估偏低類比 IC 急單湧至晶圓雙雄低階產能接單量逾 10%
E	2003/9/1	2003/9/1	晶圓雙雄外資持續大買
F	2003/9/3	2003/9/4	專業分工台積電、日月光創造雙贏
G	2003/9/8	2003/9/9	特許獲 0.18 微米以下通訊晶片急單 Q3 產能利用率逾 60%
H	2003/9/13	2003/9/13	IBM 晶圓代工再下一城取得 Intersil 電源 IC 訂單
I	2003/9/15	2003/9/19	通訊、消費性電子旺季訂單挹注，晶圓雙雄 Q4 產能利用率雙挑戰 95%
J	2003/9/21	2003/9/24	張忠謀：半導體明年景氣佳殺手級應用助威

表 3: 根據圖1意見爆量偵測對應區間的意見得出的摘要

意見問答技術更進一步利用探勘而得的意見作爲背景知識，有別於傳統常識型的問答系統，它具有回答與意見相關的問題的能力。我們將意見問題分成六類，也提出處理意見問答的前置步驟 (Ku *et al.*, 2008)。六類的意見問題列舉如下：

(1) **意見持有者 (HD)**

定義: 詢問該意見的發表者。

範例問題: 誰支持國民卡的發行？

範例回答: 人名/組織名/可持有立場的實體。

(2) **意見目標 (TG)**

定義: 詢問意見是針對何目標發表的。

範例問題: 一般認爲誰該爲空難負責？

範例回答: 可被評論的實體。

(3) **意見立場 (AT)**

定義: 詢問意見持有者對於意見目標的立場爲何。

範例問題: 輿論對於美國總統柯林頓的緋聞看法如何？

範例回答: 條列出與問題相關的意見，區分爲正面、中立、負面三類。

(4) **持該意見的原因 (RS)**

定義: 詢問特定或非特定意見持有者對於特定目標持該意見的原因。

範例問題: 爲何人們認爲不要有聯考比較好？

範例回答: 條列出持該特定立場的理由。

(5) **主流意見 (MJ)**

定義: 詢問哪一個意見是主流意見。

範例問題: 如果政府發行國民卡，它的聲望會上升或下跌？

範例回答: 主流意見及其原因。

(6) **是否類意見問題 (YN)**

