東海大學資訊工程研究所

碩士論文

指導教授：羅文聰　博士

基於主題模型應用於期刊論文分類之研究

The study of journal paper classification based on the Latent Dirichlet allocation

研究生：林修漢

中華民國 一○四 年　七　月

東海大學碩士學位論文考試審定書

東海大學資訊工程學系 研究所

研究生 林 修 漢 所提之論文

基於主題模型應用於期刊論文分類之研究

經本委員會審查，符合碩士學位論文標準。

學位考試委員會
召 集 人 _____ 簽章

委 員 _____

_____

_____

指 導 教 授 _____ 簽章

中華民國 104 年 6 月 18 日

摘 要

在資料唾手可得的時代，如何能在大量且雜亂的資料(Big data)取得有用的資訊成為一個重要的課題。利用分類的方法，協助這些資料更容易被使用是目前普遍的做法，然而同樣問題在學術上也備受討論，因此期刊論文的分類方法的研究也日益劇增，由於期刊論文分類無法年年更新，確時有改善與建議的空間。

本文將利用主題模型中的 LDA(latent dirichlet allocation)進行萃取主題，同時也達到分類的效果，利用多個年份的期刊論文做為資料集，其輸出結果在與知名期刊的分類表進行比對，針對相同及相異處進行分析及討論，並計算各類別的強度，進行資料可視化的動作，幫助研究者能更加容易的剖析資料，相信整套方法流程也能夠套用在其他專業的文件上。

關鍵詞：文件分類、主題模型 、隱含狄利克雷分布

Abstract

These days, we can easily obtain any data, but there is one issue still has to be discussed: how to get useful data from messy and numerous big data. By classification is the most common way to assistant these data to be used easier. However, the same issue is also discussed by scholars, as a result, the research on classification of journals papers is getting more and more. Because the classification of journals papers cannot be updated yearly, there are still some shortcomings have to be solved.

In this paper, we use LDA(latent dirichlet allocation) of topic model to make topic extraction and classify in the same time. Take numerous journals as the data set and compare the result with the journals' classification. Next, analyze and discuss the same and different parts, and also calculating the topic strength of every topic to make data visualization. We believe that the method can be used on academic papers and can assistant scholars to analyze easier.

Keywords：Document Classification、Topic Model、Latent Dirichlet Allocation

誌 謝

回顧這三年的時光，研究所生活實在是過得很充實，特別要感謝羅文聰老師對於我的指導，還要感謝許瑞愷老師在碩士的期間給予我的許多的協助，並且讓我學習到如何針對問題，提出解決方法，思考比較這些解決的方法是否合宜，無形中厚實我的研究能量，在這邊也要感謝陳倫奇學長，讓我到國家網路高速運算中心接受學長的指導，雖然進度緩慢並且遲遲無法有顯著的進步，但要感謝陳倫奇學長不厭其煩的帶領我如何做實驗，討論問題，然後在每次會議中激盪想法。

接下來要謝謝學長學弟同學與家人，謝謝 DBLAB 的信良、晉宇、勝傑以及雅婷在研究所這段時間裡一起分享喜怒哀樂，一起努力也一起歡笑；接著要謝謝我們實驗室的銘鴻以及建良，在艱辛的過程裡互相加油。最後要謝謝爸爸、媽媽、哥哥及小貢，畢竟我在大學的時候不是讀資工系本科，研究所的過程也不是一直都很順遂，但有你們的鼓勵讓我在挫折中更有力量去面對。感謝這所有的一切。

感謝人　修漢

目 錄

圖 目 錄

表 目 錄

第一章 緒論

1.1 研究動機與背景

在資訊爆炸的時代，隨著網際網路的發達與雲端服務的普遍，造成各式各樣的信息充斥在我們的生活周遭。面對如此巨量的資料(Big Data)也產生出許多新的問題，如何在這些雜亂無章的信息裡，找尋出對我們有用的信息也成為了一大課題。

然而這樣的課題在學術界也同樣存在，Fraser AG, Dunstan FD 於 2010 提到[1]，在 1970 年時，在科學、科技、醫學這三類共發表了 25400 篇期刊，但是在 2009 年時卻一共至少發表了 150 萬篇，並且每年還以至少成長 3.5%。 面對如此大量的期刊論文，利用文件分類的方法協助這些文件更容易被使用是目前普遍的做法，鑒於上述需求讓的分類問題變得炙手可熱。

本文將探討期刊論文的分類問題，由於更新分類是費時費力的工作，導致期刊論文的分類往往無法每年更新，研究者在投稿的時候常常會發生不知道自己投稿的論文是否放在對的分類；反之如果因為分類問題導致投稿的期刊論文分在不對的類別中，審稿人員對於研究者的期刊論文的領域不一定是最佳的選擇，相對應於此需求，許多相關的文件分類方法也日益遽增。

