國立臺灣大學管理學院資訊管理學系

碩士論文

Department of Information Management
College of Management
National Taiwan University

Master Thesis

# 應用潛藏面向評分分析於中文評論：使用局部潛藏狄利克雷分配方法

# Latent Aspect Rating Analysis on Chinese Reviews: A Local LDA Based Approach

張凱迪

Kai-Ti Chang

指導教授：盧信銘 博士

Advisor: Hsin-Min Lu, Ph.D.

中華民國 103 年 7 月

July, 2014

# 誌謝

對於一個大一到大三在臺大就讀的是政治系，大四轉到經濟系然後應屆推甄上資管所的學生來說，能完成這份論文不僅內心充滿著難以言喻的悸動，更是充滿著無限的感激。首先，最要感謝的就是我的父母和兄長。曾有人說：『家庭是幸福的泉源也是痛苦的根源。』這句話我深深地認同著，也很幸運我擁有一個幸福的家庭。我是一個酷愛閱讀的人，過去一段時間我的父母除了給予我無限的關愛外更盡力滿足我在教育資源的需求，讓我可以充分享受學習的樂趣。而大哥除了是我的好朋友外更是我學習的好榜樣，推薦的歡樂三國志更是我寫論文的良伴！

除了家人的支持外，最重要的是我在研究所遇到一位非常好的老師－盧信銘博士。猶記初入研究所時，我因為醉心於校外工作對於研究事項並未投入該有的關注，但老師還是很有耐心的因材施教，讓我建立起研究者應該有的正確心態和執行方法，真的非常感謝老師的包容和耐心指導，此外，也非常感謝口試委員曹老師和李老師願意花時間給予我們建議和指導。而實驗室的夥伴更是鞭策我繼續努力的動力來源，尤其我在研究所一起打拚的夥伴（健華、承鑫）皆為第一名從大學畢業，他們的天賦和努力不但是我積極追尋的目標，更是我在各方面學習請益的對象。從實驗室的學長姐（宇泰、取向、崇瑋、久悌、如軒）和可愛的學弟妹們身上也學到很多，相信我們 BAEIR 實驗室會越來越好！

在臺大六年的求學過程中就像武俠小說一樣充滿著曲折跌宕的過程，但幸運的是總有貴人相助，不管是我的好朋友好兄弟（庭嘉、蒸籠、布希、吳承、周生、Mori、博任），還是教導過我的師長或是業界的前輩、一起工作過的同事們（Buyble、網勁科技、Mozilla），他們教會我許多事情。

最後想說的是，台灣最近政經環境紛紛擾擾，產業升級是最根本解決台灣目前經濟問題的方法。我過去和夥伴一起參與臺大創意創業學程學生會的創辦，協助臺大創聯會、NTU Garage 的成立，創辦促進學生和業界技術交流的非營利組織 HackNTU 都是希望提供校園有更好的創新創業環境，台灣是塊充滿人情味的地方，多元自由的民主社會更是值得我們用心去守護的珍貴資產。如同狄更斯在雙城記所寫的『這是最壞的時代，也是最好的時代』，有志之士們，大家一起加油！

# 中文摘要

隨著網路科技的高速發展，網路上充滿著各式各樣的評論。如何針對這些非結構的資料進行分析也顯得日漸重要。然而在這些服務或產品評論當中，往往使用者只留下對於產品或服務的整體評論分數（overall rating），而沒有針對服務或產品的各主題面向（topical aspect）做分數的評比或是揭露使用者對於產品或服務的某一種主題面向的權重（weight），這樣對於使用者的幫助有限。而藉由分析文件的主題面向分數（topical aspect rating）和其權重（weight）的問題稱為潛藏面向評分分析（Latent Aspect Rating Analysis，簡稱：LARA）。

本研究試圖使用局部潛藏狄利克雷分配（Local Latent Dirichlet Allocation，簡稱：Local LDA）和潛藏評分迴歸模型（Latent Rating Regression，簡稱：LRR）將 LARA 分析應用於中文評論上。實驗共分為兩階段模型，第一階段使用 Local LDA 將經過前處理的評論內文進行面向的切割和和面向擷取，之後第二階段運用 LRR 模型以類似 EM 算法的形式試圖推論出文件的主題面向分數（topical aspect rating）和其權重（weight）。

本研究將使用華文最大的旅遊網站攜程網旅遊評論和全球最大的旅遊評論網站 TripAdvisor 為分析資料集，其中攜程網資料為使用網路爬蟲擷取後整理而成。實驗中我們可以發現 Local LDA 的方法比起 Bootstrap 相對較好，且 Local LDA 屬於非監督式學習，毋須人工手動設定種子關鍵詞，可以讓整個應用更加廣泛。

關鍵字：文字探勘、潛在面向評分分析、潛藏狄利克雷分配、潛藏評分迴歸模型、情感分析、意見探勘、評論分析

# Abstract

As the growth of web technology, it's an important task to mine the detailed information in the online reviews. Most reviewers only rating the entity with overall rating; however, it's not enough for users to learn more from the reviews. As a result, there is a new problem called Latent Aspect Rating Analysis in text mining which analyzes latent aspect and latent aspect weight simultaneously.

In this research, we apply the LARA on the Chinese reviews. We use the Local LDA(unsupervised learning) and LRR model to analyze the online reviews. In the first stage, we use the Local LDA method on the review contexts to conduct the aspect segmentation after preprocessing. After the aspect segmentation, we can get the aspects and aspect representative words. In the second stage, we use the LRR model to infer the latent aspect rating and latent aspect weight.

Our experiment uses the Ctrip and TripAdvisor online reviews as the dataset. The results demonstrate the Local LDA + LRR method has some advantage on Chinese LARA problems.

**Keywords:** text mining, Latent Aspect Rating Analysis, Latent Dirichlet Allocation, Latent Rating Regression Model, sentiment analysis, opinion mining, review mining

# 目錄

# 圖目錄

# 表目錄

# 第一章 緒論

## 1.1 研究動機與背景

在網際網路技術的高速發展之下，使用者可以在各種社群網站（social network site）、論壇（forum）或是部落格(blog)上發表自己的意見和看法。其中使用者在電子商務網站（e-commerce）或是評論網站所留下來的大量服務或是產品的評論（review）除了是服務商提昇服務品質的有利資訊外，更是其他使用者在消費時的重要參考依據。然而，隨著行動網路和電子商務的快速發展，大量的評論資訊已經讓使用者無法輕易的辨別評論資訊的真實性和可靠性。

因此，若能夠將資訊結構化整理，將可以更方便讓使用者閱讀和理解（Angeliki Lazaridou, 2013）。過去有許多的研究對於減輕使用者的資訊負荷和進行資訊篩選方面做出努力，因而發展出資訊擷取（information extraction）、資訊統整（information summarization）、情感分析（sentiment analysis）、意見探勘 （opinion mining）等領域的研究議題。

然而在這些服務或產品評論當中，往往使用者只留下對於產品或服務的整體評論分數（overall rating），而沒有針對服務或產品的各主題面向（topical aspect）做分數的評比或是揭露使用者對於產品或服務的某一種主題面向所在乎的權重（weight）。一般而言，若只有提供整體評論分數將對於使用者在決策上幫助有所限制，例如：圖 1.1.1 為旅館評論範例，雖然兩份評論的總體評分皆為 4.8，但從評論字詞中我們可以推測得知 ID：1358586****的評論者比較喜歡旅館的設備，而 ID：1391071****的評論者比較喜歡旅館的衛生和服務。由於每個人的偏好不同，所以在乎的面向也有所不同。這也是近年來文字探勘（text mining）研究上所遇到

的問題與挑戰，而這樣藉由分析文檔的主題面向（topical aspect）和其權重（weight）的問題稱為潛藏面向評分分析（Latent Aspect Rating Analysis，簡稱：LARA）（Wang, 2010）。因此若能在分析評論同時提供使用者各主題面向的評論分數和對於主題面向的權重的話，將可以更精確的分析應用並推薦使用者適合的資訊。

過去已有相關研究針對 LARA 問題提出相關的模型方法（Wang & Lu & Zhai, 2010；Ma & Qu, 2012），但對於模型更進一步的應用以及中文評論的分析尚不全面。本研究使用全球（TripAdviser）和華文（攜程網）最大的旅遊評論網站為資料集，主要目標希望運用 Local LDA 加上潛藏評分迴歸模型 LRR 兩階段研究方法來分析網路上的旅遊評論資料（含評論文件和整體評分）。本研究將評論文件中的句子（sentences）視為文件（document）並使用標準 LDA 進行面向（aspect）的抽取以及取出每個句子的主題面向的機率分配，可達到非監督學習（unsupervised learning）的效果，提高模型的可通用性。於取出主題面向（topical aspect）和相關統計資料後，使用 LRR 進一步分析，推導出文檔的主題面向分數及其權重。總言之，希望透過本研究分析中文評論所遇到的 LARA 問題，進而對於此相關研究議題提供貢獻。

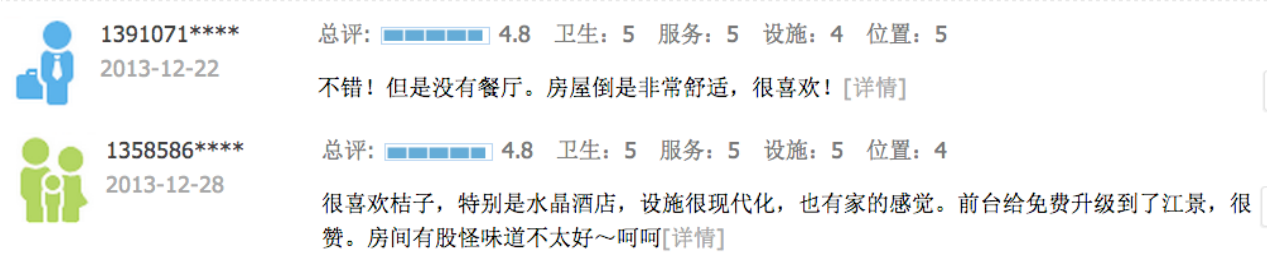



圖 1.1.1 旅館評論樣本參考（http://www.ctrip.com/）

## 1.2 研究目的

基於上述研究背景與動機，本研究希望透過 Local LDA 和 LRR 兩階段模型方法來分析網路上的中文旅遊評論資料，進而分析評論者對於評論實體（Entity）所給予的主題面向評分和權重。綜合上述，本研究希望能達成以下目的：

1. 整理過去於這個領域所提出的研究和文獻，讓之後的研究者可以參考。

2. 應用 LARA 分析於中文旅遊評論，進而分析使用者評論的主題面向評分和評分權重。

## 1.3 研究架構

本研究架構可分為六個部分，如下所述：

1. 研究背景與動機及範圍定義：確認研究動機及目標並確認研究的範圍。

2. 文獻回顧與探討：根據研究範圍蒐集 LARA 相關的文獻進行整理，分析各研究的優點和不足之處並了解相關研究方法的使用。

3. 問題定義：具體定義出本研究欲研究之問題並解釋所使用的符號和名詞定義。

4. 資料處理與系統設計：進行資料的前處理並參酌文獻定義本研究使用的 Local LDA 和 LRR 兩階段研究模型。

5. 實驗結果：定義使用的資料集並描述資料前處理之步驟和實作模型的過程，進而分析實驗結果。

6. 結論與建議：根據本研究的實驗結果和過程確認本研究的貢獻，並提供未來研究的可能方向和建議給予未來研究者參考。

# 第二章 文獻探討

本研究希望能透過 Local LDA 和 LRR 兩階段研究模型，應用於中文旅遊評論分析，進而推導出個別評論的主題面向分數和權重。本研究涵蓋範圍包括：評論的情感分析、意見分析、主題面向的評分和主題面向的擷取以及 LDA。以下各節介紹依序如下：2.1 情感分析和意見探勘研究概述、2.2 文件和句子層次的分析、2.3 主題面向層次的分析、2.4 LARA 問題研究。

## 2.1 情感分析和意見探勘研究概述

隨著網路科技的日新月異，Web 中充滿了各式各樣的非結構資料，而這些非結構資料中往往含有許多含有許多觀點和意見的資訊，例如：網路評論、推文等。由於人們在面對大量資訊時往往會受限於主觀心理偏好和客觀身體條件而產生前後不一致的結果，因此透過客觀的觀點分析和意見分析系統來協助克服主客觀限制便顯得十分重要，而過去已有大量研究投入其中，更有許多企業提供觀點分析的商業服務（Liu, 2011）。

過去研究在討論觀點分析和意見分析之前，會先定義抽象的觀點分析問題，讓問題更具結構性。圖 2.1.1 是一篇由評論網站 CNET 所擷取下來的商品評論：

(1) I love my **iPhone 5**, having purchased it after my contract was up with an android phone. (2) It's great, I love music and it produces *better* **sound** than any other phone on the market (I am an audiophile). (3) Very fast with **LTE connectivity** and **802.11n wifi**. (4) I actually like the size of the **screen**. (5) The Galaxy S4 is a great phone as well, if I had to choose I would still pick the iPhone 5. (5) But you pick what you want, it's your money.