定義: 是或否選擇題。詢問關於該意見的陳述是否正確。

範例問題: 空難的發生是否肇因於管理問題？

範例回答: 若較多人認同的想法符合陳述則答「是」，反之則答「否」。

利用意見進行的關連探索更能跳脫一般對於「具有某種關係」的定義，直接找出互相之間具有影響力的配對。以下各家公司的意見趨勢圖爲例：

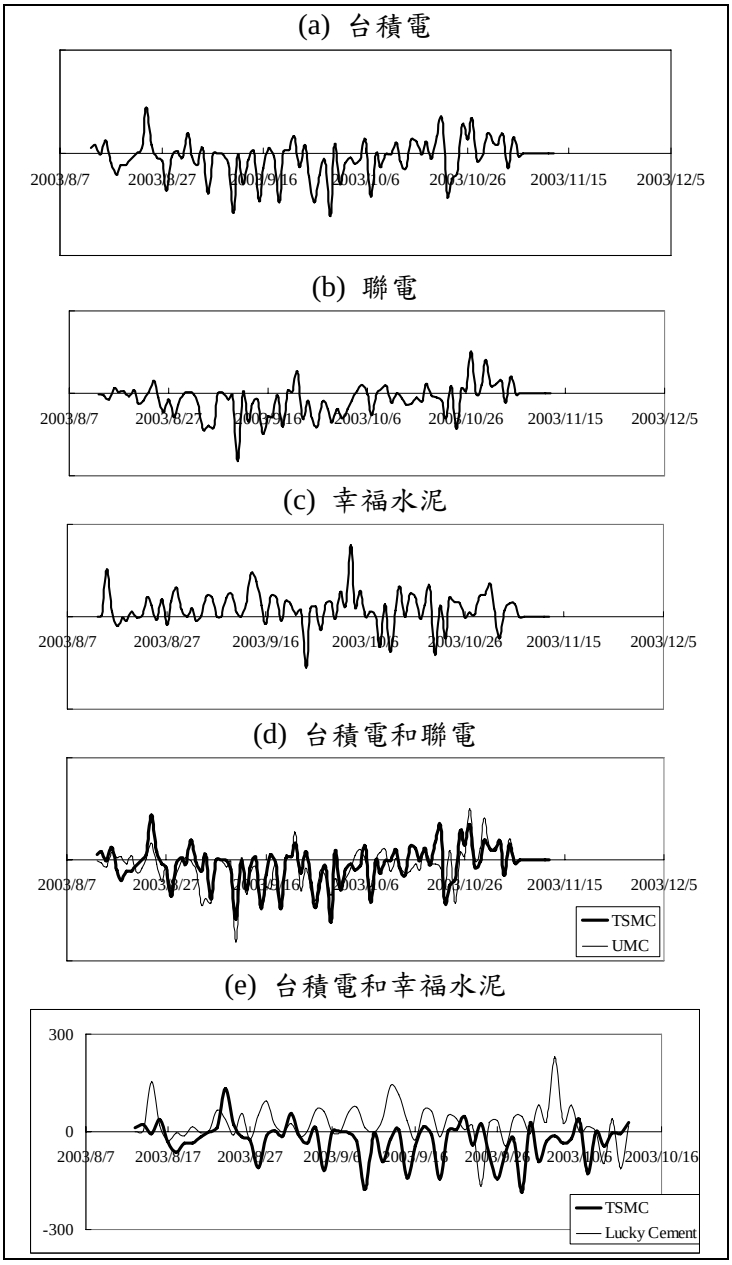

圖2: 台積電、聯電、幸福水泥之意見趨勢圖

圖2的五個小圖中，(a)、(b)和(c)圖分別是台積電、聯電及幸福水泥之意見趨勢圖。(d)圖疊合了台積電和聯電的意見趨勢圖，而(e)圖則疊合了台積電和幸福水泥的意見趨勢圖。我們假設，造成意見趨勢的背後有一連串的事件，而這些事件將對相關的實體造成類似的影響，並引起類似的評論。反過來說，在一段夠長的時間區間裡，若是兩個實體的意見趨勢圖是相似的，那麼它們很有可能具有某種關係：我們可由相似的意見趨勢圖找到相關的實體。從圖2的實例可看出，這個想法或許是可行的：在(d)中可看出台積電與聯電的意見趨勢圖相當相似，而在(e)中亦可看出台積電與幸福水泥的意見趨勢圖並不

相似，這個結果與現實狀況吻合。因此我們相信，利用意見趨勢圖來找尋相關的實體是可行的。我們的實驗結果也說明了這個做法的確可以找出相關的實體，並與傳統利用共現率 (collocation) 的概念所找出的實體不相同(Ku *et al.*, 2009)。

從我們的實驗中得知，無論是演算法或是我們提出的應用，都有令人滿意的效能。我們據此發展出一個中文意見分析系統 – CopeOpi ，是第一個中文的意見分析系統，使用者可在此介面輸入想查詢的主題，並選擇分析的時間區間與資料來源。CopeOpi 將提供使用者關於該查詢主題的意見趨勢圖，從這個趨勢圖，使用者可看到每日的總意見傾向，如有需要，還可再進一步看到當日正面與負面相關文章的標題列表，這可視爲意見摘要的一種呈現。經由列表，使用者可依意見分類閱覽每篇文章的內容。即使與國外的意見分析系統比較，我們的系統仍然提供了相當多的功能。有用的資訊與有效率的資料整理方式，使 CopeOpi 成爲目前最先進的意見分析系統之一。