一般文件分類方法主要可以分為四種模型架構，統計理論(Statistical)、利用幾何距離與向量模型等相似度(Similarity Measures)、以決策樹(Decision Trees)為分類基礎以及以類神經網路模式(Neural Networks)為基礎等四種[2]。不管常見的 SVM（Support Vector Machine）[3] 、KNN（K-Nearest Neighbors）[4]、LLSF（Linear Least Square Fit）[5]、Decision Trees[6]、naïve Bayse[7]都可以歸類於上述四類模型架構，為因應不同的分類需求及情境，相關研究人員基於上述等模型架構為基礎而衍生出其他分類模型。

在進行許多半監督式學習的文件自動分類時[8]，需要了解文件的主題，並

透過此主題大意才能給定此文件的分類。因此要將文件分類自動化，必須先給定分類時的分類規則，藉由學習這些規則，機器才能據以分類，讓機器做自動分類之前，我們必須透過分類規則的訓練，使機器得以自動學習出人工分類的經驗及規則。所謂的訓練，就是讓機器透過分析「訓練文件」，而這些訓練文件記錄了以「人工」為文件做分類的規則，機器透過這些訓練文件所給予的規則，透過學習並歸納出相對應的規則，往後看到相似的文件時，就會給予此文件適當的類別。

對於人工進行分類的部分許多文獻都提到，在人工進行分類時都有分類不一致的現象，索引詞相似的兩篇文件，卻被分在不同的類別，特別 Kao [9]也提到，就算是擁有 30 年經驗的分類專家，在面對相同的文件經過三個月後，卻發現此專家卻做了很不同的分類結果。既然人工分類都有不一致的現象，那機器透過這些訓練集學習再分類的結果也會勢必會受到影響

現今資訊傳播快速地時代裡，期刊論文分類表卻並不是年年對分類進行更新，無法年年更新可能的理由在於論文的資料量過於龐大，人工的速度不能跟上。不只這樣的問題，往往研究的信息是風馳電掣的，對於新出現的研究領域，專業人員可能無法及時地給予正確的分類，並且上述也提到就算是專業的人員也有可能有分類不一致等疑慮。對於這些問題，再再的反應出現有其論文的分類問題值得我們去探討及研究。

1.2 研究目的與方法

主題模型(Topic model)這方面的技術，是在對於大量的文檔透過相關模型的計算給於這些文檔相對應主題的方法，主題模型的想法來自於我們認為作者在寫每篇文章時都有其中心思想，而這些文章的中心思想可能有一個也可能有很多個，而我們稱這些中心思想為主題(Topic)。

主題模型的相關研究日甚一日，其中 LDA (Latent Dirichlet Allocation) 被廣泛運用在各領域中，利用貝氏模型的方式以機率分布生成主題模型，是一個有效的方法能將大量的文檔進行分類。

由於此方法為非監督學習法，我們並不用給予所謂經過人工分類的訓練集，

自然減少了因為人工分類不一致所產生的影響，並且此方法並不需要事先知道已知的類別，少了此設限可以忠實反映出現實文件中，各方的研究者所研究的主題為何，也可以更加迅速的反應出新興的研究類別，也因為 LDA 利用貝氏模型為其架構，只要給予文件或其摘要，就算是大量的文件也能迅速的輸出主題萃取結果，有了此優點要年年去更新分類表將不在是個問題。

本文將透過此方法，以 IEEE Xplore Digital Library 中的 Browse Journals & Magazines 做為我們分類依據的實驗對象，將這些大量的期刊論文利用 LDA 生成主題，透過 LDA 所輸出的分類結果比對 IEEE taxonomy 的結果進行討論及分析，最後我們會利用主題河(ThemeRiver)的方式呈現各個類別的強弱程度幫助研究者更容易看出各類別的趨勢。

1.3 章節概要

本研究將建立對於大量期刊論文的主題萃取及命名與分類方法，透過此方法能對現有的分類結果進行討論及建議，以下介紹各章節內容，第二章為文獻探討，我們將對主題模型的相關文獻進行概述，以及主題河(ThemeRiver)的介紹；第三章為研究方法，我們將詳述整個完整的分類方法流程，第四章為輸出結果與 IEEE Taxonomy 進行討論及建議，最後第五章為結論及未來工作。

第二章 文獻探討

本章節將探討主題模型的相關進程，然後再介紹本文所使用的模型 LDA(latent dirichlet allocation)，最後是可視化工具主題河(ThemeRiver)的介紹。

2.1 主題模型(Topic model)

Topic model 是一種應用十分廣泛的生成式模型(generative model)，在 IR，NLP，ML 都有廣泛的應用，Topic model 最經典的模型之一是 LDA(latent dirichlet allocation) ，我們將在下列介紹其發展進程。

2.1.1 主題模型進程介紹

前面提到 LDA(latent dirichlet allocation)為 Topic model 最經典的模型之一，然而在進程的演化中，我們可以將主題模型的看成三個階段，演變至今可簡單的分為三類模型：布林模型(Boolean Model)、向量空間模型(Vector Space Model)及機率模型(Probability Model)。其中向量模型最具代表的就是 TF-IDF（term frequency–inverse document frequency）以及 LSA (Latent Semantic Analysis)，而機率模型則為 pLSA(Probabilistic latent semantic analysis)以及 LDA(latent dirichlet allocation)。

