

國立交通大學

資訊科學與工程研究所

碩士論文



意見探勘在中文電影評論之應用

Applying Opinion Mining to Chinese Movie Reviews

研究生：邱鴻達

指導教授：梁婷 教授

中華民國 一百年七月

意見探勘在中文電影評論之應用  
Applying Opinion Mining to Chinese Movie Reviews

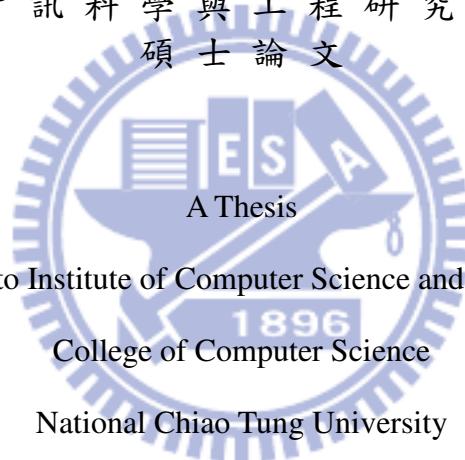
研究 生：邱鴻達

Student : Hong-Da Ciou

指導教授：梁 婷 博士

Advisor : Dr. Tyne Liang

國立交通大學  
資訊科學與工程研究所  
碩士論文



Submitted to Institute of Computer Science and Engineering

College of Computer Science

National Chiao Tung University

in partial Fulfillment of the Requirements

for the Degree of

Master

in

Computer Science

July 2011

Hsinchu, Taiwan, Republic of China

中華民國一百年七月

# 意見探勘在中文電影評論之應用

研究生：邱鴻達

指導教授：梁 婷 博士

國立交通大學

資訊科學與工程研究所

## 摘要

隨著 Web2.0 網路蓬勃發展，使用者的意見不再僅限於口耳相傳，藉由不同的網路平台，例如 mobile01、博客來、Yahoo! 奇摩等網站，使用者意見成為一股無法忽視的力量。有鑑於此，我們希望研究如何將數篇不同的電影評論整合並計算可信的分數提供使用者參考。在本篇論文中，我們實作出一個有效的電影評價系統，其中包括語料處理、屬性詞人工擷取及分類、意見詞擷取、意見詞分數計算和電影評分。首先我們透過手動收集屬性詞，再利用同義詞詞林作擴充取得較完整的電影屬性詞集。接著我們提出一個以詞性組合序列為基礎的方法擴充意見詞集。取得屬性詞及意見詞後，我們將屬性詞及對應的意見詞配對。與其他研究不同的是，我們考慮到有些只有意見詞卻缺少屬性詞的句子，因此我們利用五種特徵在支持向量機上辨識意見詞所屬的類別。最後我們提出一個考慮電影的四大屬性類別的評分，對 226 部電影做權重式評分實驗。實驗結果顯示最高的 F-score 為 83%，整體正確率為 79%。

# Applying Opinion Mining to Chinese Movie Reviews

Student : Hong-Da Ciou

Advisor : Dr. Tyne Liang

Institute of Computer Science and Engineering  
National Chiao Tung University

## ABSTRACT

In the age of Web2.0, more and more opinion platforms are developed on internet, such as Mobile01, Yahoo, Amazon etc. And more and more people express their opinions through internet platforms. It is important for industries and internet users to access huge amount of opinions quickly. In this thesis, a movie evaluation system of Chinese review corpus is developed by using opinion mining techniques. The system contains corpus processing, attribute word acquisition and classification, opinion word acquisition and opinion word score calculation, movie rating evaluation. The attribute words are manually acquire and expanded by Tongyici Cilins. The opinion words are expanded by part of speech model. For those opinion words with no corresponding attribute words, we use SVM to classify them into four categories of movie attributes. Finally, a movie scoring function is proposed to evaluate ratings of movies according to the categories of movie attributes. We evaluated the performance with a review set of 226 movies. The result shows that the best F-score is 83%, and precision is 79%.

## 誌謝

終於在這段研究所日子中完成碩士論文，首先要感謝的是我的指導教授 梁婷老師，在這兩年的碩士生活中，教導我很多做研究、處理事情的方法，對於我的碩士論文給許多建議及想法，讓我可以順利地完成這一篇碩士論文，老師也常常說一些生活經驗、對人處事的方法，在面對以後出社會有相當的助益。梁婷老師把我們當作自己的兒女一樣關心，時常問我們身體健康、生活起居...等，讓實驗室充滿到一股親切感。接著要感謝我的口試委員，鄭卜壬教授、楊武教授，感謝給予我許多寶貴建議，讓本論文可以更加完善。

此外我要感謝冠熙學長，他花了很多時間幫我修改論文紙本，在做碩論的時間，只要有問題也都會去請教他，但他都不厭其煩地教導著我，是一位非常好的學長。打程式有時候卡到，都會詢問家棋學長，他總是可以幫我一針見血地除掉問題，在此非常感謝你！還有陪伴研究所生活的奕賢、笙權夥伴，幸虧有你們生活才不至於苦悶，一起修課、打球、練嘴上功夫，當然還有課業上的交流、協助，真的很感謝你們一路互相幫忙、分享心情，讓我有動力可以完成這一篇碩士論文。

最後要感謝家人，無怨無悔地協助我完成學業，當我疲倦時，家總是一個可以充電的庇護站，讓我一直有動力去完成這碩士論文。以及在低潮時，陪我聊天解悶的朋友乳液，謝謝一路研究所生活相伴的你們。讓我面對未來也更有勇氣，努力往自己的目標前進，不會辜負你們的期望。面對未來一起加油!!

鴻達

新竹 2011/07/25

# 目錄

摘要 .....	i
ABSTRACT.....	ii
誌謝 .....	iii
目錄 .....	iv
表目錄 .....	v
圖目錄 .....	vii
第一章 緒論 .....	1
1.1 研究目的與動機.....	1
1.2 論文架構.....	2
第二章 意見探勘相關研究 .....	3
2.1 屬性詞擷取.....	3
2.2 意見詞擷取.....	4
2.3 意見極性判斷.....	5
2.4 意見詞與屬性詞的關連.....	6
2.5 意見探勘相關研究比較.....	7
第三章 研究方法 .....	9
3.1 語料處理.....	9
3.2 屬性詞人工擷取及分類.....	11
3.3 意見詞擷取.....	14
3.3.1 標記意見詞.....	14
3.3.2 意見詞詞性組合.....	16
3.3.3 建立意見詞詞性組合序列.....	19
3.4 意見詞與屬性詞類別配對.....	22
3.4.1 屬性詞與意見詞配對辨識.....	23
3.4.2 分類特徵.....	24
3.4.3 分類實驗.....	25
3.4.4 特徵選取.....	26
3.5 意見評分 .....	27
3.5.1 意見詞計分 .....	27
3.5.2 評論評分 .....	29
第四章 電影評分 .....	35
第五章 結論 .....	38
參考文獻 .....	39
附錄 .....	42

# 表目錄

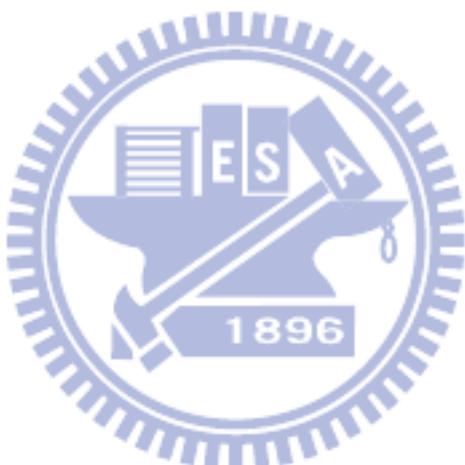
表 1:五種片語模型 (摘自[Turney, 2002]).....	4
表 2:意見詞語法規則 (摘自[Popescu and Etzioni, 2005]) .....	4
表 3:四種語法模型(摘自[Zhuang et al. , 2006]) .....	5
表 4:英文意見探勘研究比較 .....	7
表 5:中文意見探勘研究比較 .....	8
表 6:評論字數分布 .....	11
表 7:各屬性詞類別列舉 .....	13
表 8:四種屬性詞類別統計 .....	13
表 9:意見詞極性統計 .....	15
表 10:意見詞對應屬性詞類別統計 .....	16
表 11:標記意見詞詞性統計.....	16
表 12:前四大類詞性組合 .....	17
表 13:其餘詞性組合 .....	18
表 14:頻率較少的詞性組合 .....	19
表 15:意見詞雙詞統計結果 .....	19
表 16:五種雙詞詞性序列 .....	20
表 17:差異度門檻值調動 .....	21
表 18:篩選意見詞彙 .....	22
表 19:屬性詞與意見詞的配對統計 .....	23
表 20:意見詞與屬性詞的配對實驗 .....	25
表 21:分類結果第一回合 .....	26
表 22:分類結果第二回合 .....	26
表 23:分類結果第三回合 .....	27
表 24:意見詞情緒極性實驗 .....	28
表 25:意見詞分數舉例 .....	29
表 26:修飾意見詞副詞 .....	30
表 27:否定詞列表 .....	30
表 28:轉折語列表 .....	31
表 29:評論評分實驗 .....	32
表 30:評論五級分評分實驗 .....	33
表 31:評論五級分評分實驗比較 .....	33
表 32:評論五級分差距統計 .....	33
表 33:電影評分實驗 .....	35
表 34:電影評分實驗 .....	36
表 35:“新娘大作戰”評價 .....	36

表 36:“阿凡達”評價 .....	37
表 37:整體類別在標記語料中次數統計 .....	42
表 38:劇情類別在標記語料中次數統計 .....	43
表 39:特效類別在標記語料中次數統計 .....	45
表 40:電影相關人物類別在標記語料中次數統計 .....	46
表 41:擴充的意見詞 .....	47



## 圖目錄

圖 1:流程圖 .....	9
圖 2:奇摩電影首頁 .....	10
圖 3:奇摩電影評論 .....	10
圖 4:維基百科演員暱稱 .....	12
圖 5:同義詞詞林資料庫 .....	14
圖 6:標記意見詞資訊 .....	14



# 第一章 緒論

## 1.1 研究目的與動機

在 Web2.0 網路發達的時代，人們在日常生活中常常接觸到網路，並且使用網頁來搜尋自己所需要的資訊。當使用者需要去購買某一樣商品時，通常會想去網路上查看此商品的評價，當作購買商品的參考。有些人們購買完產品，使用後也會想在網路上發表心得，評論這商品使用起來的滿意程度，而網路也發展很多這種平台，讓消費者去評價產品、事物。例如: mobile01<sup>1</sup>、博客來<sup>2</sup>...等。所以在這網路資訊流行時代，網路上的意見跟人們生活是密不可分，也常重視這些網路上的評論，當作自己選擇商品、事物的參考依據。

在本篇論文中，我們應用意見探勘技術在電影評論的意見分析，讓使用者可以便利明確地收集所需要的資訊。在電影這領域中網路上的評價更是重要，因為對電影而言，沒有試用及退貨的服務，人們只能從網路上提供的預告短片當作參考，購票進場就只能觀賞兩個小時的電影，所以人們更是注重電影在網路上的評價。我們建立一個在電影評論上的資訊處理系統，包括語料收集、屬性詞擷取、意見詞擷取、意見詞與屬性詞類別分類和電影評分。我們收集奇摩網站<sup>3</sup>上的電影評論，在這獨特的電影語料中，有電影的專有名詞及新詞、評論內中英文夾雜、評論者的文學用詞素養和意見詞的極性判斷問題，這些都是處理過程中需要考慮的問題。首先擷取屬性詞與意見詞，並把屬性詞分成四種類別，期望使用者可以直接地去關注自己所需要的資訊，如使用者比較關注電影劇情、特效、演員方面，就可以直接了解這些部份的評價，而不必去查看那些細部的意見。系統可以對每部電影收集的評論產生電影評分，提供給使用者選擇電影的參考，讓使用者選擇評價較好的電影。

意見探勘近年來有許多文獻研究發表，例如: Special Interest Group on Information (SIGIR) 、 Knowledge Discovery and Date mining (KDD)...等，這些會議都有專注此領域的文獻發表，在這領域也有一些相關競賽，例如:NTCIR<sup>4</sup>、TREC<sup>5</sup>...等，這些都代表意見探勘是一個熱門議題。電影領域相較其它評論領域而言，有其獨特性的，有些常用的電影意見詞，在其他領域卻不常出現，例如：“好看”在一般的情緒字典內尋找不到，如 NTUSD<sup>6</sup>情緒字典，所以電影所使用的意

<sup>1</sup> <http://www.mobile01.com/>

<sup>2</sup> <http://www.books.com.tw/>

<sup>3</sup> <http://tw.movie.yahoo.com/>

<sup>4</sup> <http://research.nii.ac.jp/ntcir/index-en.html>

<sup>5</sup> <http://trec.nist.gov/>

<sup>6</sup> <http://nlg18.csie.ntu.edu.tw:8080/opinion/pub1.html>

見詞也比較獨特性。在電影意見詞的情緒極性，也會跟普遍其他領域不一樣，例如：“刺激”，在電影領域是比較偏於正向情緒，在其他領域則是負向情緒。在我們使用的電影語料，是來自一個開放性平台，並不是專家寫的電影評論，所以常有斷句的問題及不正常的句子結構，這些都是需要去考量的因素。

人們去查詢電影評論時，主要還是關心自己在乎的部分，例如：劇情、特效、大明星...等。所以在本篇論文中，我們將電影的屬性詞分為四種類別包括整體、劇情、特效、電影相關人物，讓使用者可以直接了解某種類的評價，而不必分析細部的屬性詞。

本篇論文所研究的電影語料是來自奇摩電影評論<sup>7</sup>，是一個自由的評論平台，人們都可以發表對某部電影的評論。我們觀察這些評論語料之後，發現意見詞所處的小句內找不到對應的屬性詞，也就是隱性屬性詞的情況比例將近四成，這也是這評論語料的問題，我們觀察先前一些論文做法[Jakeob et al., 2010]，把這些意見詞對應到固定的預設值，在這裡而言是對應到電影“整體”，但我們期望從這些沒有屬性詞配對的意見詞，不要直接讓它對應到預設值。本篇論文中我們利用一些特徵條件來辨識所屬的類別，把這些意見詞對應到屬性詞類別的四類中的其中一類。