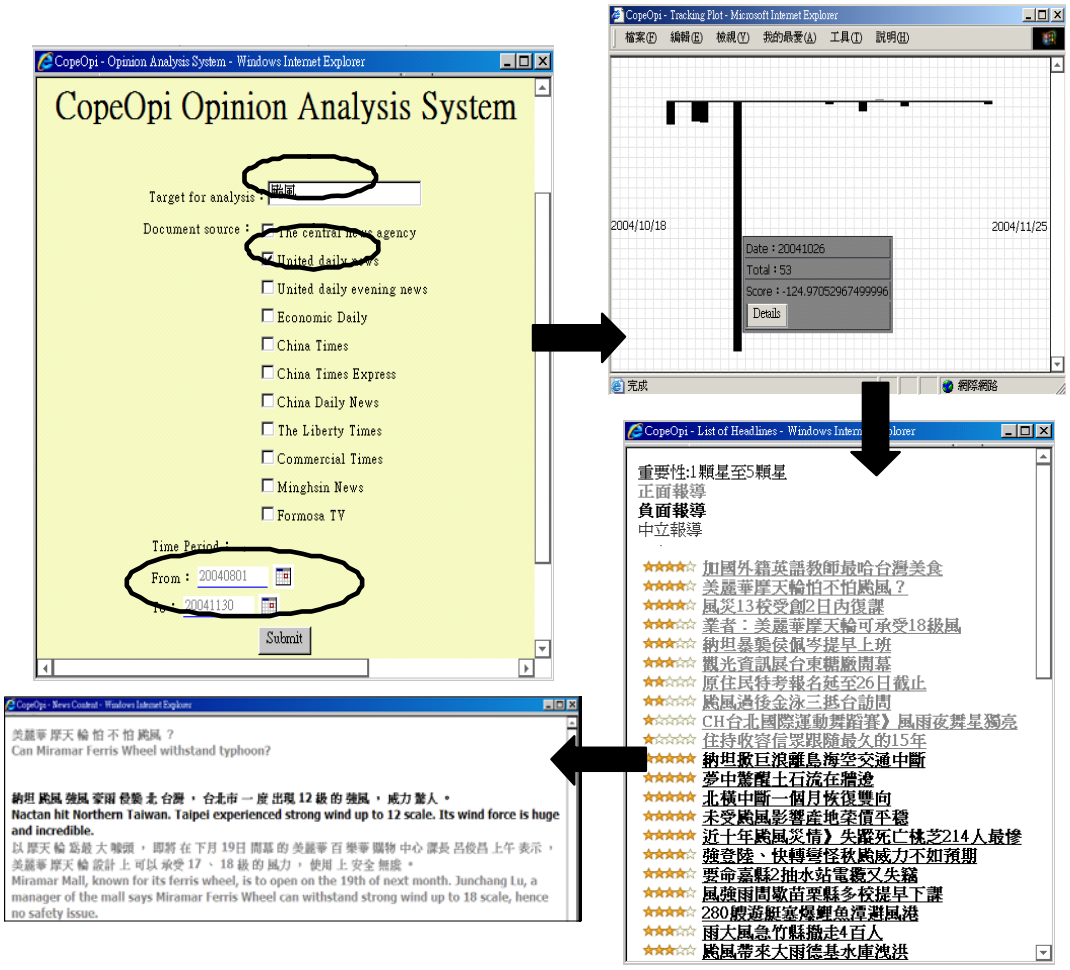

圖3: CopeOpi 系統例示

結語

本文介紹了意見分析這個受重視的研究議題目前的研究概況，並以中文語言學知識爲本，針對中文語料提出了意見分析的演算法。我們所提出的中文意見分析演算法，可達到一定的效能，但亦仍有改進的空間。

我們提出數個該技術的應用領域及應用實例，從這些實例可看出，意見分析的技術可以應用的範圍很廣。目前業界也對這個技術的應用深感興趣，我們相信，意見分析的技術有機會發展出廣爲使用的系統。

前文提到，我們對於不同國家及文化背景的人所發表的意見亦相當感興趣，因此雖然我們已爲中文提出意見分析的技術，未來我們也計劃將實驗的範圍由單語擴展到多語，並發展多語的意見分析技術。我們已從英文與日文兩個主要的國際語言切入，引用 NTCIR-7 評比會議所提供的語料，討論結構資訊在不同語言上的表現，以及自動翻譯效能在意見分析議題上的影響。未來我們希望能夠不受限地在各種文本上自動抽取並整理出有用的意見資訊，以發展出更多應用。

參考文獻

Bai, X., Padman, R. and Airoldi, E. (2005). On learning parsimonious models for extracting consumer opinions. *Proceedings of the 38th Annual Hawaii International Conference on System Sciences*, Track 3, Volume 03, page 75.2.