圖 2.1.1 CNET 商品評論（http://www.cnet.com/）

根據以上的評論我們可以定義（Liu, 2011）：

1. 實體（Entity）：實體 e 可以是一個產品或是服務。如上述產品評論中，即為描述 iPhone 5 此手機的評論，其實體即為 iPhone 5 手機。實體可以具備元件和屬性表示為 e:(T, W)，T 為實體的附屬元件、W 為實體的屬性。

2. 主題面向（Aspect）：係指實體所具有的元件和屬性。例如：sound quality、size。

3. 主題面向名稱和簡稱（Aspect Name）：係指用來稱呼主題面向的名稱以及別名。例如：sound、voice 皆可以指手機的音響狀況。

4. 實體名稱和簡稱（Entity Name）：係指用來稱呼實體的名稱以及別名。例如：iPhone、apple phone。

5. 意見擁有者（Holder）：係指發表該觀點的組織或人，於這篇評論為第一人稱。

6. 意見的傾向（Polarity）：表示觀點人對於實體的偏好。一般可以表為正面（positive）、負面（negative）、中立（neutral）。例如：I actually like the size of the screen. 即對手機螢幕表示正面觀點。

7. 客觀句：陳述客觀事實的句子。

8. 主觀句：帶有個人感受或是評價的句子。

9. 意見（Opinion）：意見探勘研究可表示為<Entity, Aspect, Polarity, Opinion Holder, Time>五元組，將非結構化的資料轉化成結構化資料方便進行研究分析（Liu, 2011）。

10. 情感（Sentiment）：根據文獻人類具有愛、喜悅、驚訝、憤怒、悲傷和恐懼等情感（Parrott, 2001）。意見的強度和情感強度有關。

進行意見探勘或情感分析主要可分為三種層次：文件（Document）、句子

（Sentence）、主題面向（Aspect），隨著分析的物件粒度（granularity）越細，分析難度也相對增加。過去已有許多文獻針對不同層次做出研究，我們將於下一節分述之。

## 2.2 文件和句子層次的分析

根據過去文獻我們在上一節已經定義了情感分析的名詞和意見探勘研究的五元組表示法，讓我們可以做進一步的情感分析。

### 2.2.1 基於監督學習的文件層次情感分析

有許多文獻將文件視為分析的基本單元，將文件分為正面、負面、中立三種類別。網路上提供了許多可以作為分析的評論，而這些評論往往會附上評論者給的分數，亦可以當做分析的基準（例如：評分範圍為 1-5 顆星，1-2 顆星者視為負面，3 顆星者視為中立，4-5 顆星者視為正面）。

由於文件層次的情感分析和基於主題的文件分類（topic-based text classification）類的問題雖有所不同但十分相似，基於主題的文件分類是透過分類方法把文本分類到所設定的主題中（例如：政治、體育、財經等），所以主題的相關詞就很重要；而情感分析中重視的是如”great”、”bad”、”best”等表達正面、負面、中立的情感詞（Liu et al., 2011）。

在前人的研究當中，已有許多利用現有的監督學習方法套用到情感分析的情境上並獲得不錯的成果，例如運用樸素貝斯分類法（Naïve Bayes Classifier）、支援向量機（SVM）將評論分成正面和負面的類別（Pang et al., 2002），在這些分類的過程中，使用了許多特徵值協助做分類（例如：詞和詞頻、詞性、情感詞、意見

規則、否定語法等）（Pang et al., 2008）。除了正面、負面的情感分類外，也有許多研究聚焦在預測整體評論的評分上（Pang et al., 2005），由於分數是可以計量的，也可以轉化成迴歸類問題。

### 2.2.3 基於監督學習的句子層次情感分析

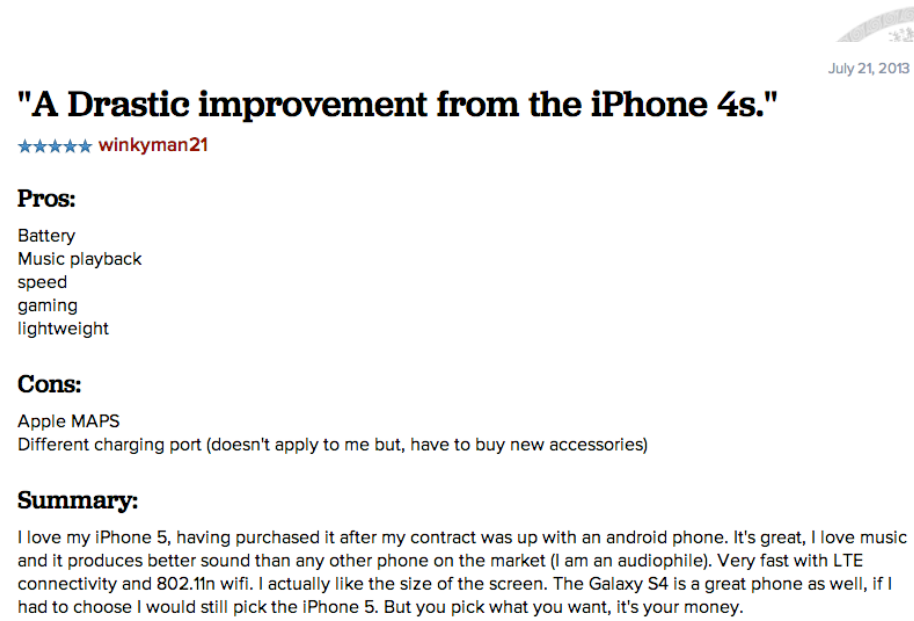
除了文件層次可以使用情感分類技術進行分類外，亦可以用於單獨的句子上。一般在進行句子的情感分類之前，必須先進行主客觀性分類，將句子分成主觀句和客觀句。若為主觀句則可以進行句子層次的情感分析，將句子分成正面、負面或是中立（Yu et al., 2003）。

## 2.3 主題面向層次的分析

一般而言評論者在評論時，往往不單只是針對一個實體的整體做評論，也會針對實體的主題面向做評論。如同圖 2.1.1 的 CNET 商品評論，評論者除了針對 iPhone 手機整體做評論外，也對音響、螢幕等主題面向做評論。因此，若能更深入針對實體的主體面向做分析，將可以提供使用者更豐富的資訊，協助使用者做決策（Liu et al., 2005）。

### 2.3.1 主題面向的擷取

針對主題的面向做擷取是進行主題面向層次的情感分析的很重要的一步。一般而言網路上的評論分為兩種形式：(1)區分評論產品優點、缺點和細節（如圖 2.3.1）、(2)讓使用者自由撰寫，較無結構性（如圖 2.3.2）。



July 21, 2013

**"A Drastic improvement from the iPhone 4s."**

★★★★★ winkyman21

**Pros:**

Battery
Music playback
speed
gaming
lightweight

**Cons:**

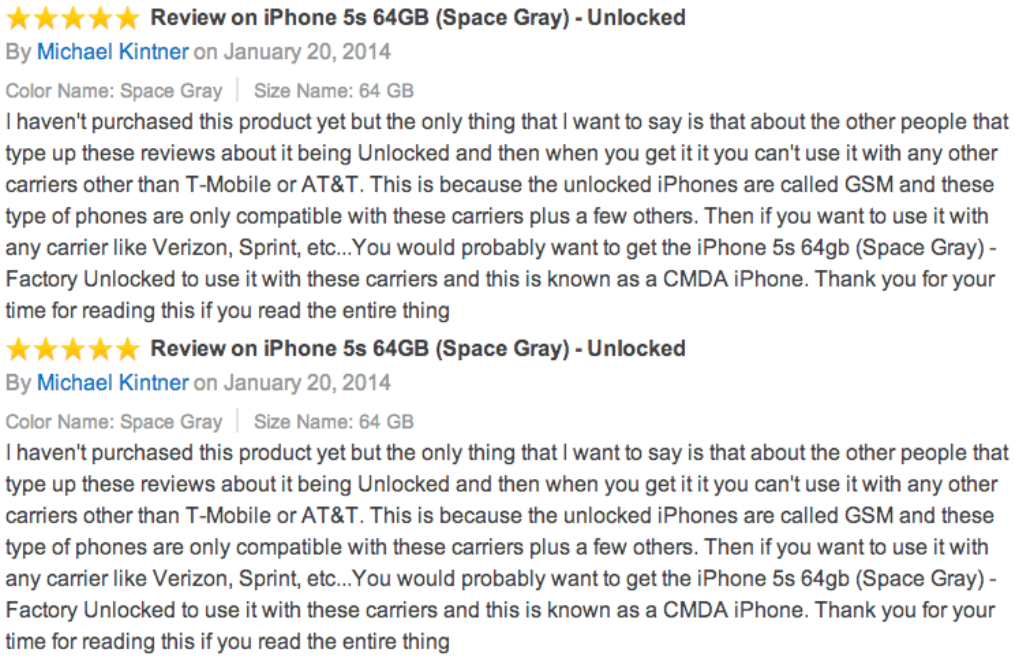
Apple MAPS
Different charging port (doesn't apply to me but, have to buy new accessories)

**Summary:**

I love my iPhone 5, having purchased it after my contract was up with an android phone. It's great, I love music and it produces better sound than any other phone on the market (I am an audiophile). Very fast with LTE connectivity and 802.11n wifi. I actually like the size of the screen. The Galaxy S4 is a great phone as well, if I had to choose I would still pick the iPhone 5. But you pick what you want, it's your money.