## 1.2 論文架構

在本論文中，我們從奇摩電影的語料中人工收集電影相關的屬性詞，並且使用同義詞詞林加以擴充，最後以人工方式去把這些屬性詞分為四類。利用標記過意見詞，建立意見詞的詞性組合序列，建立意見詞詞彙庫。最後我們利用評論中的句型結構及特徵將意見詞對應所屬的屬性類別，產生電影各個類別的評價。

本論文章節架構包括，第二章為意見探勘相關研究；第三章介紹系統架構與研究方法。我們首先介紹這些評論語料的來源，利用這些語料收集電影相關的屬性詞並且分類，使用標記的訓練語料，來建立詞性組合序列產生電影意見詞，最後收集意見詞的周圍特徵、條件，來對意見詞自動配對所屬的屬性詞類別。在第四章中我們對 226 部電影做權重式電影評分，比較系統產生的評分與網頁上的電影評價，並且對電影四種屬性類別來產生分數，第五章敘述結論和未來的發展方向。

---

<sup>7</sup> <http://tw.movie.yahoo.com/>

## 第二章 意見探勘相關研究

在意見探勘這領域中，有許多議題可以研究，例如：評論目標、持有者辨識、意見摘要、極性判斷、自動評分...等，每一個議題含有不同處理步驟去達成研究目的。我們按照這篇論文架構，把相關研究分為屬性詞擷取、意見詞擷取、極性判斷、意見詞與屬性詞的關連。

### 2.1 屬性詞擷取

屬性詞擷取有分為非監督式及監督式兩種，監督式方法需要人工去標記訊息，產生訓練語料進而擷取屬性詞，但因為需要大量人工標記，不同領域的屬性詞也需要重新地訓練，所以比較耗時間，但正確率較高。非監督式方式不需要大量的人工標記，也可以利用在不同的領域，但相較監督式而言，擷取正確率較低。

在非監督式屬性詞擷取研究中，Hu et al. [2004]觀察電子產品評論的詞性標記，發現常被當作屬性詞的詞性是名詞，所以專注評論中的名詞及名詞片語，我們把斷好詞性的文字送入資料探勘系統(CBA)，取出支持度大於 1 的屬性詞作為高頻率屬性詞，對於那些無法擷取出來的低頻率屬性詞，則是尋找意見詞周圍的名詞或名詞片語當作屬性詞。Popescu and Etzioni [2005]把屬性詞分為顯性屬性詞、隱性屬性詞，在顯性屬性詞中又分為五類，五類為：“Properties”、“Parts”、“Feature of part”、“Related Concepts” 和 “Related Concepts Feature”，首先使用 MINIPAR 對評論做結構剖析斷詞，利用斷詞後的詞性標記，設定門檻來過濾可能性不高的名詞，最後計算屬性詞之間的 Point-wise Mutual information(PMI)，觀察每個屬性詞之間的關聯性，將有關連性的屬性詞收集作為屬性詞。

在監督式屬性詞擷取的研究中，Liu et al. [2005] 使用 NLProcessor linguistic parser 斷詞後，將評論中的屬性詞用“Feature”替代掉，因為使用者會使用相似的語法結構來呈現評論，但屬性詞又不太相同，所以利用詞性標記及替代成“Feature”來尋找語法。之後使用 Association Rule Mining 產生詞性關連規則，利用此規則來擷取屬性詞，在這方法下可以達到將近九成以上的準確率、召回率。Zhuang et al. [2006]手動標記電影特徵，去除出現次數低於總次數 1% 的屬性詞，並且把電影的屬性詞分成六類，OA (overall), ST(story), CH(character design), VP(vision effects), MS (music and sound effects), Movie-related people(director and actor) ，最後比較意見極性辨識效果在電影領域與產品領域，發現在電影領域中極性辨識效果較差，因為電影評論中常會出現談論電影的情節，但這些詞彙不代表評論者的意見極性。Zhao et al. [2009]把電影屬性詞做 Ontology，具有階層式的電影屬性詞，可以在上面屬性詞階層中計算底下階層的評價，產生上層階級的屬性詞的情緒分數。

## 2.2 意見詞擷取

意見詞擷取這部分，有些論文是直接利用已存在的情緒字典，例如:NTUSD、HowNet，而有些方法則是利用詞性組合、語法結構來擷取意見詞，但當面對比較獨特性的領域時，一般的情緒字典也比較難以涵蓋整個領域的詞彙，需要利用本身的語料來擷取意見詞。

Turney [2002]判斷連續兩個詞的詞性組合，利用這些比較含有情緒的詞性組合來建立詞性規則，提出五種片語模型，擷取其中的形容詞及動詞當作意見詞。Chaovalit and Zhou [2005]及 Ye et al. [2006]也採用 Turney 的五種模型來擷取意見詞。五種片語模型如表 1 所示。

表 1:五種片語模型 (摘自[Turney, 2002])

First word	Second word
Adjective	Noun
Adverb	Adjective
Adjective	Adjective
Noun	Adjective
Adverb	Verb

Hu and Liu [2004]觀察到意見詞與屬性詞常一起出現，擷取出屬性詞之後，利用這種特性，在屬性詞周圍擷取鄰近範圍內的形容詞當作意見詞，正確率約有六成以上。此篇文章把意見詞定義為形容詞，只擷取形容詞當作有效的意見詞，但意見詞有可能是動詞或副詞所組成，例如:well、outperform，忽略這些其它詞性的意見詞，以致於實驗結果召回率較低。

Popescu and Etzioni [2005] 使用 MINIPAR 產生具有語法結構的斷詞，利用詞性標記、句子結構及已知的屬性詞，整理出十種規則，利用這些規則來擷取意見詞，規則舉例如表 2，利用這些規則擷取出的意見詞，不再限制只能為形容詞，實驗結果與[Hu and Liu, 2004]相比，發現擷取意見詞效果比較好，這也代表擷取其它詞性的意見詞，有助於提升實驗結果的正確率。

表 2:意見詞語法規則 (摘自[Popescu and Etzioni, 2005])

Extraction Rules	Examples
If $\exists (M, NP=F) \rightarrow Po = M$	(expensive) scanner
If $\exists (S=F, P, O) \rightarrow Po = O$	Lamp has (problems)
If $\exists (S, P, O=F) \rightarrow Po = P$	I (hate) this scanner
If $\exists (S=F, P, O) \rightarrow Po = P$	Program(crashed)

Po=potential opinion, M=modifier, NP=noun phrase S=subject, P=predicate,  
O=object

在[Zhuang et al., 2006]文章中使用 Stanford Parser 斷詞，產生出含有語法結構的 dependency grammar path，過濾掉低頻率的語法模型，保留四種語法模型來擷取意見詞。如表 3 所示，NN 代表是名詞，JJ 是形容詞，RB 是副詞。

表 3:四種語法模型(摘自[Zhuang et al., 2006])

Dependency relation template	Feature word	Opinion word
NN – amod – JJ	NN	JJ
NN – nsubj – VB – dobj – NN	The first NN	The last NN
NN – nsubj – JJ	NN	JJ
VB – advmod – RB	VB	RB

Jin et al. [2009]提出 Hidden Markov Model(HMM)為基礎去擷取意見，一開始先斷詞，再人工標記屬性詞、意見詞，把這些的標記語料訓練 HMM 分類器，標記測試語料上的意見詞，實驗是與[Turney, 2002]相比，發現實驗效果較好，正確率明顯地提升 3 到 7%。

### 2.3 意見極性判斷

意見詞的極性判斷，主要分為兩種方法，一個是利用網路資源，觀察兩個詞彙的關連性，進而辨識意見詞極性。另一個是利用語料，語料是來自字典提供或是自己的評論語料，藉由這些語料的輔助來判斷極性。

在使用網路資源這部分，Turney [2002]提出以網路為主的情緒極性判斷，利用詞彙共現的頻率來推斷極性，也就是正向意見詞比較常跟正向意見詞一起出現，而負向意見詞也比較常跟負向意見詞一起出現的現象，利用改良後的 PMI 來計算此意見詞與正負向意見詞的關係，此篇中取正向意見詞為“Excellent”，負向意見詞為“Poor”。經由下面的公式 1，計算出來的 SO 數值大於零就是正向意見，小於零就是負向意見，hits 是代表搜尋回傳網頁數量，極性實驗的正確率將近七成五。

$$SO(\text{phrase}) = \log_2 \left[ \frac{\text{hits}(\text{phrase NEAR "excellent"})\text{hits}(\text{"poor"})}{\text{hits}(\text{phrase NEAR "poor"})\text{hits}(\text{"excellent"})} \right] \quad (1)$$

在利用語料這部分，Hu and Liu [2004]利用 WordNet 字典辨別意見詞的情緒極性。此論文中是首先收集三十個種子詞彙，這些種子詞彙已具有正負向極性，可觀察跟未知極性意見詞的關係。利用 WordNet 的形容詞中，含有同義詞集和反義詞集，如果未知意見詞的同義詞在種子詞集內，則標記跟種子詞相同情緒極性，並加入種子詞集中。如果是反義詞在種子詞集內，則標記跟種子詞相反的情緒極性。反覆此步驟擴充種子詞彙，但對 WordNet 中沒有的詞彙，無法辨識詞彙極性，辨別句子極性的實驗正確率為 84%。朱嫣嵐等人[2006]則是利用 Hownet 來計算語義相似度及語義相關場，語義相關場是指把兩個語意概念的交集除以聯

集，個別對正負向詞彙計算數值，產生意見詞的情緒極性，在實驗中準確率有達到 78%。

Marneffe et al. [ 2010] 使用 IMDB<sup>8</sup>的電影語料，每個評論上面有評論者本身的評分，範圍是一到十分，一分是低評價，滿分十分是高評價。收集這些評論語料之後，計算未知情緒的意見詞在個別分數的語料中出現的機率，把個別的機率跟此分數互相乘積，例如：假設意見詞只出現在分數為一、二的評論，在分數為一的評論機率是 0.4，在分數為二的評論機率是 0.6，0.4 乘上 1 加上 0.6 乘上 2 當作情緒分數，利用意見詞分數的高低來判斷極性。

在不考慮屬性詞的句子層級中，大部分方式是在評論內擷取特徵，使用分類器去分類極性。Fingal et al. [2004]利用擷取出的英文評論，切成單詞及雙詞建成字典，對照所建立的字典，產生出每個句子的特徵向量，使用分類器做極性分類的實驗。Yessenov et al. [2009]把英文評論中每個字當作特徵向量，比較只擷取常頻詞跟擷取形容詞、副詞當作特徵，進行意見極性分類的效果，實驗正確率將近七成。

## 2.4 意見詞與屬性詞的關連

[Su et al. , 2008] 此文章探討屬性詞與意見詞的關聯，對某些意見詞常只會形容某類屬性詞，例如車子而言，貴與價錢、可愛與外型。所以當遇到意見詞找不到形容的屬性詞時，可利用已分群好的意見詞，辨別意見詞對應哪一類別的屬性詞。在同一個句子中，屬性詞和意見詞的共同出現建立連結權重矩陣(link weight matrix)，利用兩部分來判別意見詞與屬性詞的相似性。第一個為Intra similarity是利用傳統方式，直接判別意見詞與屬性詞之間的相似度。第二個為Inter similarity是建立的矩陣向量，利用某屬性詞與每一群意見詞集的共同出現，產生每個屬性詞的特徵向量，在跟其他屬性詞的特徵向量用餘弦定理相乘，算出兩個屬性詞之間的相似度，利用此相似度將屬性詞分群。再利用某意見詞與每群屬性詞集的共同出現，也可以將意見詞重新分群。首先要設定分成多少群，一直遞迴計算相似度加以分群，直到雙方的分群結果沒有變動，當作最後的分群結果。利用特徵向量可以得知此群意見詞集跟那一群屬性詞集比較有關聯，就算意見詞找不到形容的屬性詞，也可以得知意見詞可能對應在哪一群的屬性詞。本論文的方法比較原本抓取鄰近屬性詞方式，多抓取百分之十三的意見詞與屬性詞配對。

Li et al. [2010]收集全部意見詞與屬性詞配對，以探討在不同意見領域上，顯著的配對會隨著領域而不一樣。使用已有的屬性詞及意見詞，定義每篇評論的權重、重要性，這些評論的權重是計算評論內含有多少個配對，評論中配對越多權重越重。配對的權重也是計算在全部評論中出現頻率，出現次數越多權重越重。在相互影響遞迴之下，在各別領域中產生權重高的配對，看是否符合這個領域的配對。

---

<sup>8</sup> <http://www.imdb.com/>

## 2.5 意見探勘相關研究比較

表4列出英文意見探勘相關研究的比較，表5列出中文意見探勘相關研究的比較。

表 4:英文意見探勘研究比較

英文語料	[Liu et al. , 2005]	[Zhuang et al. , 2006]	[Ding et al. ,2008]
應用領域	電子產品	電影	電子產品
外部支援辭典	無	WordNet	WordNet
斷詞	Pos tagger	Stanford Parser	parser
實驗語料	15 個產品評論	11 部電影 各 100 篇評論 (IMDB)	445 篇評論 (Amazon)
訓練與測試語料比	10:1	4:1	無訓練語料
意見層級	產品屬性詞彙	電影屬性詞彙	產品屬性詞彙
方法	利用關連規則探勘產生模型，來擷取特徵詞。最後比較產品之間各方面的評價，例如：相機的像素、螢幕、重量...等。	手動收集特徵詞及分類，收集意見詞並用 WordNet 擴充，最後產生意見詞和屬性詞配對的模型	屬性詞與意見詞庫比對，對意見極性權重計算，最後產生屬性意見配對列表
目標	擷取意見詞所對應的產品屬性詞	意見詞與屬性詞配對辨識	擷取意見句及極性判別
實驗結果	0.84(precision)	0.53(F-score)	0.9(F-score)