Branavan, S. R. K., Chen, H., Eisenstein, J., and Barzilay, R. (2008). Learning document-level semantic properties from free-text annotations. *Proceedings of the Association for Computational Linguistics (ACL)*, .

Breck, E., Choi, Y. and Cardie, C. (2007). Identifying Expressions of Opinion in Context. *Proceedings of the 20th International Joint Conferences on Artificial Intelligence*, pages 2683-2688. Hyderabad, India.

Cardie, Claire, Farina, Cynthia, Bruce, Thomas, and Wagner, Erica. (2006). Using natural language processing to improve eRulemaking. *Proceedings of Digital Government Research* (dg.o).

Chen, Y. and Xie, J. (2008). Online consumer review:Word-of-mouth as a new element of marketingcommunication mix. *Management Science*, 54(3):477–491.

Choi, Y., Cardie, C., Riloff, E. and Patwardhan, S. (2005). Identifying sources of opinions with conditional random fields and extraction patterns. *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 355-362. Association for Computational Linguistics .

Ghose, A., Ipeirotis, P. and Sundararajan, A. (2007). Opinion Mining using Econometrics: A Case Study on Reputation Systems. *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 416-423. Association of Computational Linguistics.

Hu, M. and Liu, B. (2004). Mining Opinion Features in Customer Reviews. *Proceedings of the 19th National Conference on Artificial Intelligence*, pages 755-760.

Jin, Xin, Li, Ying, Mah, Teresa, and Tong, Jie. (2007). Sensitive webpage classification for content advertising. *Proceedings of the International Workshop on Data Mining and Audience Intelligence for Advertising*.

Kim, S.-M. and Hovy, E. (2004). Determining the sentiment of opinions. *Proceedings of the 20th International Conference on Computational Linguistics*, pages 1367-1373.

Ku, L.-W. and Chen H.-H. (2007). Mining Opinions from the Web: Beyond Relevance Retrieval. *Journal of American Society for Information Science and Technology, Special Issue on Mining Web Resources for Enhancing Information Retrieval*, 58(12), 1838-1850.

Ku, L.-W., Ho, X.-W. and Chen, H.-H. (2009). Opinion Mining and Relationship Discovery Using CopeOpi Opinion Analysis System. *Journal of American Society for Information Science and Technology.* 60(7), 1486-1503.

Ku, L.-W., Lee, L.-Y., Wu, T.-H. and Chen., H.-H. (2005). Major topic detection and its application to opinion summarization. *Proceedings of 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 627-628.

Ku, L.-W., Liang, Y.-T. and Chen, H.-H. (2008). Question Analysis and Answer Passage Retrieval for Opinion Question Answering Systems. *International Journal of Computational Linguistics and Chinese Language Processing*, 13(3), 307-326.

Ku, L.-W., Lo, Y.-S. and Chen, H.-H. (2007). Test Collection Selection and Gold Standard Generation for a Multiply-Annotated Opinion Corpus. *Proceedings of 45th Annual Meeting of Association for Computational Linguistics*, short paper, June 23-30, 2007, Prague, Czech Republic, 89-92.

Lu, Bin, Tsou, Benjamin K. and Kwong and Yee, Oi. (2008). Supervised Approaches and Ensemble Techniques for Chinese Opinion Analysis at NTCIR-7. Proceedings of the 7th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering, and Cross-Lingual Information Access, December 16-19, 2008, Tokyo, Japan. 218-225.

Martin, Lanny W. and Vanberg, Georg. (2008). A robust transformation procedure for interpreting political text. *Political Analysis*, 16(1):93–100.

Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, pages 79-86.

Seki, Y., Evans, D. K., Ku, L.-W., Sun, L., Chen, H.-H. and Kando, N. (2008). Overview of Multilingual Opinion Analysis Task at NTCIR-7. *Proceedings of the 7th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering, and Cross-Lingual Information Access*, December 16-19, 2008, Tokyo, Japan. 185-203.

Wiebe, J., Wilson, T., and Bell, M. (2001). Identify collocations for recognizing opinions. *Proceedings of ACL/EACL2001 Work*