圖 2.3.1 區分評論產品優點、缺點和細節的評論（http://www.cnet.com/）

根據過去研究，從優缺點中擷取出主題面向可以使用資訊擷取的技術，例如：隱含馬可夫模型（HMM）（Lafferty et al., 2001）、條件隨機域（CRF）（Freitag et al., 2000）。



★★★★★ **Review on iPhone 5s 64GB (Space Gray) - Unlocked**

By Michael Kintner on January 20, 2014

Color Name: Space Gray | Size Name: 64 GB

I haven't purchased this product yet but the only thing that I want to say is that about the other people that type up these reviews about it being Unlocked and then when you get it it you can't use it with any other carriers other than T-Mobile or AT&T. This is because the unlocked iPhones are called GSM and these type of phones are only compatible with these carriers plus a few others. Then if you want to use it with any carrier like Verizon, Sprint, etc...You would probably want to get the iPhone 5s 64gb (Space Gray) - Factory Unlocked to use it with these carriers and this is known as a CMDA iPhone. Thank you for your time for reading this if you read the entire thing

★★★★★ **Review on iPhone 5s 64GB (Space Gray) - Unlocked**

By Michael Kintner on January 20, 2014

Color Name: Space Gray | Size Name: 64 GB

I haven't purchased this product yet but the only thing that I want to say is that about the other people that type up these reviews about it being Unlocked and then when you get it it you can't use it with any other carriers other than T-Mobile or AT&T. This is because the unlocked iPhones are called GSM and these type of phones are only compatible with these carriers plus a few others. Then if you want to use it with any carrier like Verizon, Sprint, etc...You would probably want to get the iPhone 5s 64gb (Space Gray) - Factory Unlocked to use it with these carriers and this is known as a CMDA iPhone. Thank you for your time for reading this if you read the entire thing

圖 2.3.2 讓使用者自由撰寫，較無結構性的評論（http://www.amazon.com/）

Hu et al. (2004) 則針對自由撰寫，較無結構性的評論提出分析的方式：

(1) 找出較高頻率的名詞和名詞片語

在評論時人們的用詞表是收斂的，不重要的內容往往會發散，而反覆出現的詞彙就是真正重要的主題面向。因此找出高頻率出現的名詞是發現主題面向的第一步。

(2) 利用主題面向與情緒詞的關係找出遺漏的的主題面向

經過找出高頻率的名詞後，事實上還遺漏了那些不頻繁出現但卻是重要的主題面向。而那些出現頻率不高但卻和情感詞時常一起出現的名詞就可能是潛在的主題面向。

### 2.3.2 基於主題面向的情感分析

過去文獻提出了將商品的主題面向取出並分析屬於該主題面向的摘要和句子的情感傾向。如圖 2.3.3 所示，將 Digital_camera_1 這個商品分成 picture quality、size 等面向，各自面向含有針對該面向發表的正面或是負面的評論，實務上可以更方便使用者進行閱讀（Hu et al., 2004）。

*Digital_camera*_1:
Feature: **picture quality**
Positive: 253
<individual review sentences>
Negative: 6
<individual review sentences>
Feature: **size**
Positive: 134
<individual review sentences>
Negative: 10
<individual review sentences>
…

圖 2.3.3 主題面向層次情感分析摘要範例（Hu et al., 2004）

### 2.3.3 基於主題面向的評分分析

針對主題面向的評分分析，有研究提出了在抽取主題面向的同時也運用主題面向評分當做外顯變數（explicitly variable），並使用 ground truth 進行迴歸模型預測（Titov et al., 2008）。也有研究針使用整體評論評分和商品評論推論出面向評分和面向的摘要，希望能提供使用者更多元的資訊（Yue et al., 2009）。

## 2.4 LARA 問題研究

### 2.4.1 基於 Bootstrap 的 LARA 分析

雖然已有許多研究針對主題面向層次的情感分析做出了貢獻，但過去的研究多半只關心各個面向的整體狀況，而忽略了個別評論以及個別評論者之間的差異 ，且並不把面向評分視為潛藏變數。因此首先有研究提出了 LARA 的問題和解決方式，希望透過結合類似 Bootstrap 方法和 LRR 生成模型的兩階段的步驟可以同時推論出主題面向評分和評論者對於各面向所佔的權重（Wang, 2010）。

在進行主題面向評分和評論者對於各面向所佔的權重推論之前，我們必須先進行主題面向的擷取和分割，將句子和詞指定給所屬於的主題面向。在這篇文獻中（Wang, 2010），使用的是 Bootstrap 形式的分割方式，其演算法解釋如下：

在進行擷取主題面向演算法之前，需要先手動設定主題面向數量（K）和初始種子的關鍵字，於本篇文獻中因應資料集的 ground truth 共設計有 7 個主題面向。

表 2.4.1 主題面向分割演算法（Wang, 2010）

| 主題面向分割演算法（Aspect Segmentation Algorithm） |
| --- |
| 輸入：評論文件的集合$\{d_1, d_{2,\ldots,} d_{\lvert D\rvert}\}$、主題面向關鍵字的集合$\{T_1, T_{2,\ldots,} T_K\}$，令 |

| 字彙集合為$V$，閥值為$p$，迭代次數設為$I$ | |
|---|---|
| 第 0 步 | 將所有評論文件的集合切割成句子，令句子集合$X = \{x_1, x_2, \dots x_M\}$ |
| 第 1 步 | 每個句子若有配對到關鍵字的集合中的種子關鍵字的話，將該句子所對應的面向$Count(i)$記錄加一 |
| 第 2 步 | 將句子指定給記錄最大的面向$argmax_i\ Count(i)$，若有相同，則指定多個面向 |
| 第 3 步 | 計算每個詞彙（token）和面向（aspect）之間的相依性 $\chi^2$ |
| 第 4 步 | 取出每個面向（aspect）中$\chi^2$分數排名前$p$的詞彙（token），將其加入主題面向關鍵字的集合$T_i$ |
| 第 5 步 | 如果面向關鍵字（aspect keyword ）集合不變或是迭代超過$I$則進到第六步，否則回到第一步 |
| 第 6 步 | 輸出 sentences 和它所指定的 aspect |

在演算法中，我們利用計算詞彙（token）和面向（aspect）之間的 Chi-Square（$\chi^2$），來求得兩者之間的相依性，取出相依性前$p$名的 token 加入 keyword 的集合中，讓有代表 aspect 特性的 token 可以被擷取出來（Yang, 1997）。

$$\chi^2\left(\omega, \mathrm{A}_i\right) = \frac{C\times(C_1C_4-C_2C_3)^2}{(C_1+C_3)\times(C_2+C_4)\times(C_1+C_2)\times(C_3+C_4)} \qquad \text{公式 2.4.1}$$

公式參數定義如下：

$C$： $\omega$在所有 sentences 出現的次數

$C_1$：sentences 屬於$A_i$且$\omega$出現在 sentences 中 sentences 的數量

$C_2$：sentences 不屬於$A_i$且$\omega$出現在 sentences 中 sentences 的數量

$C_3$：sentences 屬於$A_i$且$\omega$未出現在 sentences 中 sentences 的數量

$C_4$：sentences 不屬於$A_i$且$\omega$未出現在 sentences 中 sentences 的數量

經過了 Bootstrap 的 Aspect Segmentation 演算法後，我們可以得到每個 document 都有一個$k \times n$的特徵矩陣$W_d$，其中 d 是指 document 的 index，而 i 是指 aspect 的 index，j 為 token 的 index。$W_{dij}$代表第 d 個 document 中，屬於 aspect $i$的第$j$個 token 出現的頻率，這邊用屬於 aspect $i$ 所有 token 總數進行標準化。

在進行第二階段的 LRR 模型之前，文獻假設評論者評分的行為定義為，如圖 2.4.1 所示：

（1） 當一個使用者要給予一個實體評分時，他會先決定他所希望評論的面向（aspect），然後決定代表他意思的詞彙。而這個詞彙對應了相對的情緒傾向

（2） 把所有針對這個主題面向所用的詞彙和它對應的情緒權重相加總後就會得到該主題面向的評分。而每個評論者對於各個面向（aspect）則有不同重視程度因而會給予不同的權重（weight）

（3） 將所有 aspect 的評分（rating） 和評論者所給予的權重（weight）相乘加總後，就會得到評論整體的分數（overall rating）

Wang et al. (2010) 為了成功捕捉上述使用者在評分時的行為而提出了 LRR 迴歸模型，LRR 是一個生成模型（generative model），在上一階段對於每一個 document 都有一個標準化過的列為面向（aspect）、行為詞彙（token）的頻率特徵矩陣。在 LRR 模型中將特徵矩陣$W_d$當做獨立變數，而整體評分$r_d$（overall rating）則當成預測的應變數。

為了能夠建立可以推論出 aspect rating 和 aspect weight 的模型，所以 LRR 模型不直接由特徵矩陣$W_d$決定，而是由一組潛在的面向分數（latent aspect rating）所預測，特徵矩陣$W_d$則直接預測潛在的面向分數（latent aspect rating）。由於已知有 k 個 aspect，同樣的每個文件（document）也會有 k 個潛在的面向分數（latent aspect rating）和 k 個潛在面向的權重（aspect weight）且總共有 n 個 unique token，所以將潛在面向分數（latent aspect rating）表示成線性的組合：

$$s_i \leftarrow \sum_{j=1}^{n} \beta_{ij} W_{dij} \quad \text{公式 2.4.2}$$

其中$\beta_i \in \Re$為詞彙表在$A_i$的情感傾向。

接下來透過潛在的面向分數（latent aspect rating）和潛在面向的權重（aspect weight）的加總可以產生整體評分（overall rating），表示成$\alpha_d{}^T s_d = \sum_{i=1}^{k} \alpha_{di} s_{di}$。

為了能夠模擬預測整體評分（overall rating）的不確定性，假設整體評分（overall

rating）是從平均數為$\alpha_d{}^T s_d$，變異數為$\delta^2$的高斯分配（Gaussian distribution）所抽取出來，表示為：

$$r_d \sim N(\sum_{i=1}^{k} \alpha_{di} \sum_{j=1}^{n} \beta_{ij} W_{dij,}\ \delta^2) \qquad \text{公式 2.4.3}$$

欲建構評論（review）的內容以及整體評分（overall rating）的關係，以下做更進一步的探討，發現評論者在針對不同面向給予潛在面向的權重（aspect weight）時有以下特性：

（1） 不同的評論者偏好不同，所以在乎的面向也有所不同。（例如：商務旅客可能比較在乎網路或是商務設備的完善，但新婚夫妻可能在乎的是服務或是房間的氣氛）。

（2） 不同的面向並非獨立，而會有重疊的情況。（例如：在乎乾淨（cleanliness）面向的評論者他有可能也會在乎房間（room）面向）。

因此文獻中為了考慮偏好的差異性，假設每個文件（document）中的潛在面向權重（aspect weight）為從整個文集（corpus）的先驗（prior）分佈所產生的一組隨機變數。而為了捕捉不同面向的相依性，假設潛在面向權重（aspect weight）的先驗分佈為多變量高斯分佈（multivariate gaussian distribution），其中$\mu$和$\Sigma$分別為其平均值和變異數，如下所示：

$$\alpha_d \sim N(\mu, \Sigma) \quad \text{公式 2.4.4}$$

合併 2.4.3 和 2.4.4 將問題轉化成一個貝氏迴歸問題（Bayesian regression）。給定評論文件（document）下觀察值為給定文件（document）的整體評分（overall rating）的機率如下：

$$P(r|d) = P(r_d|\mu, \Sigma, \delta^2, \beta, W_d) \quad \text{公式 2.4.5}$$

$$= \int p(\alpha_d|\mu, \Sigma) p(r_d|\sum_{i=1}^{k} \alpha_{di} \sum_{j}^{n} \beta_{dij} W_{dij}, \delta^2) d\alpha_d$$

其中$r_d$和$W_d$為文件中已知的觀察值，而文獻假設$\delta^2$與$\beta$獨立於個別的評論（review），故$\Theta = (\mu, \Sigma, \delta^2, \beta)$為文集層次的模型變數（corpus-level model parameters），整個模型示意圖如圖 2.4.1：