表 5: 中文意見探勘研究比較

中文語料	[Su et al. ,2008]	[Zhang et al. ,2009]	[Ning et al. ,2010]	本研究
應用領域	汽車	產品	旅遊	電影
外部支援辭典	WordNet	HowNet	無	同義詞詞林
斷詞	Pos tagger	HIT-IR	CKIP AutoTag	CKIP AutoTag
實驗語料	350 篇評論	485,522 篇評論 (Amazon)	9771 篇評論	226 部電影(Yahoo)
訓練與測試語料比	8:2	10:1	5:1	5000 篇評論:六萬四千篇評論
意見層級	汽車屬性詞彙	句子	句子	電影屬性詞彙
方法	首先比對屬性詞跟意見詞間的共同出現，來建立屬性詞集與意見詞集的叢集關係	斷詞產生句子中詞與詞的語法結構關係，對於副詞及否定詞來修改分數，最後利用一些特徵來計算句子在評論中的權重	建立詞組探索規則，手動篩選意見詞，利用分類特徵，LIBLinear 分類器分類極性及情感	建立電影屬性詞，產生詞性組合序列擷取意見詞，辨識意見詞與屬性詞類別配對，最後產生電影評分及屬性詞類別評分
目標	意見詞集分群	辨識評論意見極性	評論極性及情感分類	電影評分及屬性類別評分
正確率	0.65(precision)	0.76(accuracy)	0.63(F-score)	0.83(F-score)

### 第三章 研究方法

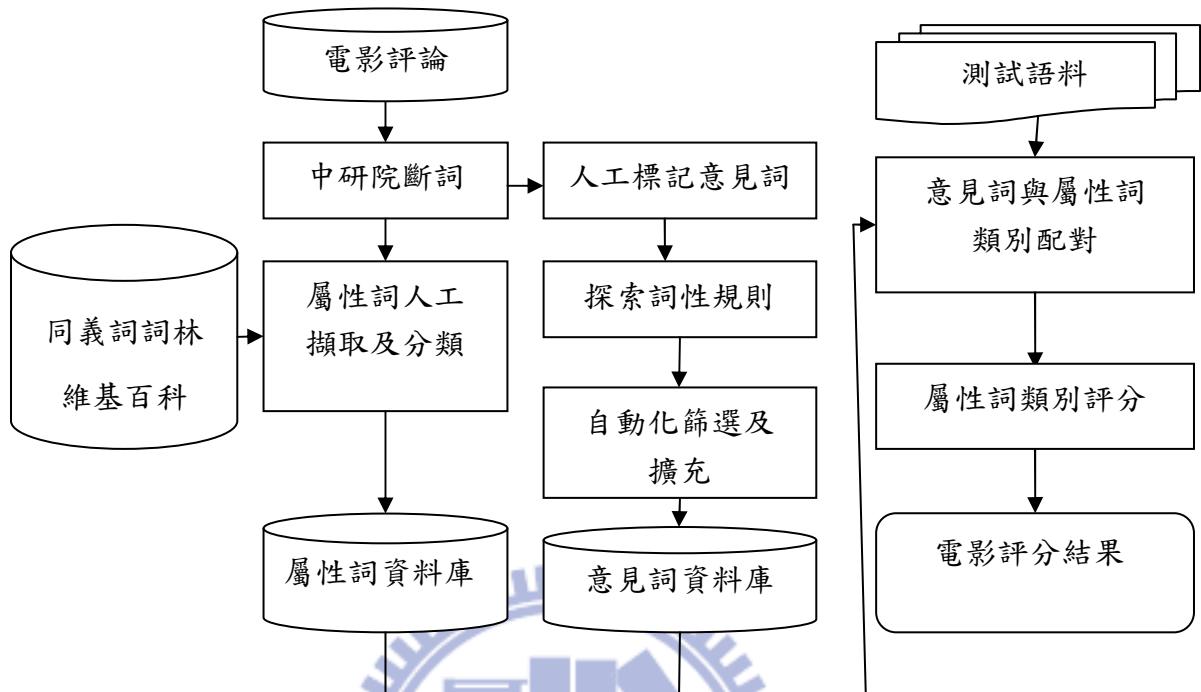


圖 1:流程圖

本論文的處理流程如圖一所示，首先我們從網路上收集中文電影評論，作為訓練及測試語料，透過自然語言處理方法，探索意見詞的詞組規則，利用這些詞性組合規則自動化意見詞擴充及篩選。在電影屬性詞部分，我們手動收集電影相關的屬性詞，再利用同義詞詞林擴充，最後把這些屬性詞人工分類。接著將意見詞與屬性詞類別配對，以便計算電影在不同屬性詞類別的評價。

#### 3.1 語料處理

我們從 Yahoo! 奇摩電影<sup>9</sup>網頁上收集電影評論，這網路平台提供電影的相關資訊，包括預告片、票房、使用者評論、使用者評分、電影類型、演員、導演...等，也是台灣網站中含有最多電影評論的網路平台，如圖二所示。

使用者觀賞完電影之後，使用者可能會來此網站打下他對於這部電影的評論，並且對這部電影來評分。奇摩電影將電影分為 1-5 顆星，1 顆星是低評價，5 顆星則是高評價。在中文電影評論而言，奇摩電影平台屬於熱門的電影網站，擁許多相關的電影評論，而想去觀賞一部剛上檔電影的人們，常會來這網路平台查看此部電影的評價，當作選擇電影的依據，使用者的電影評論特色及評分如圖三

<sup>9</sup> <http://tw.movie.yahoo.com/>

所示。



圖 2:奇摩電影首頁

評分：★★★★★ 標題：阿美~~ 好笑!!

意見詞沒有對應的屬性詞  
原來真愛就在身邊~結局雖然猜得到~但還是很感動~~so sweet..

中英文夾雜

你覺得這項短評有幫助嗎？[有][沒有][檢舉內容]

評分：★★★★☆ 標題：good !!!

空格斷句  
很好笑阿 還不錯 很少有喜劇片那麼多梗的一定要看

你覺得這項短評有幫助嗎？[有][沒有][檢舉內容]

評分：★★★★☆ 標題：

蠻多笑點的.... 電影院都是笑聲就知道有多好笑了.... 也蠻有正面意義的~~心情不好  
真的可以來看看~~哈哈

發表人：stacy\*\*\* (momo) 發表時間：2011-04-25  
發表人：m85\*\*\* (小明) 發表時間：2011-04-24  
發表人：r90\*\*\* (r90076) 發表時間：2011-04-24

重複符號

圖 3:奇摩電影評論

在奇摩電影的網頁平台中，大家都可以隨時隨地留下電影評論，比較沒有評論電影的限制，可以讓使用者自由發揮，所以含有較多的電影評論，但也因為如此這些評論比較屬於自由，口語化的寫作方式較多，而跟專家寫的專業評論不太相同。從圖三例子可發現此電影語料的特性，在標點符號這部分，有些人寫電影評論，都習慣用空格來斷句，及符號接連出現的情況，需要考量這些因素避免將電影評論斷句錯誤。處理空格的部分是用逗號來替代掉，而對符號接連出現的情

況，則縮減成一個符號。在這口語化的電影評論中，也常會出現不完整的句子結構，例如：找不到主詞的情況，如果使用語法結構樹剖析句子，也比較難整理出標準的句型，所以我們經常遇到意見詞找不到形容的屬性詞，這就是意見句中沒有主詞。這些都是評論語料的特性，也是我們要克服的問題。

在奇摩電影的電影資料庫中，總共將近有四千部電影的資訊。我們從裡面收集 226 部電影，每部電影都至少含有一百個電影評論。對於每個評論長度取一定範圍內，我們取字數 31 到 100 的評論當作訓練語料，約占全部中文評論語料 47%。統計 85636 篇電影評論，各個評論的字數統計如表 6。

我們使用中央研究院詞庫小組的中文斷詞系統<sup>10</sup>進行斷詞與詞性標記，進而對於中文語料做後續處理。我們使用標點符號「！？。」將評論分割為長句，再利用「，；。？！～. @，」將評論分割為小句。

表 6:評論字數分布

字數	1-10	11-20	21-30	31-51	51-100	101-200	201-300	301-
頻率	5940	9328	9553	16099	23892	14729	3682	2674
比率	6.9%	10.8%	11.1%	18.8%	27.9%	17.2%	4.3%	3%

### 3.2 屬性詞人工擷取及分類

電影屬性詞是一群形容電影相關的詞彙集，包括電影的整體、特色、內容...等。我們利用從奇摩電影網站上五千篇電影評論，手動標記與電影相關的屬性詞，不包含人名全部收集 411 個屬性詞。在電影演員名字及導演名字處理，因為這些演員都屬於外國人較多，在翻譯人名的辨識的研究中，李振昌等人 [1994] 使用注音音節來辨識，但辨識翻譯人名效果有限。古倫維等人 [2009] 利用出現頻率來組合斷詞後的人名，但這需要一定的訓練語料。本篇論文我們從奇摩電影網頁中自動擷取這些演員、導演名稱，經過中研院斷詞系統(CKIP)，斷詞後的每個詞彙視為人名，以影星“萊恩雷諾”為例，斷詞之後變成“萊恩 雷諾”，把萊恩和雷諾視為演員名字，評論者常會打部分的翻譯人名，在這情況下也可以尋找人名。在電影評論中，有些評論者常會用暱稱來稱呼演員，我們利用 WIKI 維基百科<sup>11</sup> 中，介紹個別演員的網頁，收集含有暱稱的演員網頁，把這些暱稱收集成為演員名字，例如：周星馳之星爺，如圖四所示。

<sup>10</sup> <http://ckipsvr.iis.sinica.edu.tw/>

<sup>11</sup> <http://zh.wikipedia.org/>

男演員	
英文名	Stephen Chow
暱稱	周星星、星爺、星仔
國籍	 香港
籍貫	浙江寧波
出生	1962年6月22日（49歲）  英屬香港
職業	演員、導演
出道日期	1982年
活躍年代	1981年 - 2008年

圖 4:維基百科演員暱稱

[Zhuang et al., 2006]文章中，電影屬性詞被分為六類，包括 overall (OA) ， screenplay (ST) , character design (CH) , vision effects (VP) , music and sound effects (MS) ， Movie-related people(director and actor)。在本篇論文中參考上述文章的屬性詞分類，及觀察中文電影評論網站<sup>12</sup>，這個網站評分項目為劇情、演員、音樂、整體，兩者對照之後把比較有關連的特徵類別合併起來，最後將電影屬性詞分類為四類，包括整體(戲、片子)、劇情(劇本、故事)、特效音效環境(3D、場景、音樂)、電影相關人物(演員、導演)等四種屬性類別，表 7 列舉電影屬性類別部分詞彙。

表 7 中斜體字是代表此屬性詞難以固定在一個屬性類別，可以同時存在兩種屬性類別。這種情況在“整體”和“劇情”類別中比較多，有些屬性詞同時可以形容兩個類別，其餘兩個特效、電影相關人物的屬性類別相對來講比較獨特，而這類的屬性詞出現數量也比較少。首先我們盡可能把屬性詞分類，如果沒有適合的屬性詞類別，則會把它分類到電影“整體”類別。如果我們遇到屬性詞可以同時存在兩個類別以上，額外在這屬性詞的前面抓取詞彙，範圍是兩個詞彙，如果有抓取到屬性詞，則重新對應新的屬性詞類別，如果沒有則對應原本的屬性詞類別。舉例：“這特效拍得很棒”，這邊屬性詞“拍”原本是屬於“電影相關人物”的類別，但額外抓取在前面的“特效”屬性詞，我們會重新指向為“特效”類別，因為屬性詞“拍”不是指向“電影相關人物”類別。

<sup>12</sup> <http://channel.pixnet.net/movie>

表 7:各屬性詞類別列舉

屬性詞類別	詞彙
電影整體	電影、影片、片子、戲、整體、總體、片、概念、手法、主線、感覺、表現、路線、題材、印象
劇情	劇本、結局、故事、情節、腳本、劇本、對白、演戲、對打、幕、打、氣勢、功夫、過招、出招
特效、音效、場景	3D、視覺、特效、顏色、色彩、音樂、歌曲、配樂、風景、動畫、運鏡、背景、景色、情景、場景、景色、佈景、環境、效果、手法、幕
電影相關人物	男主角、女主角、人物、角色、個性、導演、演員名字、導演名字、演技、武術、動作、拍出、拍到、拍

我們把屬性詞分為四種類別的目的是讓意見詞指向所評論的類別，我們使用者可以明確地知道，這部電影在四種屬性詞類別中各自評價，假設使用者比較注重“特效”這類別，就可以單獨查詢“特效”類別評價，舉例評論：這部電影的特效很好，但劇情很薄弱。使用者可能就關注“特效”這類別的好評價，而忽略劇情“類別”的壞評價。所以需要這些分類好的屬性詞類別，方便使用者查看自己關注的類別評價。

個別屬性類別統計詞性如表 8 所示，其中將近 77% 的屬性詞詞性是為名詞，20% 的屬性詞詞性為動詞，其他詞性則為 3%。我們發現電影屬性詞不全然是名詞，存在著其他詞性的屬性詞，所以不能只是單純收集名詞作為電影屬性詞。

表 8:四種屬性詞類別統計

屬性詞類別	名詞	動詞	副詞	外文	數量定詞	形容詞	屬性詞比率
整體	63	7	2	0	1	0	73 (18%)
劇情	108	55	1	1	1	1	167 (40%)
特效、音效、場景	66	8	0	5	0	0	79 (19%)
電影相關人物	78	13	0	1	0	0	92 (23%)
總合	315(77%)	83(20%)	3	7	2	1	411

我們考慮評論者會使用不同的詞彙來表達相同的屬性詞，例如：“劇本”，有些人會稱作“腳本”、“臺本”，但其實是說明同一個屬性詞。Ding et al. [2008]首先人工收集產品的屬性詞，再利用這些屬性詞的同義詞當作擴充，因此我們使用同義詞詞林[梅家駒, 2006]擴充屬性詞，我們把同義詞詞林建成資料庫如圖五，利用程式去自動抓取同義詞，我們發現有些詞彙並不適合在電影領域使用，所以透過人工檢視篩選適合電影的屬性詞。各個電影屬性詞類別的擴充數量，整體為 68 個、劇情為 237 個、特效音效為 52 個、電影人物為 49 個，總共擴充 406 個屬性詞。原本收集了 411 個屬性詞加上擴充的詞彙，我們總共收集 817 個電影屬

性詞詞彙。

The screenshot shows a table in phpMyAdmin with the following data:

	顯示為	筆記錄，開始列數:	0
	顯示為	方式及每隔	行顯示欄名
<input type="checkbox"/>		X	劇本
<input checked="" type="checkbox"/>		X	臺本
<input type="checkbox"/>		X	腳本
<input type="checkbox"/>		X	本子
<input type="checkbox"/>		X	院本
			劇本
		=	Dk25C01