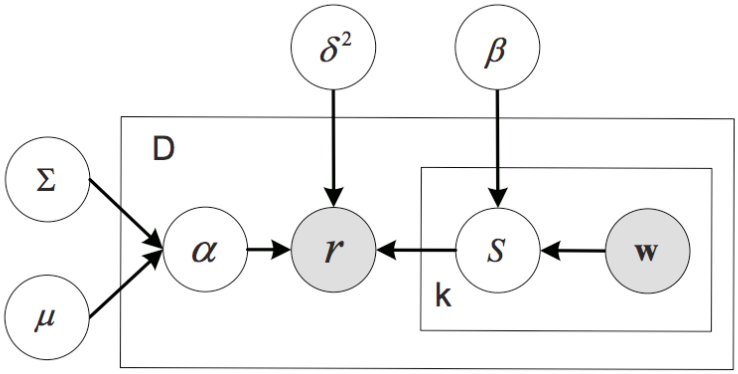



圖 2.4.1 LRR 模型示意圖（Wang et al., 2010）

LRR 模型最終的目標是希望透過給定整體評分（overall rating）和評論的內容（review content）能夠推論出潛在面向的權重（aspect weight）和潛在的面向分數（latent aspect rating）。而推論方法如下：

（1）每個文件（document）擁有的潛在面向分數（latent aspect rating）可由以定義好的 2.4.2 來計算。

（2）每個文件（document）的潛在面向的權重（latent aspect weight）則運用最大後驗機率概似法（maximum a posterior，簡稱：MAP）來計算最有可能的面向的權重（aspect weight）。其目標函數如下：

$$\mathcal{L}(d) = \log p(\alpha_d|\mu, \Sigma)p(r_d|\sum_{i=1}^{k}\alpha_{di}\sum_{j=1}^{n}\beta_{dij}W_{dij}, \delta^2) \quad \text{公式 2.4.6}$$

對應於每個文件（document）面向的權重（aspect weight），由於$\alpha_d$和$r_d$由多變量高斯分佈和高斯分佈所生成，可以將上述式子展開為：

$$\widehat{\alpha_d} = argmax\mathcal{L}(\alpha_d)$$

$$= argmax[-\frac{(r-\alpha_d{}^T s_d)^2}{2\delta^2} - \frac{1}{2}(\alpha_d - \mu)^T \Sigma^{-1}(\alpha_d - \mu)]$$

假設限制條件如下：

$$\sum_{i=1}^{k}\alpha_{di} = 1 \ ; 0 \le \alpha_{di} \le 1 \ for \ i = 1,2,\dots k \quad \text{公式 2.4.7}$$

文獻中提供了共軛梯度下降法（conjugate-gradient-interior-point），求極大值：

$$\frac{\partial\mathcal{L}(\alpha_d)}{\partial\alpha_d} = -\frac{(\alpha_d{}^T s_d - r_d)s_d}{\delta^2} - \Sigma^{-1}(\alpha_d - \mu) \quad \text{公式 2.4.8}$$

更進一步文獻使用最大概似估計法（Maximum Likelihood）找出最佳化的$\widehat{\Theta} = (\hat{\mu}, \widehat{\Sigma}, \widehat{\delta^2}, \hat{\beta})$以最大化給定評論文件（document）下觀察值為給定文件（document）的整體評分（overall rating）的機率。

對於整體評論的 log-likelihood 函數和 ML 估計式分述如下：

$$\mathcal{L}(D) = \sum_{d\epsilon D}\log p(r_d|\mu, \Sigma, \delta^2, \beta, W_d) \quad \text{公式 2.4.9}$$

$$\hat{\Theta} = arg\max_{\Theta}\sum_{d\in D}\log p(r_d|\mu, \Sigma, \delta^2, \beta, W_d) \quad \text{公式 2.4.10}$$

為了能最佳化問題，此處研究使用了類似 EM 演算法的方式，並於迭代開始

之前，先隨機初始化$\Theta_{(0)} = (\mu, \Sigma, \delta^2, \beta)$：

（1）E 步驟：已知參數可以藉由推論公式 2.4.2 和 2.4.7 求得每個文件的潛在面向分數（latent aspect rating）和潛在面向的權重（aspect weight）。

（2）M 步驟：透過求出的潛在面向分數（latent aspect rating）和潛在面向的權重（aspect weight）去更新（update）模型參數$\Theta = (\mu, \Sigma, \delta^2, \beta)$並透過極大化 complete likelihood 來得到$\Theta_{(t+1)} = (\mu, \Sigma, \delta^2, \beta)$，持續執行 E 和 M 的步驟，最後參數會收斂達到終止條件。

### 2.4.2 基於 Local LDA 的 LARA 分析

在 Wang et al. (2010) 提出用 Bootstrap 和 LRR 模型兩階段方法解決 LARA 問題之後，Ma et al.（2012）為了解決第一階段 Bootstrap 需要事先由人工決定面向種子詞的限制，改為運用 Brody et al.（2010）所提出的 Local LDA 的方式，將評論當中的句子（sentences）視為標準 LDA 模型（Blei & Ng & Jordan, 2003）中的文件（document），並將抽取出的潛在主題（latent topic）和主題詞彙（vocabulary）視為面向（aspect）的切割與代表詞的擷取，亦即將句子（sentences）分配到的主題視為屬於該面向（aspect）。

**Latent Dirichlet allocation 模型介紹**

LDA 模型是一個完整的生成模型（generative model）與主題模型（topic model）架構，其修正了之前的主題模型（例如：Latent Semantic Analysis（潛藏語意分析）、Probabilistic Latent Sematic Analysis（機率式潛藏語意分析））的缺點，解決了 PLSA

沒有辦法直接將機率分配給先前未出現(unseen)的文件；以及參數數量會隨著文件數量線性擴增等問題。

主題模型假設每一篇文件隱含了一個或多個主題，每個主題擁有特定詞彙的機率分布（probability distribution），而每篇文件是由這機率分布之下的詞彙所組成。LDA 是主題模型的一種，其假設每篇文章可能由多個主題所組成，故每篇文章擁有自己的主題機率分布，而每個主題擁有該主題下詞彙的機率分布，由這兩個機率分布決定了文件的組成內容。

用通俗的方式描述 LDA 的模型則為：

（1）一位大文豪欲寫 M 篇文章，共涉及了 K 個主題，每個主題下的詞分布為從參數為$\beta$的狄利克雷（Dirichlet）先驗分布中隨機抽樣出長度為 K 的多變量（Multinomial）分佈。

（2）對於每篇文章，他會從泊松分佈中隨機抽取一文章長度的值。

（3）再從參數為$\alpha$的狄利克雷（Dirichlet）先驗分布中隨機抽樣出長度為 M 的多變量（Multinomial）分佈當做該文章每個主題出現的機率分佈。

（4）當文豪想寫第 m 篇文章的第 n 個字時，首先先從該文章中每個主題出現的多變量機率分佈中抽取一個主題，再從這個主題所對應的多變量機率分佈中隨機取出想寫的詞。

（5）不斷重複隨機生成的過程，直到把 M 篇文章都寫完。

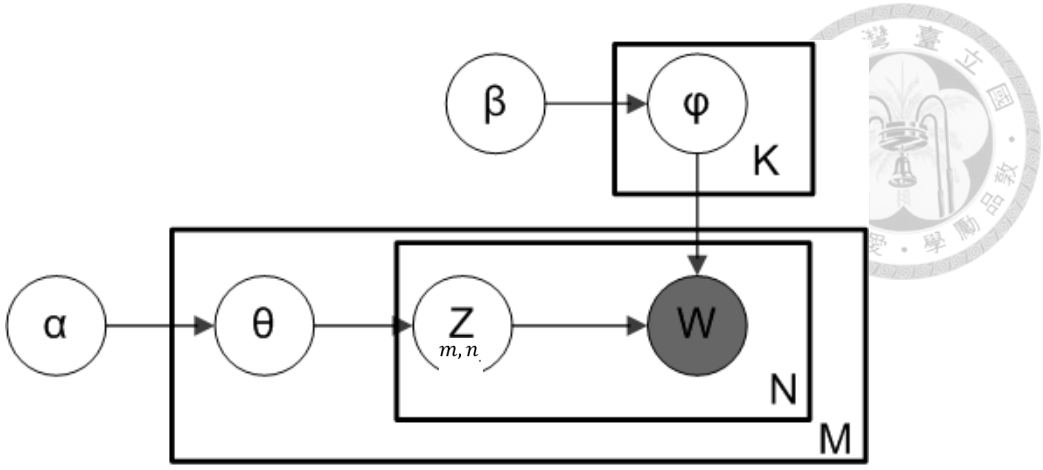



圖 2.4.2 LDA 模型示意圖（http://en.wikipedia.org/）

LDA 的模型架構如圖 2.4.2。此處假設 m 為文件的索引，n 為詞彙的索引。圖中 K 為主題個數，M 為文件總數，N 為第 m 篇文件的總字數。模型假設文件的主題先驗分佈（prior probability distribution）與主題的詞彙先驗機率分佈符合狄利克雷分配（Dirichlet Allocation）。其中$\alpha$為產生每個文件下主題多項分佈的 Dirichlet 先驗參數，以$\theta$向量表示主題先驗機率分佈，向量長度為 K；而$\beta$為產生每個主題下詞彙多項分佈的 Dirichlet 先驗參數，以$\phi$向量表示詞彙先驗機率分佈。

若文集（corpus）所有組成字以$W$向量表示，對應的主題變數以$Z$向量表示，其生成機率（generative probability）為：

$$P(W,Z,\theta,\phi|\alpha,\beta) = \prod_{k=1}^{K} P(\phi^{(k)}|\beta) \prod_{d=1}^{D} P(\theta^{(d)}|\alpha) \prod_{i=1}^{N_d} P(z_i{}^{(d)}|\theta^{(d)}) P(w_i{}^{(d)}|\phi^{(z_i{}^{(d)})})$$

公式 2.4.11

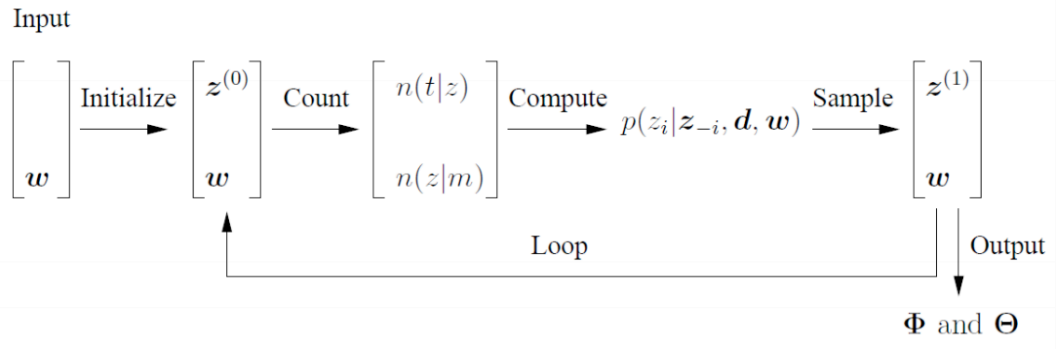



圖 2.4.3 透過 Gibbs sampling 進行 LDA 過程（Wang, 2008）

過去研究指出由於$\theta$和$\phi$都牽涉到潛在變數，LDA 在使用精確估計（exact inference）上並不容易實現，目前較常使用概似估計（approximate inference）的方法，例如：變形概似估計（Variational Approximation），馬可夫鍊蒙地卡羅法（Markov chain Monte Carlo）等方法（Blei et al., 2003）。