圖 5:同義詞詞林資料庫

### 3.3 意見詞擷取

在電影評論中的意見詞是比較屬於獨特的領域，某些意見詞是常形容在電影領域評論上，而不常形容在其他領域的，反之亦然。以台大 NTUSD 情緒字典為例，電影評論中最常使用的“好看”、“不錯看”...等意見詞詞彙，在一般的情緒字典卻不存在。另外一方面在電影領域中，意見詞的情緒極性會與其他普遍領域的意見詞極性有所不同，以台大情緒字典而言，可怕、刺激這些詞彙中，它是列在負向意見裡面，然而在電影恐怖片中，通常是形容正向意見的情況有所不同，這些因素都是做電影評論的挑戰。由於電影評論的獨特性，我們建立一個專屬電影領域的意見詞詞彙庫。

我們首先利用人工標記意見詞的方法，探索哪些詞性較可能含有意見表達。全方面地統計這些意見詞的詞性組合，整理出雙詞詞性組合，並且利用這些雙詞詞性組合序列自動去辨識更多意見詞，經過篩選存成專屬電影的意見詞詞彙庫。

#### 3.3.1 標記意見詞

我們初步先收集兩千篇評論當作練習語料，這些評論的字數在 30 到 100 範圍中，我們人工標記意見詞的位置、範圍、情緒極性(負、正、中立)，及意見詞所形容的屬性類別，標記語料如圖六所示。

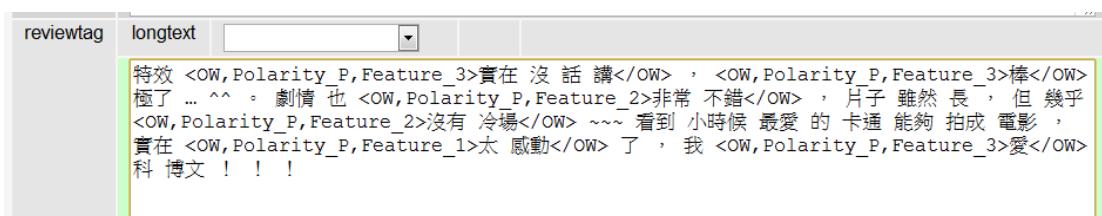


圖 6:標記意見詞資訊

標記意見詞規則：

1. 先標記意見詞的位置。
2. 在標記意見詞的範圍中，我們把輔助意見詞的一些詞彙也給標記起來。例如：副詞修飾意見詞：很、非常，動詞輔助意見詞：覺得、充滿，及否定詞反向意見詞：不、沒有。
3. 標記意見極性，辨別這意見詞對於這部電影的意見方向，標記極性包括正向、負向、中立意見。
4. 標記此意見詞所對應的屬性詞類別
  - 甲、觀察意見詞的小句範圍內是否含有屬性詞，有屬性詞則到乙步驟，沒有則到丙步驟。
  - 乙、有屬性詞的話，並且不是“整體”類別就直接當作此意見詞配對的屬性詞類別，如果是“整體”類別則需要觀察前一小句，是否含有其他類別的屬性詞，及意見詞本身的形容是比較趨向哪一類的屬性詞類別。
  - 丙、沒有屬性詞的話，再觀察前面一個句子，盡量尋找特徵詞，及觀察意見詞本身的形容比較趨向哪一類的屬性詞類別，如果都找尋不到，就對應到“整體”類別。

在兩千篇電影評論中，去除掉沒有意見詞的電影評論，最後剩下 1790 篇為有效評論，在這些評論中標記了意見詞共有 5396 個，平均一個評論有 3.01 個意見詞，極性統計如表七，屬性類別統計如表 9。

表 9: 意見詞極性統計

意見詞極性	正向意見詞	負向意見詞	中立意見詞
次數	3968	1356	74
百分比	73.5%	25.12%	1.3%

在意見詞極性統計中，可以發現負向意見詞比正向意見詞數目少了許多，我們可以觀察評論者使用的負向意見詞，這些負向意見詞很多時候是利用正向意見詞加上否定詞來組合負向意見詞，例如：不好看、不清楚，所以造成標記負向意見詞數量比較少的情況。

表 10:意見詞對應屬性詞類別統計

屬性詞類別	整體	劇情	特效、音效、背景	電影相關人物
次數	2432	1894	485	585
百分比	45.07%	35.1%	8.98%	10.84%

我們觀察意見詞形容的屬性類別情況，如表 10 所示。在每部不同特性的電影，會影響各屬性類別所占的比率，例如：愛情片的評論都會偏向“劇情”類別，若是科幻片性質的電影，就有比較多關於特效的評論，而大卡司、大導演的電影，則會有較多提及演員、導演…等屬性的評論，但評論者大多都會評論整體、劇情這類別，以至於這兩類別所占的比例較高。

### 3.3.2 意見詞詞性組合

在[Turney,2002]、[Ning et al. , 2010]...等文章中，都是利用詞性來探索意見詞的組合，利用這些詞性組合序列，進而擷取意見詞，甚至抓出所形容的屬性詞。本論文使用中研院斷詞系統對中文電影評論斷詞及標記詞性，再對應已標記的訓練語料，觀察意見詞標記的詞性，進而整理出中文意見詞的詞性組合。

我們首先取出訓練語料中被標記意見詞的詞彙，並把這些意見詞斷詞後的詞性整理歸納。表 11 統計全部意見詞的詞性組合，這些標記的意見詞可能含有詞性為副詞、動詞…等輔助意見詞的詞彙，再按照標記意見詞的詞數及出現頻率依序展示。

表 11:標記意見詞詞性統計

	意見詞詞性組合	標記頻率	涵蓋率
單詞	V	2158	40%
	N	285	5.3%
	ADV	50	0.92%
	FW	24	0.44%
雙詞	ADV+V	1780	33%
	V+V	275	5.1%
	V+N	43	0.8%
	ADV+N	28	0.52%
	ADV+ADV	21	0.39%
	A+V	19	0.35%
	ADV+FW	19	0.35%

	V+ Nep	16	0.3%
	V+ADV	16	0.3%
	N+V	15	0.25%
三詞	ADV+ADV+V	54	1%
	V+ADV+V	44	0.81%
	ADV+V+V	20	0.37%
	ADV+V+N	17	0.31%
	V+ Nep + N	12	0.22%

在中研院斷詞系統中，動詞、副詞、名詞可分為很多子類別，例如動詞有VC(動作及物動詞)、VH(狀態不及物動詞)等，但是我們觀察訓練語料中的統計結果，意見詞並無特別偏向哪一小類的動詞或名詞，還是有其他小類的動詞出現，所以在這部分我們忽略子類別詞性，只考慮大類的詞性。

首先我們觀察在語料中出現次數前四多的詞性組合，總共占九成多的標記意見詞頻率，幾乎涵括全部意見詞的詞性組合，所以先探討這四種詞性標記，如表12所示。

表 12:前四大類詞性組合

	詞性組合	例子	相關詞性組合	次數	涵蓋率
單詞	1. V	好看、感動、推薦、不錯	ADV+N	2158	40%
	2. N	普普、刺激、超帥、超棒、老梗、超讚、爛片		285	5.3%
雙詞	3. ADV+V	很值得、非常好笑、很好看、太失望、很喜歡	1. ADV+ADV+V 2. V+ADV+V 3. ADV+V+V 4. ADV+V+N	1942	36%
	4. V+V	覺得好笑、充滿溫暖、值得推薦、超喜歡、有感人		275	5.1%

#### 意見詞四大詞性組合:

1. V:由單一動詞所組成的意見詞，約占全部 40% 的意見詞詞性是最大比例的詞性，利用訓練語料收集當作意見詞。藉由統計可以得知出現頻率較頻繁的詞性 V，有將近八成以上的詞彙量會重複在 ADV+V 詞性組合中的動詞，所以可以利用 ADV+ V 的詞性組合序列擷取含有意見的單一動詞(V)。
2. N:由單一名詞組成的意見詞。評論者有時會使用一些名詞來表達情緒，例如：刺激、爛片...等詞彙，我們利用訓練語料擷取具有意見的名詞詞彙。相關的詞性組合(ADV+N)，例子：超正、最經典。

3. ADV+V:由副詞加動詞所組成的意見詞詞性標記，與單一動詞(V)重疊率高，我們取後面動詞當作意見詞。副詞通常是一些加強輔助意見詞的副詞，例如：很、太、非常。相關詞性組合則有四組，如表 12 所示。
4. V+V: 我們取出第二的動詞來跟 ADV+V 的動詞比較，將近有八成詞彙的重複，具有很大的相似性，我們取句型中的第二個動詞當作意見詞。第一動詞通常是輔助形容後面的意見詞，例如：覺得、充滿。

我們從標記評論語料統計可看出意見詞大多以動詞為主，前三組比例高的詞性組合都與動詞 V 相關，約占標記總次數 86%，出現次數第四高的名詞也是主要表達意見的詞性，所以接下來重點是放在動詞及名詞這些比較含有意見成分的詞性，以它們為中心來尋找詞性組合規則。我們在這步驟中收集 588 個單一動詞、121 個單一名詞當作意見詞。

除了上述這四類詞性組合之外，還有一些數量較少但也有固定詞性規則可以整理，如表 13 所展示的三種詞性組合。

表 13:其餘詞性組合

詞性組合	例子	相關詞性組合	次數	涵蓋率
1. FW	Nice、Good	ADV+FW	43	0.8%
2. V+N	沒有冷場、沒有看頭、有點、沒有劇情		43	0.8%
3. A+V	超級好看、超級感人		19	0.35%

1. FW:外國語文標記。相關詞性組合:ADV+FW，例子: 很 nice，我們從標記語料中取這些外文標記當作意見詞。
2. V+N:在這詞性組合中的動詞大多是否定詞或肯定詞，我們統計出現次數前五高的動詞如下無、絕無、沒、沒有、有。這種詞性組合常會出現“沒有劇情”這類否定詞加上屬性詞的組合。當只有單一名詞“劇情”在句子中是沒有含有意見，通常就是“沒有”、“有”這些動詞加上電影屬性詞才會有表達意見，這部分我們藉由否定詞加上屬性詞擷取整個意見詞。
3. A+V:非謂形容詞加動詞組合，動詞與先前收集的單一動詞幾乎重疊，我們取後面動詞當作意見詞。

上述的詞性組合是比較有規則性可循，出現次數也是前七多，這些詞性組合占全部次數九成以上為主要意見詞的詞性組合。

剩下的詞性組合是標記頻率較少及寫法比較口語化的，這一些意見詞不是只有單一詞彙所組成，我們利用訓練語料挑選作為有用的意見詞，如表 14 所示。

表 14:頻率較少的詞性組合

詞性組合	例子	相關詞性組合	次數	涵蓋率
1.ADV	草草、沒話說	ADV+ADV	70	1.3%
2.V+Nep	沒什麼、沒啥、沒啥重點、沒什麼內容	V+Nep +N	28	0.52%
3.N+V	熱血沸騰、闔家觀賞		15	0.25%
4.V+ADV	悶一點、超正		16	0.29%
5.V+N+V	令人感動、沒話講		19	0.35%

1. ADV:我們收集這些單一副詞作為意見詞，舉例:草草、沒話說。
2. V+Nep: 由動詞加指代定詞組合而成，這些詞彙組合幾乎屬於負向的意見，我們收集這些詞組當作意見詞。
3. N+V: 這組合幾乎屬於成語性的句子，我們收集這些詞彙當作意見詞。
4. V+ADV: 在這句型中的 V 將近六成會出現在 V 單一詞中，但對有些口語化的詞組，例如:“超正”的詞性組合不太符合斷詞規則，我們利用訓練語料收集這些詞彙作為意見詞。
5. V+N+V: 這詞性組合中的第二動詞將近有七成是跟 V 單一詞重疊，我們取最後動詞當作意見詞，在前面“V+N”，都大多是“讓人”、“令人”的字眼。特殊用法例如:沒話講，我們從訓練語料收集這一些詞彙。

### 3.3.3 建立意見詞詞性組合序列

在上一節語料統計中可得知動詞及名詞是比較具有意見的詞性，所以以動詞、名詞為主要來建立詞性組合序列。我們在雙詞、三詞為基礎對標記語料進行探索。觀察標記語料統計的三詞，發現三詞的詞性組合無顯著偏向固定組合，所以我們以雙詞詞性建立詞性組合序列，在標記單詞意見詞的部分，是利用前面一個詞及後面一個詞結合產生雙詞，在標記雙詞意見詞的部分是直接搭配使用，在三詞意見詞的部分，是把前面兩個詞及後面兩個詞當作雙詞。我們把這些收集的雙詞詞性加以統計，前十多次數的詞性組合如表15所示。

表 15:意見詞雙詞統計結果

雙詞詞性組合	次數	覆蓋率
ADV+V	2481	51.02%
V+V	627	12.9%
N+V	396	8.14%
V+N	316	6.50%
DE+V	144	2.96%
Nf+V	100	2.06%
SHI+V	78	1.60%

N+N	66	1.36%
FW+V	62	1.28 %
ADV+N	53	1.09 %

我們觀察表15並考慮上一節討論到動詞、名詞為主要意見詞性，在動詞的部分我們取次數前四多的雙詞組合當作擷取詞性組合序列，這些詞性組合是比較符合擷取動詞意見詞的需求，四種詞性組合序列如下:ADV+V、V+V、N+V、V+N，我們擷取裡面的動詞當作意見詞。另外一方面在名詞部分，我們則是把ADV+N當作意見詞雙詞詞性序列擷取名詞。利用這些詞性序列擷取詞彙來擴充意見詞。在本篇論文中我們使用上述五種組合來擷取意見詞，各個詞性序列的詳細資料如表16所示，表格中原本詞彙是指原本意見詞出現次數與比例，擴充詞彙則是後來擴充得到的意見詞。

表 16:五種雙詞詞性序列

詞性組合	原本詞彙	擴充詞彙	新增 詞彙量	次數	Bigram 覆蓋率
1. ADV+V	3623 (88%)	505(12%)	240	2481	51.02%
2. V+V	699(89%)	82(11%)	59	627	12.89%
3. N+V	673(89%)	87(11%)	56	396	8.14%
4. V+N	551(88%)	73(12%)	48	316	6.49%
5. ADV+N	56(64%)	32(36%)	29	53	1.09%