Gibbs Sampling 可視為 Markov-Chain Monte Carlo 演算法的特例，根據 Griffiths et al.（2004）的研究，其方法如下所述：初始值為給定每個詞彙隨機主題$z^{(0)}$，然後統計每個主題$z$下出現詞彙$t$的數量以及每個文件$m$下出現在主題$z$中的詞彙數量。每一次排除目前的詞彙計算$p(z_i|z_{-i}, d, w)$，根據所有其他詞的主題分配估計目前詞分配各主題的機率。當得到目前詞彙屬於所有主題$z$的機率分佈後，根據此機率分佈為該詞彙隨機取出一個新的主題，然後用同樣的方式持續迭代更新下一個詞的主題$z^{(1)}$，直到每個文件下主題分佈$\theta$和每個主題下詞彙分佈$\phi$收斂為止。

我們的目標$P(Z|W) = \frac{P(Z,W)}{P(W)} \propto P(Z,W)$，可以透過分別對$P(W|Z)P(Z)$前後項

積分，完成吉伯斯抽樣（Gibbs sampling），使得整條馬可夫鍊的主題變數最終收斂趨於穩定，並可藉此估計$\phi$和$\theta$參數：

$$\hat{\phi}_v^{(k)} = \frac{n_{(-i)}^{(.)} + \beta}{n_{(-i)}^{(.)} + V\beta}$$

$$\hat{\theta}_k^{(d)} = \frac{n_{(.)}^{(d)} + \alpha}{n_{(.)}^{(d)} + K\alpha}$$

其中 n 表示計數（count），其對應的左上標為計數的範圍，若以（.）表示則為全部範圍，即整個文集；左下標用來標記是否排除第 i 個位置，若以（.）表示則為將所有位置納入計數範圍；右上標是指定要計數的詞彙，若以（.）表示則為將所有詞彙納入計數範圍；右下標是指定要計數的主題，若以（.）表示則為將所有主題納入計數範圍。所以，上面$\phi$公式的$n_{(-i)}^{(.)}$是除了位置 i 外，$Z$向量出現在主題 k 的次數；而$\theta$公式的$n_{(.)}^{(d)}$是 d 文件出現在主題 k 的次數。有了上述估計值，可根據現有的 W, Z 條件進行新文件的主題機率分配的預測。

## 2.5 小結

上面幾個小節我們討論了前人在意見探勘（opinion mining）和情感分析（sentiment analysis）方面的研究。從最早的文件層次的分類問題到後來的整體評論評分分析再到主題面向的分析都希望能夠透過找到一個好的方式去分析這些大量的非結構資料，最後我們討論到分析 LARA 問題，希望能夠藉由整體評論內容

和整體評論分數推論出每個文件（document）中的潛在面向權重（aspect weight）和潛在面向分數（latent aspect rating），提供給使用者更多的決策資訊。本研究主要參酌了 Ma et al.（2012）所應用的 Local LDA 模型和 Wang et al. (2010)所提出的 LRR 模型進行調整，希望能解決中文評論遇到的 LARA 問題。

# 第三章 問題定義與系統設計

## 3.1 問題定義

本研究使用網路上蒐集的旅遊評論當做資料集，以旅館當做研究討論的實體（Entity）。在輸入項當中，資料集提供了針對實體的評論內容（content text）和整個實體的整體評分（overall rating），結合 LDA 模型的面向（aspect）抽取和 LRR 模型的分析，進而運用這些資料推導出文件的潛在面向分數（latent aspect rating）和潛在面向權重（aspect weight）。

令 $D=\{d_1,d_2,d_3...,d_{|D|}\}$ 為評論文件集合，包含了欲分析的文件實體（Entity）。而每一個文件 $d\in D$ 都擁有整體評論分數 $r_d$，$r_d\in[r_{\min},r_{\max}]$（介於最大值和最小值之間，本研究中假設 $r_{\min}=1$，$r_{\max}=5$）。$V=\{w_1,w_2,...,w_n\}$ 則是由文件集合 D 中取出的字彙。令 $X=\{sentence_1,sentence_2,...,sentence_m\}$ 為從文件集合 D 中取出的句子。

我們假設評論者（reviewer）在評論一件實體（Entity）時會先決定要評論實體（Entity）的哪一個面向（aspect），然後選擇適合的用詞進行評論並根據這些字詞給予這個面向（aspect）評定分數。最後，評論者根據不同面向在乎的程度，給予面向（aspect rating）評定分數不同的權重（weight）進而組合成一個整體的評論分數值（overall rating）。

因此，我們假設文件中有 $k$ 個面向（aspect），令 $s_d$ 為 d 文件面向評定分數（aspect rating），$s_{di}$ 則為 d 文件的第 $i$ 個面向評定分數（aspect rating），視為潛藏變數（latent variable）。同時我們也要考慮權重的部份，令 $\alpha_d$ 為面向的權重（weight），一樣為 $k$ 維度的向量，$\alpha_{di}$ 則表示 d 文件的第 $i$ 個面向評定分數的權重（weight）。假設

$\alpha_{di} \in [0,1]$ 和 $\sum_{i=1}^{k} \alpha_{di} = 1$。

## 3.2 系統設計

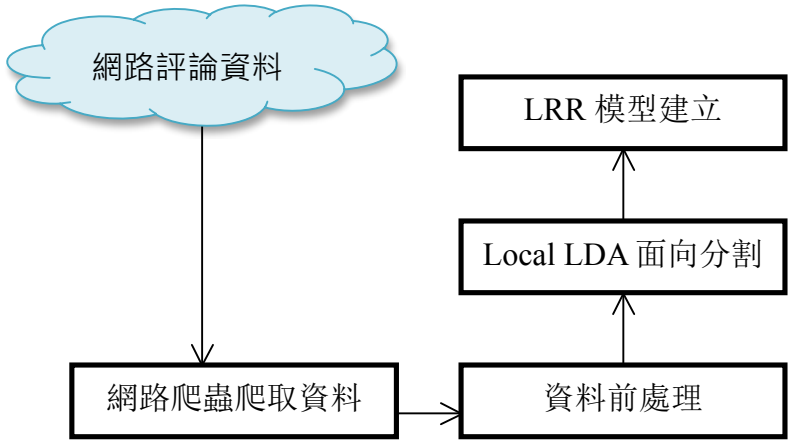



圖 3.2.1 系統設計示意圖

前一小節定義了我們的研究問題和名詞假設，此處我們將更進一步定義我們的模型。我們參酌了 Brody et al.（2010）所提出的 Local LDA 模型和 Wang et al. (2010)所提出的 LRR 模型進行調整，進行中文評論的 LARA 問題分析，以下詳述模型建立過程：

第一階段我們先使用運用 Local LDA 的方式將前處理完(有 n 個 unique token）的評論當中的句子（sentences）視為標準 LDA 模型（Blei & Ng & Jordan, 2003）中的文件(document)，並將抽取出潛在的主題(latent topic)和主題詞彙(vocabulary）視為面向（aspect）的切割與面向字詞（aspect word）的擷取結果，亦即將句子（sentences)分配到的主題視為屬於該面向(aspect)，而後並對每個文件(document）都建立一個 k×n 的特徵矩陣（元素已使用屬於面向（aspect） i 所有詞彙（token）總數進行標準化）。

在進行非監督學習之前，需要決定總共要將資料分成幾群才是最佳的狀況，亦即先確認所欲抽取的面向（aspect）數量 k，文獻中 Ma et al.（2012）運用了 Niu et al.（2007）所提出的 consistency function 去進行 cluster validation，每次完成 Local LDA 後會將每一個句子（sentences）標註一個屬於的面向（aspect），希望透過每次檢驗標註結果的一致性去決定最佳的 k 值，具體演算法和公式如下：

$$F\left(C, C^{'}\right) = \frac{\sum_{i,j} 1\{C_{i,j} = C^{'}_{i,j} = 1, d_i, d_j \epsilon S^{'}\}}{\sum_{i,j} 1\{C_{i,j} = 1, d_i, d_j \epsilon S^{'}\}} \quad \text{公式 3.2.1}$$

其中定義$C$和$C^{'}$為行列為句子（sentences）的 connectivity matrix，其內容值為若行列對應兩者 sentences 所屬的 aspect 相同則值為 1

驗證步驟：

表 3.2.1 cluster validation 驗證步驟（Niu et al., 2007）

| |
|---|
| （1）執行參數為 k 的 Local LDA 模型於 sentences $X$獲得 connectivity matrix $C_k$ |
| （2）建立均勻分佈隨機指定給$X$元素的 connectivity matrix $R_k$ |
| （3）隨機從$X$大小為$\xi\|X\|$的子集$X^{'}$ |
| （4）執行參數為 k 的 Local LDA 模型於 sentences $X^{'}$獲得 connectivity matrix $C_k^{'}$ |
| （5）建立均勻分佈隨機指定給$X$元素的 connectivity matrix $R_k^{'}$ |
| （6）根據公式 3.2.1 計算$score_{\xi}(k) = F\left(C_k^{'}, C_k\right) - F(R_k^{'}, R_k)$ |

（7）重複$q$次步驟 3 到 6

（8）$q$次迭代後回傳平均分數，取最大者為決定的 k 值

圖 3.2.2 為演算法步驟求出的 connectivity matrix 示意圖：

| sentences | $S_1$ | $S_2$ | … | $S_m$ |
|---|---|---|---|---|
| $S_1$ | 1 | 1 | 0 | 0 |
| $S_2$ | 0 | 1 | 1 | 1 |
| … | 1 | 0 | 1 | 0 |
| $S_m$ | 1 | 1 | 0 | 1 |

圖 3.2.2 connectivity matrix 示意圖

第二階段我們使用 Wang et al. (2010)所提出的 LRR 模型進行調整：

為了能夠建立可以推論出 aspect rating 和 aspect weight 的模型，所以 LRR 模型不直接由特徵矩陣$W_d$決定，而是由一組潛在的面向分數（latent aspect rating）所預測，而特徵矩陣$W_d$則直接預測潛在的面向分數（latent aspect rating）。由於已知有 k 個 aspect，同樣的每個文件（document）也會有 k 個潛在的面向分數（latent aspect rating）和 k 個潛在面向的權重（aspect weight）且總共有 n 個 unique token，所以將潛在面向分數（latent aspect rating）表示成線性的組合：

$$s_i \leftarrow \sum_{j=1}^{n} \beta_{ij} W_{dij} \quad \text{公式 3.2.2}$$

其中$\beta_i \in \Re$為 unique token 在$A_i$的情感傾向。由圖 3.2.3、圖 3.2.4 示意圖可以得知特徵矩陣$W_d$和$\beta_i$情感傾向為$k \times j$的矩陣。圖 3.2.5 呈現$i$列的潛在面向分數（latent aspect rating）的示意圖。

| | $W_1$ | $W_2$ | … | $W_j$ |
|---|---|---|---|---|
| $A_1$ | 0.0113 | 0.0723 | | 0.0413 |
| $A_2$ | 0.0023 | 0.0215 | | 0.0113 |
| … | | | | |
| $A_i$ | 0.0343 | 0.0223 | | 0.0113 |

圖 3.2.3 特徵矩陣$W_d$示意圖

| | $W_1$ | $W_2$ | … | $W_j$ |
|---|---|---|---|---|
| $A_1$ | -0.01113 | 0.0723 | | 0.0213 |
| $A_2$ | 0.0013 | 0.0215 | | -0.0313 |
| … | | | | |
| $A_i$ | 0.0373 | -0.0123 | | 0.0413 |

圖 3.2.4 詞彙情感傾向$\beta$示意圖

| | |
|---|---|
| $s_1$ | 0.001 |
| $s_2$ | 0.34 |
| $s_3$ | 0.5 |
| … | |
| $s_i$ | 0.1 |

圖 3.2.5 面向評分$s$示意圖

接下來透過潛在的面向分數（latent aspect rating）和潛在面向的權重（aspect weight）的加總可以產生整體評分（overall rating），表示成$\alpha_d{}^T s_d = \sum_{i=1}^{k} \alpha_{di} s_{di}$。