我們利用這五種詞性組合序列對其他三千篇電影評論自動尋找意見詞，取得可能含有意見成分的動詞、名詞詞彙，把這些詞彙當作候選意見詞，這些詞彙經過篩選後建立專屬電影領域的意見詞庫。我們觀察擷取的動詞、名詞，發現許多詞彙不是屬於意見詞，但這些詞彙都符合這詞性序列中，所以我們需要再進一步對這些候選詞篩選，選出具有意見成分的意見詞彙。Qiu et al. [2009]、陳立 [2010]利用詞彙在各別正負語料出現的頻率計算差異性篩選候選詞，我們採用這方式篩選這些候選詞之後，把這些詞彙收集為意見詞。

奇摩電影語料中每篇電影評論都有評論者本身所打的分數，分數是一顆星到五顆星，在[Chaovalit et al., 2005]文章中，把1、2顆星當作負向語料，另外把4、5顆星當作正向語料，因此我們從奇摩電影中各取四千篇當作正負評論。這些候選詞必須至少出現在正向或負向語料兩次以上，之後再對這些初步擷取出的意見候選詞，計算個別在正向評論出現的次數，利用下面公式2，觀察意見詞在正負評論次數的差異性， $F_P$ 、 $F_N$ 各代表候選詞在正負向語料出現的次數。

$$\text{Difference} = \frac{|F_P - F_N|}{(F_P + F_N)/2} \quad (2)$$

我們對每個候選詞利用公式1計算之後，產生候選詞對正負語料的差異度，原本Qiu et al. [2009]設定的差異度的門檻是一，若差異度小於一則代表這候選詞在正負評論的出現次數相近，代表它是一個通用詞並沒有意見成分，所以我們把

這時候選詞篩選過濾掉。若候選詞的差異度大於等於一，代表這詞彙有偏向正向或負向語料，則把這時候選詞收集當作意見詞。

我們調動門檻值觀察篩選候選詞為意見詞的效果，門檻值範圍是0.6到1.2之間，如表17所示。一開始對於這些候選詞，我們手動去標記含有意見極性的詞彙作為實驗比對，總共標記442個詞彙。我們觀察到門檻值越高，雖然準確率越高，但召回率越低，也會淘汰掉原本含有正負情緒的詞彙，反之門檻值越低，則召回率雖然提高，但準確率會下降，造成篩選之後的意見詞並不是真的含有意見極性成分。我們在兩者之間挑選一個平衡點，挑選出F-score較高的門檻值當作篩選條件，最後我們設定門檻值為0.7，當意見候選詞的差異度大於等於0.7時，我們就將此候選詞視為電影相關意見詞。

表 17: 差異度門檻值調動

門檻大於等於	人工標記結果	系統正確標記	系統標記	Recall	Precision	F-score
0.6	442	316	514	71.5%	58.4%	64.3%
<b>0.7</b>	<b>442</b>	<b>290</b>	<b>427</b>	<b>63%</b>	<b>70%</b>	<b>67%</b>
0.8	442	279	398	63.1%	70.1%	66.4%
0.9	442	257	352	58%	73%	64.7%
1.0	442	247	325	55.9%	76%	64.4%
1.1	442	207	270	46.9%	76.7%	58.1%
1.2	442	198	257	44.8%	77%	56.7%

原始意見詞數量為804個，我們經過此這步驟額外擴充403個動詞意見詞詞彙，29個名詞意見詞詞彙。

我們在篩選作後的意見詞，列舉出幾個辨別含有意見極性正確及不正確的詞彙例子，如表18所示。我們可以發現某一些單一詞彙，單獨觀察此詞彙是比較不含有意見極性，但確實這些詞彙是常常出現在負向語料，例如：死光、幹嘛，也有詞彙常常出現在正向語料，例如：透露、猜測。在這情況下這些不直接有意見極性的詞彙也會通過我們所設定的門檻值。

表18:篩選意見詞彙

擴充的意見詞	意見詞彙舉例
含正負極性的擴充意見詞	籠統、適當、緊繃、絕望、臭屁、可笑、純真、猜想到、美麗、乖乖、受不了、善良、不足以、稱讚、不像話、退錢、拍手、噴淚、猜中、貼近、奇幻、期望、破爛、警示、不適、隨便、羨慕、不值、牽動、改觀、開懷、逗趣、順暢、美好、丟臉
不含正負極性的擴充意見詞	看似、顯現、收集、有錢、小聲、觀賞、拉去、幹嘛、死光、說給、製片、透露、猜測、成為、獨立、操控、相當、請去、排除、結果、飾演、打出來、跳舞、帶出來、出手、放到、扯上

五個雙詞詞性序列統計資料:

1. ADV+V: 我們利用此詞性序列擷取 693 個動詞，經過公式篩選過程後剩下 240 個動詞。
2. V+V: 在此詞性組合，通常第一個動詞是輔助形容第二個動詞，例如:感覺、充滿...等詞彙，第二個動詞才是意見詞，擷取出 283 個動詞，經過公式篩選過程後剩下 59 個動詞。
3. V+N: 在這詞性組合序列中的名詞大部份屬於電影屬性詞。我們擷取出 364 個動詞，經過公式篩選後擷取 48 個動詞。
4. N+V: 在這詞性組合序列中的名詞大部份屬於電影屬性詞。我們擷取出 406 個動詞，經過公式篩選後擷取 56 個動詞。
5. ADV +N: 在這詞性組合，格外擷取出 123 個名詞，經過公式篩選後擷取 29 個名詞。

我們自動擷取新的意見詞詞彙在其他三千篇的奇摩電影評論，擷取出的新詞彙跟之前兩千篇收集的意見詞詞彙來做比較。發現在雙詞詞性序列部分，原本意見詞詞彙占將近九成的意見詞數量，而我們探索出來新的意見詞詞彙只占了一成左右，尤其是出現頻率較高的意見動詞更是明顯。在這觀察之下，我們得知電影評論普遍使用一些固定的意見詞形容電影，這也是為什麼我們要建立一個專屬電影領域的意見詞彙庫，而這些詞彙也可以涵蓋大部分的電影意見詞。

### 3.4 意見詞與屬性詞類別配對

我們利用上一章節收集而成的電影意見詞彙庫，辨識評論句子內是否含有意見詞，如果有含有意見詞，則尋找此意見詞所形容的屬性詞。在意見詞所處的小句範圍內尋找對應屬性詞，我們以屬性詞靠近意見詞距離較近的為優先，而對於意見詞後面接“的”、“之”這些詞彙，則優先辨別後面接下來的詞彙是否為屬性詞。

### 3.4.1 屬性詞與意見詞配對辨識

我們考慮到句子結構因素，一個句子中可能會有單個意見詞形容多個屬性詞的情況，因此提出下列三種句型結構去辨識對應的屬性詞。

#### 1. 句型 1: 屬性詞+連接詞+屬性詞+意見詞

例子:特效和劇情都很棒。

在意見句子中，我們發現屬性詞與屬性詞之間含有並列關係的連接詞，例如：“和”、“與”、“跟”、“及”...等。符號中的逗號“、”也是有並列連接的效果。

在這評論例子中，則我們會尋找到兩對屬性詞與意見詞配對，兩對配對為:(特效，棒)、(劇情，棒)。

#### 2. 句型 2: 屬性詞+意見詞+連接詞+屬性詞

例子:這部電影都很棒除了女主角之外。

句型 2 的連接詞是指含有轉折關係的連接詞，例如：“除了”、“但是”、“不過”、“然而”...等。在這評論例子中，我們會把前面的意見詞加上否定詞，對應到後面的“女主角”，產生兩對屬性詞與意見詞配對，兩對配對為(電影，棒)、(女主角，不棒)。

#### 3. 句型 3: 意見詞+連接詞+屬性詞

例子:真的很棒，比方特效方面。

這例子中意見詞真正所對應的屬性詞，不在同一個小句，就算在前一個小句也找尋不到對應的屬性詞。這些連接詞屬於承接關係的連接詞，例如：“於是”、“比方”、“像”、“例如”...等。我們考慮意見詞下一小句含有這些連接詞並且後面接著屬性詞，則把這屬性詞對應到前面意見詞。在這例子中我們會產生屬性詞與意見詞配對為(特效，棒)。

在奇摩電影的評論語料中，意見詞常會尋找不到所形容的屬性詞的情況。表 19 展示將近五成的意見詞在周圍詞彙內尋找不到屬性詞，而在意見詞的小句範圍內將近四成尋找不到屬性詞，我們這電影語料比較多這種情況。Jakeob et al. [2010]面對這種情況的做法是將意見詞對應到預設值，在電影領域的預設值是指電影“整體”，但這樣做法會使“整體”類別相較其他屬性詞類別占的比例高上許多，事實上有些沒有配對的意見詞也不是真正指向“整體”類別，這種做法會影響到各類屬性詞類別的評分。Zhuang et al.[2006]單純觀察意見詞常形容某些屬性詞，利用這特性把意見詞對屬性詞配對。本篇論文方法是擷取一些意見詞特徵，使用 Libsvm 分類器辨識意見詞所屬的屬性類別。

表 19: 屬性詞與意見詞的配對統計

意見詞位置	特徵詞前面	特徵詞後面	尋找不到
周圍一個詞範圍	1071(20%)	286(5.4%)	3901(74.1%)
周圍兩個詞範圍	1786(33.9%)	836(15.8%)	2485(47.2%)
意見詞小句範圍	3450(64%)		1947(36%)

在意見詞小句內找不到對應屬性詞的情況下，我們期望可以將這些意見詞辨識所屬的類別中。我們首先收集分類特徵再利用Libsvm[Chang et al. , 2001]分類器處理這些特徵，將這些意見詞自動辨識所形容的屬性類別。

分類步驟:

1. 判別意見詞在小句內有無配對的屬性詞，有則到步驟二，沒有到步驟三。
2. 意見詞有屬性詞配對，則觀察屬性詞是否為整體類別，是的話到甲步驟，不是的話到乙步驟。
  - 甲、若屬性詞為“整體”類別，這情況下有間接評論對象及直接評論對象的問題，因此對“整體”類別的屬性詞不可以直接配對。  
舉例:這電影很好笑(劇情)，正常寫法為:這電影的劇情很好笑。  
電影:間接評論對象(目標)，劇情:直接評論對象(屬性)，好笑:意見詞
  - 乙、若屬性詞不是“整體”類別，則意見詞直接配對屬性詞。
3. 意見詞沒有屬性詞配對，則產生意見詞的分類特徵，利用分類器自動辨識所屬的類別。

### 3.4.2 分類特徵

我們使用五種分類特徵辨識意見詞所對應的屬性類別，分別定義把這些分類特徵定義為 F1 到 F5。意見詞分類特徵，列表如下:

#### 1. 意見詞在評論中的第一小句

意見詞位在第一小句的情況下，如果意見詞找不到配對的屬性詞，則大部分是形容“整體”類別。有些電影評論開頭就開門見山直接寫意見詞，通常這情況的意見詞是指向電影“整體”。舉例:好看!!這部電影真棒。

特徵向量為一個維度，如果意見詞在第一小句，則向量為 $<1>$ ，否則向量為 $<0>$ 。

#### 2. 意見詞前一小句的屬性詞

我們考慮文章連貫性，前一小句出現的屬性詞常常是目前意見詞所形容的屬性詞。如果意見詞前一小句含有屬性詞，則抓取這屬性詞當作分類特徵。

特徵向量為四個維度，如果意見詞的前一小句含有第一類別的屬性詞，則向量為 $<1,0,0,0>$ ，若前一小句沒有屬性詞，則向量為 $<0,0,0,0>$ 。

#### 3. 一個電影評論內，評論者常會評論到某一種類別的屬性詞。

評論者通常會比較關注某一類別的屬性詞類別評論。例如:專注特效。這個分類特徵是計算此篇評論中各類屬性詞出現的數量，產生各種類別屬性詞出現的比例。

特徵向量為四個維度，假設評論者講第一類別較多，則向量為 $<1,0,0,0>$ 。

#### 4. 一個意見詞常形容哪一類屬性類別。

某些意見詞會有偏向形容某類的屬性詞的情況。例如:“帥”這意見詞，通常形容演員人物方面。“好聽”這意見詞，則是形容音效、音樂等。我們利用訓

練語料，統計這些意見詞所對應的屬性詞類別，觀察這個意見詞對應各個類別屬性詞的比例。

特徵向量為四個維度，每個維度代表意見詞對應各個類別屬性詞的比例。

##### 5. 利用網路資訊，計算意見詞和屬性類別間的關連性

在網頁內容中，常有關連的兩個詞彙會共同出現在一個網頁，例如：女主角與漂亮。我們利用查詢詞的回傳網頁數量計算 PMI 值，進而觀察兩個詞彙的關連性，如公式 3。

$$PMI(O, C) = \frac{Hits(O+C)}{Hits(O) \times Hits(C)} \quad (3)$$

O 是指所要分類的意見詞，C 是指這類別 C 中的屬性詞。我們取各個類別中五個最高頻率的屬性詞計算，Hits 代表搜尋詞回傳的網頁數目。

特徵向量為四個維度，每個屬性類別取前五個最常出現的屬性詞當作代表，把這些屬性詞跟意見詞做 PMI 計算並取出五個數值平均。每個維度代表這個屬性類別的 PMI 數值占全部類別的 PMI 數值比例。

#### 3.4.3 分類實驗

實驗語料為兩千篇的電影評論，這些評論的意見詞已標記配對的屬性詞類別，總共有 2671 個意見詞與屬性詞類別配對，這些配對是意見詞在小句內無法找到所對應的屬性詞。我們將其中 2141 個配對當作訓練語料，530 個配對當作測試語料，訓練語料跟測試語料的比例大約是 4:1。

我們以支持向量機當作基礎，把上一節所述的分類特徵化成維度向量，支持向量機是 Vapnik et al. [1995] 提出以統計學習理論為基礎，可以應用目前的資料分類及圖形辨識等。我們使用 Chang et al. [2001] 所開發的 Libsvm 來作為意見詞分類的分類器，其中核心函數部分我們選用 RBF。在調最佳參數 gamma 和 cost 的部分，是利用 Libsvm 中的 grid 程式來反覆測試，尋找出最佳的 gamma 和 cost，再經由分類器對訓練語料找出最佳效能，各類屬性詞類別實驗結果如表 20，整體分類器正確率 86%。