為了能夠模擬預測整體評分(overall rating)的不確定性，假設整體評分(overall rating）是從平均數為$\alpha_d{}^T s_d$，變異數為$\delta^2$的高斯分配（Gaussian distribution）所抽取出來，表示為：

$$r_d \sim N(\sum_{i=1}^{k} \alpha_{di} \sum_{j=1}^{n} \beta_{ij} W_{dij,}\ \delta^2) \qquad \text{公式 3.2.3}$$

為了考慮偏好的差異性，假設每個文件(document)中的潛在面向權重(aspect weight）為從整個文集（corpus）的先驗（prior）分佈所產生的一組隨機變數。而為了捕捉不同面向的相依性，假設潛在面向權重（aspect weight）的先驗分佈為多變量高斯分佈，其中$\mu$和$\Sigma$分別為其平均值和變異數，公式和示意圖 3.2.6 如下所

示：

$$\alpha_d \sim N(\mu, \Sigma) \quad \text{公式 3.2.4}$$

| | |
|---|---|
| $\alpha_1$ | 0.001 |
| $\alpha_2$ | 0.34 |
| … | |
| $\alpha_i$ | 0.1 |

圖 3.2.6 面向權重$\alpha$示意圖

合併 3.2.3 和 3.2.4 將問題轉化成一個貝氏迴歸問題（Bayesian regression）。給定評論文件(document)下觀察值為給定文件(document)的整體評分(overall rating)的機率如下：

$$P(r|d) = P(r_d|\mu, \Sigma, \delta^2, \beta, W_d) \quad \text{公式 3.2.5}$$

$$= \int p(\alpha_d|\mu, \Sigma) p(r_d|\sum_{i=1}^{k} \alpha_{di} \sum_{j}^{n} \beta_{dij} W_{dij}, \delta^2) d\alpha_d$$

其中$r_d$和$W_d$為文件中已知的觀察值，而文獻假設$\delta^2$與$\beta$獨立於個別的 review，故$\Theta = (\mu, \Sigma, \delta^2, \beta)$為文集層次的模型變數（corpus-level model parameters），整個模型示意圖如下：

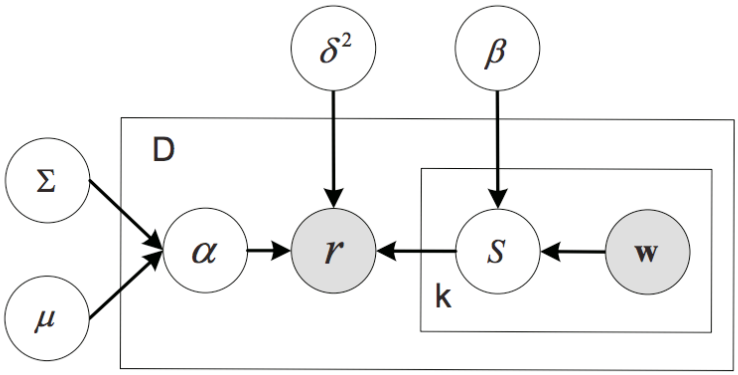



圖 3.2.6 LRR 模型示意圖（Wang et al., 2010）

LRR 模型最終的目標是希望透過給定整體評分（overall rating）和評論的內容（review content）能夠推論出潛在面向的權重（aspect weight）和潛在的面向分數（latent aspect rating）。而推論方法如下：

（1）每個文件（document）擁有的潛在面向分數（latent aspect rating）可由以定義好的 3.2.2 來計算。

（2）每個文件（document）的潛在面向的權重（aspect weight）則運用最大後驗機率概似法（maximum a posterior，簡稱：MAP）來計算最有可能的面向的權重（aspect weight）。其目標函數如下：

$$\mathcal{L}(d) = \log p(\alpha_d|\mu, \Sigma)p(r_d|\sum_{i=1}^{k}\alpha_{di}\sum_{j=1}^{n}\beta_{dij}W_{dij}, \delta^2) \qquad \text{公式 3.2.6}$$

對應於每個文件（document）面向的權重（aspect weight），由於$\alpha_d$和$r_d$由多變量高斯分佈和高斯分佈所生成，可以將上述式子展開為：

$$\widehat{\alpha_d} = argmax\mathcal{L}(\alpha_d)$$

$$= argmax[-\frac{(r-\alpha_d{}^T s_d)^2}{2\delta^2} - \frac{1}{2}(\alpha_d - \mu)^T \Sigma^{-1}(\alpha_d - \mu)]$$

假設限制條件如下：

$$\sum_{i=1}^{k} \alpha_{di} = 1 \quad ; 0 \leq \alpha_{di} \leq 1 \; for \; i = 1,2, \ldots k \qquad \text{公式 3.2.7}$$

我們採用 L-BFGS-B 方法，求極大值：

$$\frac{\partial \mathcal{L}(\alpha_d)}{\partial \alpha_d} = -\frac{(\alpha_d{}^T s_d - r_d) s_d}{\delta^2} - \Sigma^{-1}(\alpha_d - \mu) \quad \text{公式 3.2.8}$$

並使用最大概似估計法(Maximum Likelihood)找出最佳化的$\widehat{\Theta} = (\hat{\mu}, \widehat{\Sigma}, \widehat{\delta^2}, \hat{\beta})$以最大化給定評論文件（document）下觀察值為給定文件（document）的整體評分（overall rating）的機率。

對於整體評論的 log-likelihood 函數和 ML 估計是分述如下：

$$\mathcal{L}(D) = \sum_{d \epsilon D} \log p(r_d | \mu, \Sigma, \delta^2, \beta, W_d) \qquad \text{公式 3.2.9}$$

$$\hat{\Theta} = arg \max_{\Theta} \sum_{d \in D} \log p(r_d | \mu, \Sigma, \delta^2, \beta, W_d) \quad \text{公式 3.2.10}$$

為了能最佳化問題，此處研究使用了類似 EM 演算法的方式，並於迭代開始之前，先隨機初始化$\Theta_{(0)} = (\mu, \Sigma, \delta^2, \beta)$：

（1）E 步驟：已知參數可以藉由推論公式 3.2.2 和 3.2.7 求得每個文件的潛在面向分數（latent aspect rating）和潛在面向的權重（aspect weight）。

（2）M 步驟：透過求出的潛在面向分數（latent aspect rating）和潛在面向的權重（aspect weight)去更新(update)模型參數$\Theta = (\mu, \Sigma, \delta^2, \beta)$並透過極大化 complete likelihood 來得到$\Theta_{(t+1)} = (\mu, \Sigma, \delta^2, \beta)$，持續執行 E 和 M 的步驟，最後參數會收

斂達到終止條件。

其中更新（update）模型參數如下：

$$\mu_{(t+1)=arg\max_{\mu}-\sum_{d\epsilon D}(\alpha_d-\mu)^T\Sigma^{-1}(\alpha_d-\mu)}$$ 公式：3.2.11

$$=\frac{1}{|D|}\sum_{d\epsilon D}\alpha_d$$

$$\Sigma_{(t+1)=arg\max_{\Sigma}[-|D|\log\Sigma-\sum_{d\epsilon D}(\alpha_d-\mu_{(t+1)})^T\Sigma^{-1}(\alpha_d-\mu_{(t+1)})]}$$ 公式 3.2.12

$$=\frac{1}{|D|}\sum_{d\epsilon D}(\alpha_d-\mu_{(t+1)})\left(\alpha_d-\mu_{(t+1)}\right)^T$$

$$\delta^2_{(t+1)}=arg\max_{\delta^2}[-|D|\log\delta^2-\frac{\sum_{d\epsilon D}(r_d-\alpha_d{}^Ts_d)^2}{\delta^2}]$$ 公式 3.2.13

$$=\frac{1}{|D|}\sum_{d\epsilon D}(r_d-\alpha_d{}^Ts_d)^2$$

這裡我們使用複迴歸分析（multiple regression analysis） 方法替代 Wang et al.（2010）提出的共軛梯度下降法（conjugate-gradient-interior-point）來進行$\beta$的估計。

我們將迴歸問題定義為整體評分（overall rating）$r_d$為應變數$y$，自變數則為$\beta_{dij}$迴歸係數為已知的$W_{dij}\alpha_{di}$，d 為文件（document）的索引，i 為面向（aspect）的索引，j 為 unique token 的索引，如下式。

$$r_d = W_{d11}\alpha_{d1}\times\beta_{d11}+W_{d12}\alpha_{d1}\times\beta_{d12}+\cdots+W_{dij}\alpha_{di}\times\beta_{dij}$$ 公式 3.2.14

我們可以藉由求解$(XX^T)^{-1}X^TY$，推斷出$\beta$的更新值。

## 3.3 基準線模型

在基準線的選擇上，我們使用於文獻探討中 Wang et al.（2010）所提出的 Bootstrap＋LRR 兩階段模型。雖然 Bootstrap＋LRR 兩階段模型在進行擷取主題面向演算法之前，需要先手動設定主題面向數量（k）和初始的種子字，因而在應用上會受到不少的限制，但其為最早開始討論 LARA 相關議題的研究，也有一定的預測準確度，故選擇其為基準線模型。

# 第四章 資料處理

本章將描述實驗中所使用的資料集的特性和前處理步驟以及模型建立的過程，以建立後續對於實驗結果的了解。

## 4.1 資料來源

本研究將使用華文最大的旅遊網站攜程網（www.ctrip.com）旅遊評論和全球最大的旅遊評論網站 TripAdvisor 為分析資料集，其中攜程網資料為使用網路爬蟲擷取後整理而成。

每個資料集（data set）都有具備評論內容和總體評論分數（overall rating），此外，攜程網評論包含已定義的 4 個面向和其評分（位置、設施、服務、乾淨），同樣地 TripAdvisor 評論也包含已定義的 7 個面向和其評分（location、value、room、cleanliness、check in/front desk、service、business service），故將兩資料集的已定義面向評分當做實驗的 ground truth，方便評估接下來的實驗結果。以下圖 4.1.1、4.1.2 分別為攜程網、TripAdvisor 評論範本。

{"helpful": "0", "service": "5", "facility": "5", "review": "不错酒店位于市中心，交通便利，闹中取静，房间很干净，布置的也很温馨，相比同档次酒店，性价比很高。服务员很热情，还推荐了酒店免费的 SPA 体验，下次还会入住。", "author": "E4646****", "user_type": "独自出游", "review_overall_rating": "5.0", "location": "5", "clean": "5", "date": "2014-02-13", "room_type": "豪华客房"}

圖 4.1.1 攜程網資料集（http://www.ctrip.com）

<Overall Rating>4

<Avg. Price>$302

<URL>http://www.tripadvisor.com/ShowUserReviews-g60878-d100504-r22932337-Hotel_Monaco_Seattle_a_Kimpton_Hotel-Seattle_Washington.html

<Author>selizabethm

<Content>Wonderful time- even with the snow! What a great experience! From the goldfish in the room (which my daughter loved) to the fact that the valet parking staff who put on my chains on for me it was fabulous. The staff was attentive and went above and beyond to make our stay enjoyable. Oh, and about the parking: the charge is about what you would pay at any garage or lot- and I bet they wouldn't help you out in the snow!

<Date>Dec 23, 2008

<No. Reader>-1

<No. Helpful>-1

<Overall>5

<Value>4

<Rooms>5

<Location>5

<Cleanliness>5

<Check in / front desk>5

<Service>5

<Business service>-1

圖 4.1.2 TripAdivisor 資料集（Wang et al., 2010）

## 4.2 資料前處理

在進行模型建立之前我們需要先分別對兩個資料集的原始資料進行前處理（preprocessing）。

### 4.2.1 攜程網資料

我們在分析相關工具後使用 jieba 分詞（https://github.com/fxsjy/jieba）當做中文斷詞工具。前處理具體步驟如下：

（1）由於評論是簡體中文故將 gb2312 轉碼成 utf-8

（2）去除中文的標點符號和停止詞（http://www.cnblogs.com/ibook360/）

（3）經調測後增加去除出現頻率高，但意思較不明顯的單詞

| ['很', '住', '一个', '小', '房', '老', '差', '帮', '里', '算', '适合', '太', '订', '很多', '日本', '每次', '好好', '值得', '再', '挺','这家', '好好', '日航', '推荐', '喜欢', '酒店', '不错', '好', '都', '还', '新', '高', '价格', '总体', '不', '感觉', '大', '很大', '免费', '入住', '人', '说', '没', '房间', '携程', '时', '上', '去', '后', '才', '很', '下次', '还会', '出行', '选择', '满意', '会'] |
|---|

表 4.2.1 去除字詞

（4）使用結巴分詞（Jieba）中進行中文分詞

（5）取頻率最高的前 970 為 unique token list 並建立 $sentence\ ID - token\ ID$ 表

| #sentence ID – token ID<br>1 – 3 4 7 13 21<br>2 – 1 2 5 19 45 |
|---|

圖 4.2.2 $sentence\ ID - token\ ID$ 表示意圖

### 4.2.2 TripAdvisor 資料

英文斷詞我們進行以下步驟：

（1）將單詞轉成小寫

（2）去除停止詞（http://www.lextek.com/manuals/onix/stopwords1.html）

（3）使用 Porter Stemmer 進行 stemming

（4）使用 NLTK 工具進行斷詞、斷句

（5）取頻率最高的前 970 為 unique token list 並建立 $sentence\ ID - token\ ID$ 表

## 4.3 實驗描述

我們的研究模型主要有分兩階段，第一階段 Local LDA 模型使用於前處理完成的$sentence\ ID - token\ ID$ 表，除了希望能自動發現面向（aspect）的詞彙並將 sentence 給定屬於的面向（aspect），完成面向切割的工作。

為了避免稀疏性以及有單篇評論無法涵蓋所有面向，所以我們將針對同一個旅館所撰寫的評論視為同一篇評論，形成虛擬的評論，並將評論分數合併並平均。

由於運算資源的限制，於攜程網資料及我們將 143,082 個詞，取頻率閥值（threshold）1,873，將出現次數前 970 取為 unique token。而於 TripAdvisor 則從 113,529 詞中取 3,147,024 為閥值，也將出現次數前 970 取為 unique token。

其資料統計特性經合併如下表所示：

表 4.3.1 資料統計特性

| | 攜程網（Ctrip） | TripAdvisor |
|---|---|---|
| 文件數量 | 527 | 1850 |
| 句子數量 | 1,205,050 | 2,095,763 |
| 單詞數量 | 6,963,373 | 14,639,225 |
| 詞彙表（unique token） | 970 | 970 |
| 句子平均長度 | 5.778493 | 6.985153 |
| 句子長度標準差 | 6.431855 | 5.430736 |
| 資料集評論時間 | 2010-08 ~ 2014-02 | 2009-02 ~ 2009-03 |

### 4.3.1 Local LDA 模型

（1）攜程網

在進行非監督學習之前，我們需要決定總共要將資料分成幾群才是最佳的狀況，亦即先確認所欲抽取的 aspect 數量 k，透過公式 3.2.1 演算法，從 k=1 到 k=5 依序測試，我們得到 k=4 為最適合的分群數量。

令 LDA 模型的參數($k = 4$；$\alpha = 0.1$；$\beta = 0.1$；$n = 1{,}000$；$burnin = 500$)。

運行結果如表 4.3.2 所示：

表 4.3.2 攜程網 Local LDA 面向切割結果

| 推論面向（Inferred Aspect） | 代表詞（Representative Words） |
|---|---|

| | |
|---|---|
| 1. 位置（Location） | 位置、近、很近、机场、西湖、地铁、吃饭、交通、周边、购物、分钟、早餐、地理位置、距离、地方、干净、步行、旁边、打车、服务 |
| 2. 設施（Facility） | 设施、服务、早餐、位置、装修、舒服、旧、陈旧、卫生、点、硬件、床、环境、不好、卫生间、地理位置、隔音、五星、标准、干净 |
| 3. 服務（Service） | 服务、前台、早餐、服务员、客人、升级、这次、送、晚上、带、退房、吃、特别、大堂、餐厅、态度、人员、热情、看、真的 |
| 4. 清潔（Clean） | 交通、服务、干净、便利、环境、设施、位置、出差、卫生、市中心、位于、地理位置、舒适、热情、早餐、朋友、商务、温馨、齐全、安静 |

（2）TripAdvisor

透過公式 3.2.1 演算法，從 k=3 到 k=9，我們得到 k=8 為最適合的分群數量。令 LDA 模型的參數（$k = 4$；$\alpha = 0.1$；$\beta = 0.1$；$n = 1{,}000$；$burnin = 500$）。運行結果如下所示：

表 4.3.3 TripAdvisor Local LDA 面向切割結果

| 推論面向（Inferred Aspect） | 代表詞（Representative Words） |
|---|---|
| value | hotel, stay, review, would, good, price, room, great, expect, book, servic, like, better, much, place star, read, rate, locat, money |
| staff | staff, friend, help, hotel, servic, great, desk, front, us, english, nice, clean, excel, alway, extrem, speak, room, good, pleasant, concierg |
| view | beach, pool, resort, get, go, peopl, like, day, time, one, bar, beauti, water, want, realli, lot, even, area, place, also |
| service | breakfast, restaur, food, good, buffet, bar, servic, also, coffe, free, great, drink, includ, eat, dinner, room, nice, day, excel, serv |
| cleanliness & room | room, clean, bed, bathroom, view, nicev, floor, comfort, small, hotel, larg, wellv, shower, size, great, window, spacious, decor, suit, area |
| location | hotel, walk, locat, great, street, |

| | |
|---|---|
| | shop, minut, close, right, away, restaur, block, station, area, citi, squar, distanc, within, also, easi |
| check/in & front desk | room, us, check, day, get, arriv, one, ask, cal, time, hotel, would, back, night, told, even, got, desk, front, could |
| business service | stay, hotel, would, night, great, recommend, time, return, back, love, place, trip, year, go, definit, new, day, locat, week, visit |

經過 Local LDA 的 Aspect Segmentation 演算法後，我們可以得到兩個資料集中每個 document 都有一個$k \times n$的特徵矩陣$W_d$，其中 d 為 document 的 index，而 i 是指 aspect 的 index，j 為 token 的 index。$W_{dij}$代表第 d 個 document 中，屬於 aspect $i$的第$j$個 token，我們用屬於 aspect $i$ 所有 token 總數進行標準化。

### 4.3.2 LRR 模型

藉由上述步驟我們可以進行 LRR 模型的建立並推論出潛在面向的權重(aspect weight）和潛在的面向分數（latent aspect rating）。

一開始給定各參數隨機初始值，給定初始值如圖 4.3.1、圖 4.3.2，經過類似 EM 的迭代和參數的更新和執行完迴歸模型後，可以得到收斂後的參數

$\Theta = (\mu, \Sigma, \delta^2, \beta)$和潛在面向的權重（aspect weight）。藉由公式 2.4.2 我們可以獲得潛在的面向分數（latent aspect rating）。

| | $A_1$ | $A_2$ | … | $A_i$ |
|---|---|---|---|---|
| $A_1$ | 1 | 0 | 0 | 0 |
| $A_2$ | 0 | 1 | 0 | 0 |
| … | 0 | 0 | 1 | 0 |
| $A_i$ | 0 | 0 | 0 | 1 |

圖 4.3.1 $\Sigma$ 初始值示意圖

| | |
|---|---|
| $\mu_1$ | 0.125 |
| $\mu_2$ | 0.125 |
| … | |
| $\mu_i$ | 0.125 |

圖 4.3.2 $\mu$初始值示意圖

## 4.4 基準線模型實驗

不同於使用 Local LDA 模型來進行面向切割的步驟，基準線模型使用了類似 Bootstrap 的方法，透過人工手動先挑選適合的面向種子關鍵詞（keywords），透過不斷迭代建立出面向代表詞（Representative Words）和建立$k \times n$的特徵矩陣$W_d$。

以下表 4.4.1、表 4.4.2、表 4.4.3、表 4.4.4，依序為攜程網種子詞，攜程網 Bootstrap 面向切割結果、TripAdvisor 種子詞、面向切割結果。

（1）攜程網

中文：

表 4.4.1 攜程網種子詞

| 面向（Aspects） | 種子詞（Seed Words） |
|---|---|
| 位置 | 交通, 位置, 地理位置, 市中心 |
| 設施 | 设施, 商务, 硬件, 酒店设施 |
| 服務 | 服务, 客人, 早餐, 服务员 |
| 乾淨 | 干净, 舒服, 卫生, 整洁 |

表 4.4.2 攜程網 Bootstrap 面向切割結果

| 推論面向（Inferred Aspect） | 代表詞（Representative Words） |
|---|---|
| 位置 | 交通, 位置, 地理位置, 市中心, 便利, 位于, 环境, 闹中取静, 布置, 温馨, 极为, 周边, 出差, 舒适, 天河, 西面, 好近, 预定, good, 天河区 |
| 設施 | 设施, 商务, 硬件, 酒店设施, 出差, 齐全, 软件, 很近, 高架, 配套, 人性化, 中, 面积, 安静, 收费, 一家, 确实, 相比, 优越, 客户, 日式 |

| | |
|---|---|
| 服務 | 服务, 客人, 早餐, 服务员, 前台, 热情, 态度, 餐厅, 人员, 广州, 一流, 问, 点, 大堂, 希望, 特别, 到位, 元, 晚上, 行李, 管理 |
| 乾淨 | 干净, 舒服, 卫生, 整洁, 床, 宽敞, 设计, 漂亮, 旁边, 唯一, 火车, 设备, 少, 整体, 距离, 服务周到, 豪华, 改进, 餐饮, 购物 |

（2）TripAdvisor

表 4.4.3 TripAdvisor 種子詞

| 面向（Aspects） | 種子詞（Seed Words） |
|---|---|
| value | valu, price, qualiti, worth |
| room | room, suit, view, bed |
| location | locat, traffic, minut, restaur |
| cleanliness | clean, dirti, maintain, smell |
| check in | stuff, check, help, reserv |
| service | servic, food, breakfast, buffet |
| business service | busi, center, comput, internet |

表 4.4.4 TripAdvisor Bootstrap 面向切割結果

| 推論面向（Inferred Aspect） | 代表詞（Representative Words） |
|---|---|

| | |
|---|---|
| value | valu, price, qualiti, wort, hotel, money, good, stay, compar, nice, place, probabl, manhattan, littl, better, nyc, giraff, pay, cheap, expens |
| room | room, suit, view, bed, balconi, comfort, small, trip, spacious, bathroom, floor, larg, well, bedroom, cheesstandard, wine, book, doubl, river |
| location | locat, traffic, minut, restaur, walk, within, close, central, park, union, away, conveni, also, distanc, downtown, block, easi, near, uptown, madison |
| cleanliness | clean, dirti, maintain, smell, modern, night, appoint, kept, spotless, decent, extrem, high, furnitur, lobbi, alway, stylish, came, effici |
| check in | stuff, check, help, reserv, desk, us, found, love, bag, arriv, enjoy, put, front, upon, without, attent, pleasant, general, make, direct |
| service | servic, food, breakfast, buffet, continent, morn, includ, coffe, complimentari, fruit, excel, bagel, day, fresh, hour, offer, cereal, juic, although, love |

| | |
|---|---|
| business service | busi, center, comput, internet, travel, access, tri, trip, connect, use, wireless, work, need, much, pleasur, laptop, though, district, street, like |