表 20：意見詞與屬性詞的配對實驗

屬性詞類別	人工 標記	系統正確 標記	系統 標記	Recall	Precision	F-score
整體	346	332	389	95.95%	85.34%	90.32%
劇情	137	105	115	76.64%	91.3%	83.33%
特效	14	5	5	35.71%	100%	52.62%
電影人物	33	16	21	42.42%	76.19%	51.84%

在表 20 的實驗結果中，我們可以發現特效及電影相關人物辨識效果比較不好，原因是在電影語料中這兩部分的屬性詞類別次數占的比例較低，以至於辨識效果也比較不好。而整體、劇情類別中訓練語料比較多，所以辨識屬性類別有比較好的實驗效果。

### 3.4.4 特徵選取

我們觀察哪些特徵的組合是對辨識屬性類別比較有幫助的，共有五個分類特徵，使用 Le et al. [20004]所提出的 Forward Sequential Selection Algorithm 來做特徵選取，這個特徵選取的方法是首先定一個特徵的空集合 SF(Selection Feature)，然後我們對每一個分類特徵做分類實驗，挑一個最高正確率的特徵  $F_i$  加入 SF 集合中。接著下一步我們將  $F_i$  跟每一個分類特徵搭配，挑選最高分類正確率的組合來當作第二個要放入 SF 中的特徵，一直反覆這步驟直到最後正確率不再比前面特徵組合高為止，最後 SF 就是我們最佳的特徵組合。

我們利用 Forward Sequential Selection Algorithm 比較分類器分類效果，首先對單一特徵做實驗找出最佳的分類特徵，如表 21 所示。

表 21: 分類結果第一回合

單一特徵	正確率
F1. 意見詞在評論中的第一小句	65.00%
F2. 意見詞前一小句的屬性詞	72.22%
F3. 在一個電影評論內，評論者常會評論到某一種類別的屬性詞	65.00%
F4. 一個意見詞常形容哪一類屬性類別	85.87%
F5. 利用網路資訊，計算意見詞和屬性類別的關連性	78.83%

我們根據表 21 的實驗結果，我們在第一回合挑選出 F4“一個意見詞常形容哪一類屬性類別”這個特徵，將此特徵加入我們的 SF 集合，這時候  $SF=\{F4\}$  再與其它特徵配對，進行第二回合實驗。

表 22: 分類結果第二回合

雙特徵	正確率
F4+F1	86.67%
F4+F2	86.67%
F4+F3	87.78%
F4+F5	86.11%

在表 22 第二回合中，我們挑選 F3“在一個電影評論內，評論者常會評論到某一種類別的屬性詞”這個特徵，將此特徵加入我們的 SF 集合，這時候  $SF=\{F4、F3\}$  再與其它特徵配對，進行第三回合實驗。

表 23:分類結果第三回合

三特徵	正確率
F4+F3+F1	86.94%
F4+F3+F2	87.22%
F4+F3+F5	86.39%

我們可以看出在表 23 第三回合中沒有更高的正確率，所以特徵選取結束，我們以第二回合選出的  $SF=\{F4, F3\}$  為最佳的特徵組合。在這特徵組合下，我們得到正確率 87.22%，而在使用全部特徵的情況下正確率是 86%，代表這特徵選取可以得到更好的分類效果。

我們觀察選取特徵過程，發現 F4“一個意見詞常形容哪一類屬性類別”為主要增加分類效果的特徵，以及搭配 F3“在一個電影評論內，評論者常會評論到某一種類別的屬性詞”特徵可以達到最好的分類效果。我們觀察其它分類特徵，F1 特徵在意見詞在第一小句，這種情況相對來說比較少數，對分類效果沒有很大的助益，而 F2 特徵在意見詞前一小句的屬性詞，也不一定是此意見詞所形容的屬性詞，F5 特徵利用網路資訊計算詞彙關連性，沒有明顯的助益效果。

### 3.5 意見評分

在這節中我們會利用 3.3 節擷取的意見詞，對每個意見詞會給它一個意見分數。意見分數範圍是一到五分，一分是最低評價並偏向負向意見，而五分偏向正向情緒。我們考慮修飾意見詞的副詞及意見詞前的否定詞的影響，所以在這些影響下去修改意見詞分數，最後我們觀察意見詞所處的小句位置、小句內的轉折語當作這意見詞在整個評論的權重，最後產生整個評論的評分。

#### 3.5.1 意見詞計分

我們參考 [Marneffe et al., 2010] 的方法產生每一個意見詞的分數，此篇方法是收集奇摩電影評論一到五顆星各約兩千篇，計算這意見詞在各個星等的評論出現的機率，機率乘上目前星等的顆星數加總為此意見詞分數，但我們觀察產生的意見分數，不少意見詞的分數都在三分左右，這些意見詞分數沒有很明顯地偏向五分或者一分，而這些分數在後續的評論自動評分實驗中效果也不太好。所以我們修改產生意見詞分數的方法。

我們單純把評論語料中四、五顆星當作正向語料，一、二顆星當作負向語料，各別算出意見詞在正負語料中出現的機率，如果是正向語料的機率是乘上 5 分，而負向語料機率則是乘上 1 分，兩個數值相加當作意見詞分數。如果意見詞含有強烈的正向意見會靠近 5 分，如果意見詞比較沒有強烈正向意見則會遠離 5 分，反觀負向意見也是如此，而意見分數越靠近 3 分則代表意見極性越中立，意見分數產生如公式 4、5、6。

$$F(w_i, t) = \frac{\text{count}(w_i, t)}{\text{count}(t)} \quad (4)$$

$$P(w_i, t) = \frac{F(w_i, t)}{\sum_{t' \in T} F(w_i, t')} \quad (5)$$

$$\text{Word\_score}(w_i) = \sum_{t \in T} S \times P(w_i, t) \quad (6)$$

正面語料為 8000 篇，負面語料為 7189 篇。在公式 4 中是計算意見詞在語料  $t$  中出現頻率， $w_i$  是指這意見詞， $t$  是指正負語料，而  $\text{count}(t)$  是指在語料  $t$  中全部詞彙出現數量， $\text{count}(w_i, t)$  代表在語料  $t$  中詞彙  $w_i$  出現數量。公式 5 利用  $w_i$  在語料  $t$  出現頻率，來計算意見詞  $w_i$  在正負語料出現的機率。公式 6 在正面語料  $S$  為 5，負面語料  $S$  為 1，把意見詞出現正面語料的機率乘上 5 分，加上出現在負面語料的機率乘上 1 分加總起來，當作此意見詞的意見分數。

計算意見詞分數範例：

假設正面語料  $t$  詞彙量為 10000 個，負面語料  $t$  詞彙量為 8000 個

意見詞  $w_i$  在正面語料出現數量為 100 個，負面語料為 20 個

$t$  是正面語料:  $F(w_i, t) = 100/10000 = 0.01$

$t$  是負面語料:  $F(w_i, t) = 20/8000 = 0.0025$

$t$  是正面語料:  $P(w_i, t) = 0.01/(0.01+0.0025) = 0.8$

$t$  是負面語料:  $P(w_i, t) = 0.0025/(0.01+0.0025) = 0.2$

計算意見詞  $w$  的分數:  $\text{Word\_score}(w_i) = (5 \times 0.8) + (1 \times 0.2) = 4.2$

產生各個意見詞的意見分數之後，我們做一個意見極性判斷實驗，利用之前兩千篇評論標記的評論語料，收集正面意見詞及負面意見詞當作測試語料，正面意見詞為 451 個，負面意見詞為 255 個，把這些標記意見詞與自動產生的意見詞分數比較，我們定  $\text{Word\_score}(w_i)$  大於等於三視為正面意見詞， $\text{Word\_score}(w_i)$  小於三則視為負面意見詞，觀察意見分數的極性辨別效果，實驗結果如表 24。

表 24: 意見詞情緒極性實驗

意見詞極性	人工標記結果	系統正確標記	系統標記	Recall	Precision	F-score
正面意見詞	451	396	435	87.8%	91%	89.3%
負面意見詞	255	216	271	84.7%	79.7%	82.1%

實驗結果中我們發現單純判別意見詞極性效果 F-score 都有 80% 以上甚至 90%，這說明自動產生的意見詞分數在意見極性辨識有相當的準確度。

我們列舉一些意見詞及它的意見分數，如表 25 所示，我們可以發現利用意見分數辨別極性的錯誤情況中，這些意見詞分數大多都是靠近三分左右，某方面

來說這些意見詞都靠近中立分數，以致於會有辨識極性錯誤的情況。

表 25: 意見詞分數舉例

人工標記意見詞極性	系統產生意見詞分數	意見詞分數舉例						
正向意見詞	系統正向意見詞	完美	超好	震撼力	享受	歡樂	好笑	歡笑
		4.1	4.2	4.2	3.9	3.9	3.5	4.5
負向意見詞	系統負向意見詞	好評	分明	清晰	典型	奇妙	頗好	尖叫
		2.6	2.5	2.8	2.9	2.1	2.0	2.5
負向意見詞	系統正向意見詞	醜女	遺憾	不屑	機車	倉促	暎草	複雜
		3.3	3.8	3.1	3.2	3.1	3.1	3.4
負向意見詞	系統負向意見詞	牽強	糟糕	抄襲	無理	納悶	乏味	狗屁
		2.5	1.8	1.0	2.6	2.8	2.2	1.2

我們利用訓練語料標記的結果，找出同時有兩種極性的意見詞，這些意見詞需要前後文判斷情緒極性。例如：老、芭樂、大、深、少、簡單、平淡、太多、平平淡淡、誇張...等，及情緒中立詞，例如：普通、還可以、還算可以、普普通通...等，我們收集這些詞彙作為後續評分處理。

### 3.5.2 評論評分

3.5.1 節產生出每個意見詞的意見分數，我們參考 Zhang et al. [2009]考慮修飾意見詞的副詞及否定詞的影響。我們在處理副詞部分，對意見詞極性有加強作用的副詞作處理，如表 26 所示，定義這些副詞詞彙為副詞集合。從 3.3.3 節詞性序列 ADV+V 收集而來，這些副詞對意見詞的極性有加強的作用，所以對意見詞的分數也有所影響。我們設參數為  $\alpha$  來對意見詞分數修改。在否定詞部分，否定詞出現在意見詞前面，這個否定詞有反轉這意見詞的意見極性效果，我們設參數為  $\beta$ 。如果沒有副詞、否定詞，則參數預設值為 1。否定詞列表如表 27，定義這些詞彙為否定詞集合。

表 26:修飾意見詞副詞

很	非常	最	太	好	還
滿	真是	都	有夠	夠	真的
真	那麼	挺	更	這麼	比較
蠻	絕對	極力	越	實在	極為
更加	如此	多麼	一直	頗	多
十分	相當				

表 27:否定詞列表

不覺得	不要	不可能	不	不是	勿
缺乏	缺	了無	很少	不會	難
否認	比不上	全無	難以	絕無	無
缺乏	不至於	少	毫無	毫不	萬萬不
請勿	從不	不再	決不	不用	怎會
沒有	哪有	原本	本來	以為	別
沒	無	不會	過於	哪裡	

利用公式 7 修改意見詞分數，原本 Word\_score 範圍是[1,5]轉換成[-1,+1]，之後乘上 $\alpha$ 、 $\beta$ 參數修改意見詞分數，產生 Modified\_score 來達到加強或反轉意見極性的作用。

$$\text{Modified\_score}(w_i) = \alpha_i \times \beta_i \times \text{Word\_score}(w_i) \quad (7)$$

$$\alpha_i = \begin{cases} 1.5, & \text{如果意見詞 } w_i \text{ 前面兩個詞中有包含副詞集合的詞彙} \\ 1, & \text{如果意見詞 } w_i \text{ 前面兩個詞中沒有包含副詞集合的詞彙} \end{cases}$$

$$\beta_i = \begin{cases} -0.4, & \text{如果意見詞 } w_i \text{ 前面三個詞中有包含否定詞集合的詞彙} \\ 1, & \text{如果意見詞 } w_i \text{ 前面三個詞中沒有包含否定詞集合的詞彙} \end{cases}$$

我們考慮意見詞所位處的小句及轉折語。意見詞所處的小句位置，在整個評論意見詞位於在開頭或結尾的句子中，通常會比中間句子有重要性，更能表達整個評論的意見方向，所以我們對位在開頭或結尾的意見詞給予較大的權重，評論中開頭、結尾的句子定義為評論前面兩小句及後面兩小句，權重值設為 $\lambda_i$ ，如果句子位置不是位在開頭或結尾，則權重 $\lambda_i$ 預設值為 1。

在轉折語方面，我們收集開頭含有這些轉折語的句子，例如：含有假設語氣的詞彙等，如表 28 所示，定義為轉折語集合。[林宇中, 2003]論文中，提出含有這類轉折語的句子通常不是整個評論所要主要表達的語意，而 Narayanan et al. [2009]也探討著轉折語對意見的影響。例子：“雖然這部電影演員都不是巨星，但整部電影很棒！”實際上訴的評論主要表達的語意是“電影很棒”，而不是“演員都不是巨星”這句話。在意見詞小句中如果含有這類轉折語，則給予這意見詞較

小的權重，權重值設為 $\gamma_i$ ，如果句子中沒有這些轉折語，則權重 $\gamma_i$ 預測值為 1。

表 28:轉折語列表

縱使	原本	假設	否則	要不	要不然
儘管	不然	縱	雖然	雖	縱然
如果	假設	想說	以為	本來	

我們計算整個評論的分數如公式 8，假設一個評論內有 n 個意見詞，每個意見詞分數都已考慮過副詞、否定詞修改產生 Modified\_score，我們考慮這些意見詞所處的句子位置是開頭或結尾及是否含有轉折語，將這些分數乘上權重值，再除以全部意見詞產生的權重值，算出意見詞平均的分數，最後再配合網路評論的評分範圍，把分數範圍[-1,+1]轉換成[1,5]，當作整個評論的自動評分。

$$\text{Document\_score}(d_j) = \frac{\sum_{i=1}^n \text{Modified\_score}_i \times \lambda_i \times \gamma_i}{\sum_{i=1}^n \lambda_i \times \gamma_i} \quad (8)$$

$$\lambda_i = \begin{cases} 1.5, & \text{如果意見詞 } w_i \text{ 所在的位置是評論開頭或結尾} \\ 1, & \text{如果意見詞 } w_i \text{ 所在的位置是評論中間} \end{cases}$$

$$\gamma_i = \begin{cases} 0.3, & \text{如果意見詞 } w_i \text{ 所在的小句有含轉折語集合的詞彙} \\ 1, & \text{如果意見詞 } w_i \text{ 所在的小句沒有含轉折語集合的詞彙} \end{cases}$$

我們收集一到五顆星評論當作實驗語料，把五、四顆星的評論當作正向評論，總共有 1928 篇，三顆星的評論當作中立評論，共有 953 篇，一、二顆星的評論當作負向評論，共有 1912 篇。利用公式 7 產生出整個評論的評分，把評論的評分四捨五入後自動對應到這三種評論。

我們首先對副詞、否定詞參數來處理，我們設定副詞參數是 1.1 到 2.0，代表這意見詞分數最低是原本分數的 1.1 倍，最高為 2 倍，而否定詞部分我們把範圍定為 -1 到 -0.1，分數乘以 -1 的話代表意見詞分數完全正負向顛倒。我們利用分三種評論的實驗，調整副詞、否定詞的參數，藉由比對分類效果挑選最佳的參數，副詞參數  $\alpha$  為 1.5，否定詞參數  $\beta$  為 -0.4，如果乘上參數超過最大範圍，例如：數值超過 1 則以最大範圍的數值 1 為限。

我們產生出副詞、否定詞的參數後，再對句子的位置、是否含有轉折語的權重給一定範圍，句子在開頭或結尾的權重  $\lambda_i$  範圍為 1.1 到 2，代表這意見詞最多比其它意見詞多一倍權重，在句子含有轉折語的權重  $\gamma_i$  範圍為 0.1 到 0.9 之間，代表這意見詞至少有其它意見詞 0.1 倍的權重。我們利用評分實驗結果算出權重的最佳解，最佳權重  $\lambda$  為 1.5， $\gamma$  為 0.3。

計算範例:

評論例子:雖然這部電影的卡司“不是很強”，但整部電影“很棒”!