完成了面向的切割之後我們同樣使用 LRR 模型進行潛在面向的權重（aspect weight）和潛在的面向分數（latent aspect rating）的推論，獲得面向的權重（aspect weight）和潛在的面向分數（latent aspect rating）。

# 第五章 實驗結果

本章說明 Local LDA 模型加上 LRR 模型研究的實驗結果並和 Bootstrap 加 LRR 基準線模型實驗結果進行比較。本實驗共使用華人最大和全球最大的旅遊評論網站（攜程網）和 TripAdvisor 兩個網路評論資料集進行實驗。以下是使用的評估項目：

（1）mean square error

比較面向評分（Aspect Rating）預測值和 ground truth 的差異。

$$\Delta^{2}_{aspect} = \sum_{d=1}^{|D|}\sum_{i=1}^{k}(s_{di} - s^{*}_{di})^{2}/(k\times|D|) \quad \text{公式 5.1}$$

（2）aspect correlation inside reviews

比較評論內面向的相關性（correlation）。

$$\rho_{aspect} = \sum_{d=1}^{|D|}\rho_{s_d,s^{*}_d}/|D| \quad \text{公式 5.2}$$

（3） aspect correlation across reviews

比較面向之間的相關性（correlation）。

$$\rho_{review} = \sum_{i=1}^{k}\rho_{\frac{(\vec{s_i},\vec{s^{*}_i})}{k}} \quad \text{公式 5.3}$$

由於需要對照 ground truth 所對應的面向，我們將 Local LDA 中 check in 和 front desk 合併，Local LDA 和基準線模型的 cleanliness 和 room 合併。而 business service 在面向切割結果中表現不理想，所以這邊先不考慮 business service 這個面向。實驗結果如表 5.1、表 5.2 所示：

表 5.1 攜程網中文評論資料集實驗結果表

| | $\Delta^2_{aspect}$ | $\rho_{aspect}$ | $\rho_{review}$ |
|---|---|---|---|
| Local LDA + LRR | 10.46249 | **0.5743762** | **0.2082056** |
| Bootstrap + LRR | **9.567472** | 0.3107433 | 0. 1944556 |

表 5.2 TripAdvisor 英文評論資料集實驗結果表

| | $\Delta^2_{aspect}$ | $\rho_{aspect}$ | $\rho_{review}$ |
|---|---|---|---|
| Local LDA + LRR | **6.280913** | **0.5123924** | **0.3145933** |
| Bootstrap + LRR | 6.675879 | 0.4790101 | 0.2421156 |

表 5. 3 為整體評論同樣都為 4.6 的評論，我們可以發現雖然使用者對於實體的整體評分相同，但事實上評論者對於不同的面向各自有不同的權重。

表 5. 3 相同整體評分的評論權重

| | | |
|---|---|---|
| 服務 | (0.14) | (0.17) |
| 設施 | (0.15) | (0.16) |
| 位置 | (0.53) | (0.64) |
| 乾淨 | (0.78) | (0.64) |

我們可以從表 5. 4、5. 5 以及表 5. 6、5. 7 分別觀察到個別評論面向分

數和面向權重，其中面向分數表格裡括弧為 ground truth。同時由表 5.1 得知評論內面向的相關性表現較好，亦即個別評論中的面向分數排序正確性表現較佳，但面向分數在分布上差距較大，造成 mean square error 表現較差，我們推測在 EM 步驟求極大化參數時，若能將參數優化將會有更好的表現。

表 5.4 個別評論面向分數範例一

| | |
|---|---|
| 服務 | 0.2067060 (4.7) |
| 設施 | 0.7633797 (4.6) |
| 位置 | 1.3250827 (4.7) |
| 乾淨 | 4.8801735 (4.8) |

表 5.5 個別評論面向權重範例一

| | |
|---|---|
| 服務 | 0.1611887 |
| 設施 | 0.1627941 |
| 位置 | 0.5261116 |
| 乾淨 | 0.7653878 |

表 5.6 個別評論面向分數範例二

| | |
|---|---|
| 服務 | 0.1956294 (4.5) |
| 設施 | 0.7329265 (4.6) |
| 位置 | 1.3534558 (4.6) |
| 乾淨 | 4.9262485 (4.7) |

表 5.7 個別評論面向權重範例二

| | |
|---|---|
| 服務 | 0.1668522 |
| 設施 | 0.1632480 |
| 位置 | 0.6413038 |
| 乾淨 | 0.6437909 |

# 第六章 結論與建議

## 6.1 實驗結論

從實驗結果可以發現，雖然因為運算資源和時間的限制無法將結果最佳化，尤其中文資料集比起英文資料集更面對了歧義詞、領域詞彙和未見過詞彙等問題，造成 MSE 結果上表現並不理想，但可以發現不管是中文評論或是英文評論在使用 Local LDA＋LRR 模型的面向評分相關性表現上均優於基準線 Bootstrap＋LRR 的模型，更重要的是使用 Local LDA 於做面向分割時不必事先人工手動設定關鍵字，這樣可以讓研究的應用更加廣泛。

## 6.2 研究貢獻

本研究試圖解決中文評論所遇到的 LARA 問題，運用 Local LDA 和 LRR 兩階段模型，希望透過給定整體評分（overall rating）和評論的內容（review content）能夠推論出潛在面向的權重（aspect weight）和潛在的面向分數（latent aspect rating），提供使用者評論中所潛藏的面向等更深入的決策資訊。從實驗中我們可以看到中文評論或是英文評論資料集在實驗結果上都表現的比基準線模型來的優異。

此外，我們也整理過去研究者在意見探勘（opinion mining）和情感分析（sentiment analysis）方面的研究。讓對於此方面研究有興趣的研究者可以更進一步的深入探討相關的議題。

## 6.3 未來研究方向

在應用 LARA 分析於中文評論後，未來研究可以朝三個方向繼續深入探討相關議題。第一方面是嘗試優化 Local LDA 模型和 LRR 模型的

各項參數，雖然最佳化參數並非本研究之目標，但若能在運算資源和時間的許可下增加 unique token 的數量，讓整個模型實驗的結果可以表現的更好；第二個方向是更深入的將研究中的模型應用於個人化商品推薦、客製化資訊檢索等領域；第三個方向是加入時間、評論者性別等其他可以探討的變數進行其他更進一步的分析研究。

# 參考文獻

[1] Onix text retrieval toolkit stopword list.
http://www.lextek.com/manuals/onix/stopwords1.html

[2] M. Porter. An algorithm for su±x stripping. Program,14(3):130 - 137, 1980.

[3] Wang, H., Lu, Y., & Zhai, C. (2010, July). Latent aspect rating analysis on review text data: a rating regression approach. In Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 783-792). ACM.

[4] Ma, G., & Qu, Y. (2012, October). A local LDA based method for Latent Aspect Rating Analysis on reviews. In Signal Processing (ICSP), 2012 IEEE 11th International Conference on (Vol. 3, pp. 2240-2245). IEEE.

[5] D. Blei, A. Ng, and M. Jordan., (2003), Latent dirichlet allocation. The Journal of Machine Learning Research, 3:993-1022,2003.

[6] Dave, K., Lawrence, S., & Pennock, D. M. (2003, May). Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. InProceedings of the 12th international conference on World Wide Web (pp. 519-528). ACM.

[7] Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and trends in information retrieval, 2(1-2), 1-135.

[8] Pak, A., & Paroubek, P. (2010, May). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In LREC.

[9] Hu, M., & Liu, B. (2004, July). Mining opinion features in customer reviews. InAAAI (Vol. 4, pp. 755-760).

[10] Chaovalit, P., & Zhou, L. (2005, January). Movie review mining: A comparison between supervised and unsupervised classification approaches. In System Sciences, 2005. HICSS'05. Proceedings of the 38th Annual Hawaii International Conference on

(pp. 112c-112c). IEEE.

[11] Hu, M., & Liu, B. (2004, August). Mining and summarizing customer reviews. InProceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 168-177). ACM.

[12] Titov, I., & McDonald, R. (2008). A joint model of text and aspect ratings for sentiment summarization. Urbana, 51, 61801.

[13] Lu, Y., Zhai, C., & Sundaresan, N. (2009, April). Rated aspect summarization of short comments. In Proceedings of the 18th international conference on World wide web (pp. 131-140). ACM.

[14] Lin, C., & He, Y. (2009, November). Joint sentiment/topic model for sentiment analysis. In Proceedings of the 18th ACM conference on Information and knowledge management (pp. 375-384). ACM.

[15] Brody, S., & Elhadad, N. (2010, June). An unsupervised aspect-sentiment model for online reviews. In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (pp. 804-812). Association for Computational Linguistics.

[16] Jo, Y., & Oh, A. H. (2011, February). Aspect and sentiment unification model for online review analysis. In Proceedings of the fourth ACM international conference on Web search and data mining (pp. 815-824). ACM.

[17] Su, Q., Xu, X., Guo, H., Guo, Z., Wu, X., Zhang, X., ... & Su, Z. (2008, April). Hidden sentiment association in chinese web opinion mining. In Proceedings of the 17th international conference on World Wide Web (pp. 959-968). ACM.

[18] Angeliki Lazaridou, Ivan Titov and Caroline Sporleder, (2013), A Bayesian Model for Joint Unsupervised Induction of Sentiment, Aspect and Discourse Representations.

[19] Bing Lu, (2013), Web Data Mining: Exploring Hyperlinks, Contents, and Usage

Data (Data-Centric Systems and Applications)

[20] CNET

http://www.cnet.com/products/apple-iphone-5/user-reviews/

[21] W. Gerrod Parrott, (2001), Emotions in Social Psychology: Essential Readings

[22] Bo Pang and Lillian Lee, (2008), Opinion Mining and Sentiment Analysis

[23] Bo Pang and Lillian Lee, Shivakumar Vaithyanathan, (2002), Thumbs up? Sentiment Classification using Machine Learning Techniques

[24] Bo Pang and Lillian Lee, (2005), Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales, Proceedings of ACL 2005.

[25] Peter D. Turney, (2002), Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews

[26] B Liu, M Hu, J Cheng, (2005), Opinion Observer: Analyzing and Comparing Opinions on the Web

[27] Minqing Hu and Bing Liu, (2004), Mining and Summarizing Customer Reviews

[28] Yiming Yang ,Jan O. Pedersen, A Comparative Study on Feature Selection in Text Categorization

[29] Samuel Brody, Noemie Elhadad, (2007), An Unsupervised Aspect-Sentiment Model for Online Reviews

[30] Niu, Zheng-Yu, Dong-Hong Ji, and Chew-Lim Tan., (2007) I2r: three systems for word sense discrimination, chinese word sense disambiguation, and English word sense disambiguation. In SemEval '07: Proc. of the 4th International Workshop on Semantic Evaluations. ACL, Morristown, NJ, USA, pages 177-182

[31] Yue Lu, ChengXiang Zhai, Neel Sundaresan, (2009), Rated Aspect Summarization of Short Comments

[32] Yi Wang, (2008), Distributed Gibbs Sampling of Latent Topic Models: The Gritty

Details

[33] Prasanth Lade, (2011), Study of Latent Dirichlet Allocation (LDA) models and their application to Human Affective state recognition

[34] 中文停止詞表

http://www.cnblogs.com/ibook360/archive/2011/11/23/2260397.html