假設意見詞“強”分數為 0.25，“棒”分數為 0.5

意見詞“不是很強”的修改分數= $-0.4 \times 1.5 \times 0.25 = -0.15$

意見詞“很棒”的修改分數= $1.5 \times 0.5 = 1$

整個評論評分 Document\_score(d<sub>j</sub>)

$$=[(-0.15 \times 1.5 \times 0.3) + (1 \times 1.5 \times 1)] / [(1.5 \times 0.3) + (1.5 \times 1)] = 0.73$$

我們在參數  $\alpha$  為 1.5、 $\beta$  為 -0.4，權重  $\lambda$  為 1.5、 $\gamma$  為 0.3 的情況下，實驗出表 29 的評論評分結果，下面無句型處理是完全沒有考慮這些情況帶預設值 1 進行評分，我們可以看得出在考慮副詞、否定詞的影響及句子的位置、轉折語之後，對自動評分不管哪一種類的評論都有增加評分的準確度，相對原本方式的自動評分有助益的效果。

表 29:評論評分實驗

		網頁 標記	系統正 確標記	系統 標記	Recall	Precision	F-score
有句型 處理	正向評論	1928	1330	1778	68.99	74.80	71.78
	中立評論	953	389	1376	40.81	28.27	33.40
	負向評論	1912	1162	1639	60.77	70.9	65.45
無句型 處理	正向評論	1928	1295	1822	67.17	71.08	69.07
	中立評論	953	412	1586	43.23	25.98	32.45
	負向評論	1912	994	1385	51.99	71.77	60.3

我們做一個分類五級分評論的實驗，把我們系統自動產生的評論評分，四捨五入對應到五級分中，比較與使用者自己所打的分數是否有吻合，實驗結果如表 30，我們可以發現在要單獨分類五、四分及一、二分是有一點難度的，因為五、四分比較屬於正向評論，使用者在評論正向評論時，會打四分或五分比較取決自己主觀因素，所以我們可以看出在正向極端五分及負向極端一分的分類效果比較好，越靠近中間模糊地帶分類效果越差，在表 31 的中間斜線部分代表正確分類到五級分的評論數量。我們可以觀察系統產生的分數，發現在表格右上角有很多錯誤分類級分的評分，說明這些系統評分大多都是比使用者所給分數還要高。

表 30:評論五級分評分實驗

評論網路評價	系統正確標記	系統標記	Recall	Precision	F-score
5 分	52	138	0.52	0.38	0.47
4 分	27	132	0.27	0.20	0.23
3 分	27	100	0.27	0.27	0.27
2 分	30	97	0.3	0.31	0.31
1 分	20	33	0.2	0.6	0.3

表 31:評論五級分評分實驗比較

		系統評論評價				
		1 分	2 分	3 分	4 分	5 分
網路評論評價	1 分	20	47	16	11	6
	2 分	7	30	34	21	8
	3 分	3	16	27	35	19
	4 分	1	3	16	27	53
	5 分	2	1	7	38	52

最後我們單純看系統評分跟實際上評論者所給的分數差距，如表 32 所示，發現九成評分差距是在兩分以內，說明我們系統自動評分大部分還是靠近使用者所給的分數。

表 32:評論五級分差距統計

系統評分與評論者評分差距	差距小於等於 1	差距小於等於 2	差距小於等於 3	差距小於等於 4
正確率	332(67%)	456(91%)	488(98%)	400(100%)

我們觀察這一些錯誤評分的評論，發現評論者發表的評論與給予的評分往往不一定具有相關性。有些評論內容幾乎都是好評，評論者卻給予不高的分數，這都與評論者主觀因素相關。有一些評論者比較嚴苛，評論中大部分都是好評但，還是給予不高的分數，如下面評論例子。而有一些評論者則比較寬鬆，評論中雖然有批評，但還是給予最高分的評價，這都是使我們分類級分效果變差的因素。

#### 評論五級分評分例子:

很感動的一部電影,尤其是從周杰倫跑出畢業堂外開始到最後他回到 1979 年時..我在家裡哭到我爹都一直笑!! 桂綸鎂演的真的很好!! 我從那天看,到現在我還是

無法釋懷..我想我可能中毒太深了吧?總覺得這幾天看完之後心中有一股鬱悶感,  
然後又想到小雨寫完立可白趴在桌上死去的畫面,我就想哭..

使用者評分:2

系統評分:4.8



## 第四章 電影評分

在 3.5.2 章節，我們對每篇評論計算評分，這一章則是自動產生每部電影的評分。對我們所要自動評分的電影，收集這部電影的所有評論，之後產生每個評論的評分，把這些已評分的評論加總平均當作這部電影的總評分，如公式 9，N 代表這部電影有 N 篇評論，Movie\_score(M) 代表電影 M 的總評分。我們收集每部超越一百篇評論的電影並擷取這些電影的網路評價，我們將這網路評價四捨五入定義為 2 到 5 顆星，利用這些網路評價進一步去跟我們系統自動產生的評分比較，實驗如表 33。

$$\text{Movie\_score}(M) = \frac{\sum_{i=1}^N \text{Document\_score}(d_i)}{N} \quad (9)$$

表 33:電影評分實驗

電影網路評價	電影數量	系統正確標記	系統標記	Recall	Precision	F-score
5 分	86	67	67	0.78	1.0	0.88
4 分	109	84	103	0.77	0.82	0.80
3 分	20	20	51	0.95	0.37	0.54
2 分	11	2	3	0.27	0.6	0.38

在表 34 實驗中，我們發現在實驗語料充足的四分和五分電影時，這部分有比較好的實驗效果。系統辨別高評價的電影效果會比低評價的電影好，在電影自動評分，某些電影評分產生三分的比例較高，但實際上電影評分是在四、二分範圍，以致於準確率較低，而召回率較高。在兩分的電影評分也會被評為三分的情況以致於召回率較低。整個評分正確率約 77%。

在另外一種方法電影評分方式中，我們利用此電影全部評論，計算出每個屬性詞類別評分，公式 11 中  $S(C_j)$  代表屬性詞類別  $C_j$  的總分，代表 p 個意見詞形容這類屬性類別。我們把配對屬性類別的意見分數加總平均，當作這電影屬性類別的分數。之後把這四類屬性類別分數，乘上對應的權重  $Q_j$ ，產生電影 M 的總評分，如公式 10， $S(C_j)$  代表每類屬性類別的分數。每個屬性類別的權重  $Q_j$  是計算在整個電影類別所占的比例，在公式 12 中  $\text{Count}_j$  代表意見詞數量形容屬性類別 j。我們利用自動產生的電影評分去跟網路評價比較，實驗如表 26。

$$\text{Movie\_score}(M) = Q_1 \times S(C_1) + Q_2 \times S(C_2) + Q_3 \times S(C_3) + Q_4 \times S(C_4) \quad (10)$$

$$S(C_j) = \frac{\sum_{i=1}^p \text{Modified\_score}(w_i)}{p} \quad (11)$$

$$Q_j = \frac{Count_j}{\sum_{i \in C}^C Count_i} \quad (12)$$

表 34:電影評分實驗

電影網路評價	電影數量	系統正確標記	系統標記	Recall	Precision	F-score
5 分	86	56	56	0.65	1.0	0.79
4 分	109	99	129	0.90	0.77	0.83
3 分	20	19	36	0.95	0.53	0.68
2 分	11	4	5	0.36	0.8	0.5

我們把這個實驗跟第一個實驗比較，發現系統評分比較沒有靠攏三分的情況，雖然在五分的實驗效果比第一個方法低，但整體在辨識四、三、兩分效果比較好。這方法整體的正確率 79% 比較高，具有比較好的電影評分效果。

我們觀察兩部電影，一部屬於愛情劇情片為“新娘大作戰”，一部特效片為“阿凡達”，利用這兩部不一樣類型的電影，我們觀察電影各類屬性類別的評分。表 35 為“新娘大作戰”，表 36 為“阿凡達”，我們首先觀察類別涵蓋率部分，愛情片在劇情類別所占的比例會比特效片高，說明評論者對愛情片比較會評論劇情這個類別，相對“新娘大作戰”的特效類別，我們也可以看得出懸殊的比例差別，愛情片幾乎沒有人會去評論特效類別，而在特效片就有比較多關於特效的評論。“阿凡達”在特效的評價比劇情的評價高，代表特效是這部電影的強項，而“新娘大作戰”的劇情分數也會比特效分數高。我們列出每個類別中前五個最常出現意見詞，提供使用者了解這部電影各方面的優缺點，讓使用者當作挑選電影的依據，幫助使用者挑選自己所喜愛的電影，例如感動，讓喜愛感人劇情的使用者可以選擇這部電影。

表 35:“新娘大作戰”評價

新娘大作戰	整體	劇情	特效、音效	導演、演員
系統類別評分	4.8	<b>4.63</b>	<b>4.28</b>	4.85
類別涵蓋率	36%	<b>49%</b>	<b>2%</b>	14%
常用意見詞	好笑 好看 感人 好 讚	感動 好笑 好 好看 哭	少 哭 不錯 好 加分	適合 好 爆笑 漂亮 正
系統電影評分	4.7			
網路電影評分	4.5			

表 36：“阿凡達”評價

阿凡達	整體	劇情	特效、音效	導演、演員
系統類別評分	4.35	<b>4.10</b>	<b>4.4</b>	4.08
類別涵蓋率	36%	<b>30%</b>	<b>26%</b>	8%
常用意見詞	好看 好 值得 棒 讚	好 好看 感動 棒 值得	好看 好 棒 讚 震撼	好 鄙視 好看 大 美
系統電影評分	4.26			
網路電影評分	4.8			



## 第五章 結論

本論文提出並實作一個應用意見探勘的方法在中文電影評論語料上，系統自動對電影做評分與網路上已有的電影評價來比較。本論文的主要貢獻如下：

1. 建立 226 部電影資料庫，資料庫包含評論內容、評論評價和電影網路評價。
2. 人工收集電影屬性詞並利用同義詞詞林將這些屬性詞擴充，最後我們將這些電影屬性詞分成四種類別。
3. 我們利用標記語料建立意見詞詞性序列，並使用這些詞性序列擷取意見詞。
4. 建立專屬電影領域的意見詞及產生每個意見詞的分數。
5. 在意見詞分數方面，考慮副詞加強及否定詞反轉的效果。在中文語言句型方面，提出句型結構來更準確地找出屬性詞與意見詞配對。
6. 比較系統自動產生每部電影總評分與網路上已有的電影評價。

在本論文的未來研究，有下列幾個方向：

1. 屬性詞自動化擷取及分類。
2. 在屬性詞分類方面，屬性詞不一定固定屬於一個類別而已，屬性詞可以同時屬於不同類別並解決屬性詞語義歧義的問題。
3. 利用收集每部電影的意見詞，可以對電影加以分類。例如：電影常出現感動...等意見詞，在這情況這部電影可能屬於劇情片。
4. 把電影評分系統應用在專家寫的電影評論，比較在不同語料、知識下的評分效果。

## 參考文獻

- [1] 梅家駒等編著，同義詞詞林，臺灣東華書局股份有限公司，1997 年 3 月
- [2] 李振昌，李禦璽，陳信希，“中文文本人名辨識問題之研究”，第七屆計算語言學研討會論文集，1994 年，pp.203- 222.
- [3] NTUSD(National Taiwan University Semantic Dictionary)  
<http://nlg18.csie.ntu.edu.tw:8080/opinion/pub1.html>
- [4] 古倫維，李佳穎，陳信希，“意見持有者辨識之研究”，中文計算語言學期刊，2009年，第十四卷第四期
- [5] CKIP，中央研究院中文斷詞系統，<http://ckipsvr.iis.sinica.edu.tw/>，2011年
- [6] Li Zhuang, Feng Jing, and Xiao-Yan Zhu (2006), "Movie review mining and summarization", In the Proceedings of the CIKM Conference
- [7] X. Ding, B. Liu, and P. Yu (2008), "A Holistic Lexicon-Based Approach to Opinion Mining" , Proceedings of the first ACM International Conference on Web search and Data Mining (WSDM)
- [8] Pimwadee Chaovalit and Lina Zhou(2005), "Movie Review Mining: a Comparison between Supervised and Unsupervised Classification Approaches", Proceedings of the 38th Annual Hawaii International Conference on System Sciences (HICSS)
- [9] Qiang Ye, Wen Shi, and Yijun Li(2006) , "Sentiment Classification for Movie Reviews in Chinese by Improved Semantic Oriented Approach", Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS)
- [10] 窩格致，賴昆棋，“基於網路社群之旅遊經驗及對應情境之情感意見分析研究”，Proceedings of the 22nd Conference on Computational Linguistics and Speech Processing (ROCLING) ，2010 年
- [11] Turney(2002), "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews", Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 417-424
- [12] Likun Qiu, Weishi Zhang, Changjian Hu, and Kai Zhao(2009), "SELC : A Self-Supervised Model for Sentiment Classification", CIKM
- [13] 陳立，“中文情感語意自動分類之研究”，國立臺灣師範大學資訊工程所碩士論文，2010年
- [14] Minqing Hu and Bing Liu(2004), "Mining and summarizing customer reviews", KDD, pp.168-177
- [15] Oren Etzioni and Ana-Maria Popescu(2005), "Extracting product features and opinions from reviews", Proceedings of the conference on Human Language

Technology and Empirical Methods in Natural Language Processing(HLT)

- [16] Bing Liu, Minqing Hu, and Cheng Junsheng(2005), "Opinion Observer: Analyzing and Comparing Opinions on the Web", 14th international conference on World Wide Web(www), pp. 342–351
- [17] Marie-Catherine de Marneffe, Christopher D. Manning, and Christopher Potts(2010) ,“Was it good? It was provocative.”Learning the meaning of scalar adjectives”, 48th Annual Meeting of the Association for Computational Linguistics(ACL)
- [18] Qi Su, Xinying Xu, Honglei Guo, Zhili Guo, XianWu, Xiaoxun Zhang, Bin Swen, and Zhong Su(2008), "Hidden Sentiment Association in Chinese Web Opinion Mining", 17th international conference on World Wide Web(WWW)
- [19] Binyang Li, Lanjun Zhou, Shi Feng, and Kam-Fai Wong(2010) , "A Unified Graph Model for Sentence-based Opinion Retrieval", 48th Annual Meeting of the Association for Computational Linguistics(ACL) , pp. 1367–1375
- [20] Changli Zhang, Daniel Zeng, Jiexun Li, Fei-Yue Wang, and Wanli Zuo(2009) , "Sentiment Analysis of Chinese Documents: From Sentence to Document Level", Journal of the American Society for Information Science and Technology, pp.2474-2487
- [21] Chih-Chung Chang and Chih-Jen Lin, LIBSVM : a library for support vector machines (2001), Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>
- [22] Corinna Cortes and Vladimir Vapnik (1995), "Support-Vector Networks", Machine Learning, Vol. 20, pp. 273-297.
- [23] 林宇中，“基於語意內容分析之情緒分類系統”，國立成功大學，碩士論文，2003 年
- [24] Niklas Jakob and Iryna Gurevych(2010) , "Using Anaphora Resolution to Improve Opinion Target Identification in Movie Reviews", Proceeding ACLShort '10 Proceedings of the ACL Conference Short Papers, USA
- [25] Ramanathan Narayanan, Bing Liu, and Alok Choudhary(2009). "Sentiment Analysis of Conditional Sentences" , Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP-09), August 6-7, Singapore.
- [26] Valentin Jijkoun, Maarten de Rijke, and Wouter Weerkamp(2010), "Generating Focused Topic-Specific Sentiment Lexicons", In Proceedings of ACL. pp.585~594
- [27] Wei Jin, Hung Hay Ho, and Rohini K. Srihari(2009), "OpinionMiner: A Novel Machine Learning System for Web Opinion Mining and Extraction", In Proceedings of KDD, pp.1195-1204

- [28] Lili Zhao and Chunping Li(2009), "Ontology Based Opinion Mining for Movie Reviews ", Lecture Notes in Computer Science, pp.204-214
- [29] Kuat Yessenov and Sa'sa Misailović(2009) , "Sentiment Analysis of Movie Review Comments" , Massachusetts Institute of Technology
- [30] Dan Fingal, Je Michels, and Jamie Nicolson(2004), "Summarizing Movie Reviews", Stanford University
- [31] 朱嫣嵐, 閔錦, 周雅倩, 黃萱菁, 吳立德, “基於HowNet的詞彙語義傾向計算”, 中文信息學報, 上海, 2006年
- [32] Changli Zhang, Daniel Zeng, Jiexun Li, Fei-Yue Wang, Wanli Zuo(2009), "Sentiment analysis of Chinese documents: From sentence to document level", JASIST, pp. 2474-2487
- [33] Le Cuong Anh, Shimazu Akira. (2004), "High WSD Accuracy Using Naïve Bayesian Classifier with Rich Features", PACLIC 18, Waseda University, Tokyo, pp. 105-113



## 附錄

表 37:整體類別在標記語料中次數統計

整體			
部 1776	心情 57	驚悚片 19	質感 5
片 1244	表現 55	迪士尼 18	精采度 4
集 776	從頭 45	武打片 16	金馬獎 4
感覺 642	票房 43	商業片 15	大體 4
電影院 389	動作片 43	特色 15	意境 4
整個 259	水準 42	大致上 14	議題 3
整體 161	片名 39	噱頭 14	排隊 3
國片 155	片長 38	結論 13	愛情戲 3
看電影 143	期望 35	配合 13	路線 3
戲 139	劇 35	概念 11	功夫片 3
其他 118	基本 31	教育 9	歷史片 2
時間 115	題材 30	主題 9	主線 2
預告片 111	作品 28	戰爭片 7	情感戲 1
戲院 103	風格 25	銀幕 7	靈異片 1
手法 96	印象 22	完結篇 6	現代劇 1
預告 96	紀錄片 22	記錄片 6	人數 1
續集 88	災難片 19	翻譯 6	戰爭劇 1
影片 76	設計 19	總體 5	

表 38:劇情類別在標記語料中次數統計

劇情			
劇情 1408	張力 46	內心戲 15	轉折點 3
最後 376	功夫 43	看書 15	情結 3
結局 346	那邊 43	轉折 14	話語 3
內容 293	對話 42	戰 14	句點 3
開始 265	節奏 40	刻畫 13	交戰 3
地方 254	事情 39	片尾 12	轉場 3
後面 230	編劇 38	追逐 12	連接 3
故事 192	人性 38	描寫 11	過招 3
愛 178	娛樂 36	架構 11	對戰 3
段 159	橋段 35	掛 11	段落 3
前面 157	詮釋 34	氣勢 10	插曲 2
笑點 150	改編 34	鋪 10	門琴 2
死 149	感情 31	鋪成 9	構思 2
打 132	舞蹈 30	笑料 9	涵意 2
打鬥 118	現實 30	結構 8	作戰 2
幕 114	邏輯 29	剪接 8	書本 2
結束 113	劇本 29	講話 8	賽車 2
想像 109	台詞 29	片頭 7	笑場 2
武打 101	氣氛 28	時空 7	決鬥 2
講 98	對白 28	招式 7	對打戲 1
交代 97	開頭 28	飛車 7	攻城 1
小說 97	情感 27	形象 7	出招 1
部分 93	戰爭 27	連結 7	情面 1
中間 89	冒險 25	全程 6	愛情劇 1
梗 88	原著 23	創作 6	闡述 1
幽默 84	愛情片 22	刻劃 6	前戲 1
書 75	對決 22	涵義 6	鋪成 1
愛情 71	細節 21	溫馨片 6	癡情 1
片段 68	搞笑片 21	拳法 6	解謎 1
東西 65	情緒 21	步調 6	安插 1
一段 62	伏筆 21	旁白 6	感動片 1
句 59	寫實 20	古裝 5	射擊 1
真實 55	安排 19	情義 5	刻化 1
歷史 53	死法 17	真愛 5	
鋪陳 52	打架 17	環節 5	
情節 50	敘述 17	打仗 5	

意義 50	對打 16	主軸 4	
過程 50	演戲 15	故事性 4	
結尾 48	描述 15	床戲 4	
爆點 46	刻 15	營造 4	
關係 46	編排 15	ending 3	



表 39:特效類別在標記語料中次數統計

特效、音效、場景			
特效 405	3d 23	風景 8	特技 2
3D 386	角度 22	主題曲 8	景觀 2
畫面 183	中文 21	環境 5	顏色 2
效果 181	拍得 21	歌曲 5	曲子 1
場面 129	配音 19	色彩 5	景 1
動畫 96	歌 19	奇幻 5	化妝術 1
場景 96	科幻 19	聲效 4	景色 1
音樂 86	聲光 18	V8 4	布景 1
拍攝 72	服裝 17	照片 4	風光 1
音效 65	科幻片 16	運鏡 4	編曲 1
配樂 52	爆炸 16	美景 4	影音 1
變形 49	唱歌 12	影像 4	音響 1
鏡頭 43	造型 12	美術 4	光影 1
視覺 38	臨場感 11	取景 3	景象 1
螢幕 37	畫質 11	剪片 3	情景 1
機器人 35	境 9	聽覺 2	歌聲 1
背景 25	特效片 9	美術 4	古典樂 1
聲音 24	IMAX 9	剪片 3	
怪物 24	發音 9	聽覺 2	

表 40:電影相關人物類別在標記語料中次數統計

導演、演員			
主角 517	男生 24	武術 11	歌手 3
演 327	王子 22	客串 10	編導 3
拍 211	個性 21	妹 10	臉部 3
演技 188	拍成 21	功力 10	執導 2
導演 170	戲份 21	敵人 10	天份 2
動作 144	魅力 20	花栗鼠 9	舉手投足 2
演員 140	大哥 19	父親 8	肢體 2
角色 97	男人 18	年紀 8	SAM 1
長 96	拍出 18	將軍 8	外型 1
人物 57	拍片 18	船長 7	武僧 1
女生 52	公主 18	主演 7	WILL 1
卡司 52	女人 17	選角 6	醉仙 1
英雄 43	明星 14	腳色 6	一舉一動 1
狗 30	美女 13	雙胞胎 5	手勢 1
演出 30	男主角 12	指導 5	律師 1
小丑 29	妙麗 12	拍到 5	松鼠 1
表情 27	身材 11	身手 4	
配角 24	陣容 11	演成 3	

表 41:擴充的意見詞

擴充意見詞						
籠統	相遇	火辣	還給	嚴格	互動	心動
看似	亂殺	砸了	集中	吃	敷衍	飛來飛去
徹底	尷尬	無恥	震驚	濃濃續	概念	地震
適當	補充	眨眼	團結	持續	單薄	無限
緊繃	做出來	猶豫	妙麗	改編	慎選	在意
絕望	極端	不如	冷靜	串起來	環繞	長大
顯現	超乎	威風	仔細	失眠	警惕	哀哀
臭屁	打造	直接	煩悶	優美	恰如其分	巧妙
收集	諷刺	宣傳	跌破	全新	媚俗	火大
有錢	缺少	打算	走來走去	愛上	維持	夢遊
小聲	具備	鋪陳	批評	永幹	不耐煩	哪裡
可笑	快明	打完	俗套	掉合	睡到	追求
純真	成長	換成	擔心	超配	花錢	淺痛
猜想到	有效	平均	有感	深厚	表現出	快感
打發	受傷	進衡	玩味	乾淨	毛	甜蜜
容讓	拿出來	念念不忘	輸給	入迷	追殺	聳動
觀賞	播完	神經	存在	求婚	幸運	叫做
拉去	倒下	關掉	戲劇化	假婚	心疼	差不多
美麗	害羞	欺騙	溫馨	愉悅	麥可	差強人意
乖受	變換	注重	青澀	希望	保持	檢舉
不良	發現	充分	撲朔迷離	高高	毀滅	講出來
善良	重現	必要	卡通	拍好	李奧納多	了得
不同	膽小	兒童	自私	猜到	重視	啟動
不斷	反觀	放進去	有情	說服	亮眼	兩極
製片	回想	灌籃	搞混	革命	講究	停止
死光	延續	晃動	贊同	絕出	孤單	面對
說給	不已	講話	檢討	超出	呼籲	殘酷
毫無	入微	高潮	略顯	變突	死人	達到
幹嘛	客串	暗淡	三流	拍合	精準	鋪路
透露	去世	怕怕	現實	迷糊	流露	幽默
飾演	瞎搞	了解	寧願	超過	回神	個人
成為	瞧瞧	意料	絕命	用過	引人	場景
不足以	嚇死	隱藏	失去	無是	莫名其妙	離場
稱讚	表達出	截然不同	欽佩	算座	動到	誠懇
無所謂	集合	套好	難熬	賣超值	回饋	帶入
獨立	穿著	開門	細心		低能	讚嘆

不像話	處於	感受到	訝異	惋惜	看出	捨不得
操控	注意到	違背	佈置	煩死	超瞎	怎麼辦
退錢	出乎	安排	晃來晃去	爽快	頭暈	強力
拍手	刻畫	神經病	大陸	騙人	無止盡	一模一樣
噴淚	做夢	矯情	葉問	看念	吃到	無窮
相當	哭到	目不轉睛	出去	懷念	萬歲	壞掉
呼應	咬人	衝到	幸福	改變	音樂	絕無
思考	超想	中肯	竟是	超炫	拖到	帶出
請去	耍帥	作出	乾脆	無關	貫穿	鼓掌
排除	血脈	專注	回到	感人	忘不了	隨便
結果	可取	完成	打成	中英文	出場	羨慕
猜測	神奇	渴望	了無	悲觀	推推推推	不值
猜中	不少	拍下	容入	超猛	偷懶	牽動
打出來	超悶	融入	白癡	猜出	雜亂	改觀
超超	上去	新穎	放慢	演完	計較	經歷
貼近	濫用	一頭霧水	有意思	不過	深情	沉重
跳舞	讚不絕口	亂扯	預知	沮喪	礙於	靈異
帶出來	莫過於	穿插	撥完	關心	陷入	扯上
出手	汙辱	養眼	倒胃口	復古	捧腹大笑	有心
奇幻	四起	告知	打架	走人	荒誕	屬於
期望	黑暗	描繪	埋下	充滿	當紅	身材
破爛	制式化	亂拍	拿掉	放到	記憶	低頭
警示	重建	知名	悲慘	開懷	紮實	見到
不適	有失	有所	不自覺	逗趣	好好玩	丟臉
做完	美好	坐立難安	沒關係	順暢	體會到	濃厚