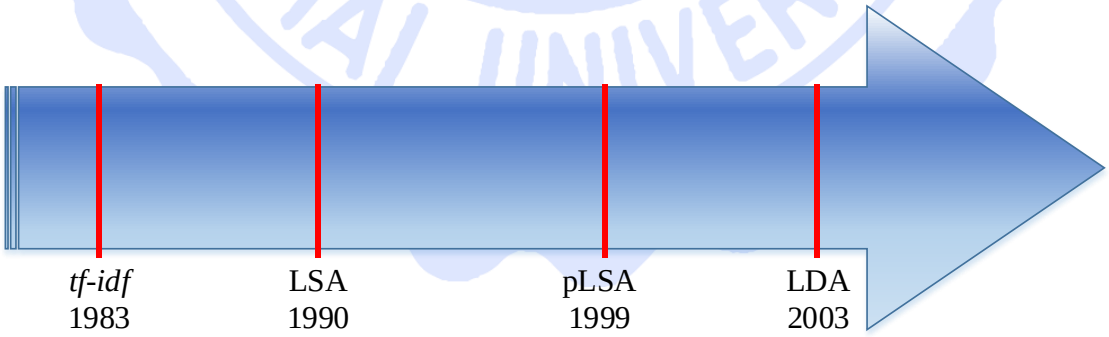
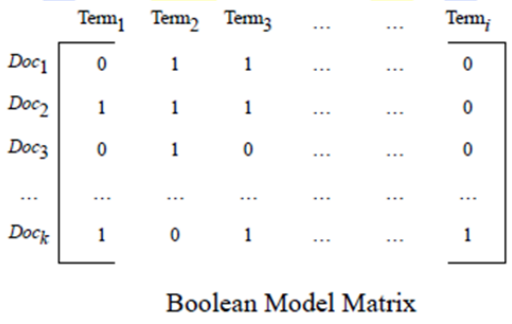


圖 2.1 主題模型發展進程

2.1.2 布林模型(Boolean Model)

布林模型屬於早期的建模方法，其方法簡單的來說就是先將文檔視為數個字彙(term)所組成，然後對文檔進行「字串比對」，找出"完全符合"檢索條件的文檔集合，檢索後只會有完全符合(Exact Match) 不會有部分符合(Partial Match)的情形。其中檢索的條件通常是單字或片語，這些條件可以使用 AND/OR/NOT 等布林運算式將其加以連接，而布林模型的特點在於相當的簡當容易瞭解，但並不具有排名(Ranking)的能力，在檢索時又必須完全符合檢索條件，因此缺乏彈性。綜觀上述特點，導致布林模型在運用上有所限制。



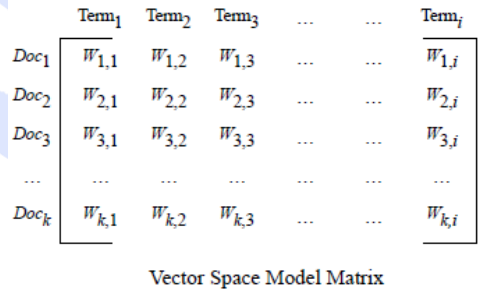
	$Term_1$	$Term_2$	$Term_3$	...	...	$Term_i$
Doc_1	0	1	1	...	...	0
Doc_2	1	1	1	...	...	0
Doc_3	0	1	0	...	...	0
...	...	...	...	...	...	...
Doc_k	1	0	1	...	...	1

Boolean Model Matrix

圖 2.2 布林模型表示方法

2.1.3 向量空間模型(Vector Space Model)

向量空間模型，最早是由 Salton 等人在 1975 年提出[10]，是一種文檔的表示模型，將字彙-文檔(term by document)表示成關係矩陣的方式，以方便計算文檔與要檢索的項目之間相似的程度。



	$Term_1$	$Term_2$	$Term_3$	...	...	$Term_i$
Doc_1	$w_{1,1}$	$w_{1,2}$	$w_{1,3}$	...	...	$w_{1,i}$
Doc_2	$w_{2,1}$	$w_{2,2}$	$w_{2,3}$	...	...	$w_{2,i}$
Doc_3	$w_{3,1}$	$w_{3,2}$	$w_{3,3}$	...	...	$w_{3,i}$
...	...	...	...	...	...	...
Doc_k	$w_{k,1}$	$w_{k,2}$	$w_{k,3}$	...	...	$w_{k,i}$

Vector Space Model Matrix

圖 2.3 向量模型表示方法

對於文檔我們可以表達成文檔向量，其中$d_i = (w_{i,1}, w_{i,2}, \ldots, w_{i,n})$，$w_{i,j}$表示在文檔$d_i$中字彙j的權重$w_{i,j}$。而$w_{i,j}$權重的產生最常使用的是利用*tf-idf* 來計算 $wi;j = tf\,(i,j)\cdot idf(j)$ 。

$$w_{ij} = tf(i,j) \cdot idf(j)$$

$$tf(i,j) = \frac{w_{ij}}{\sum_{n=1}^{N} w_{i,n}}, \quad idf(i) = log\frac{N}{n_i}$$

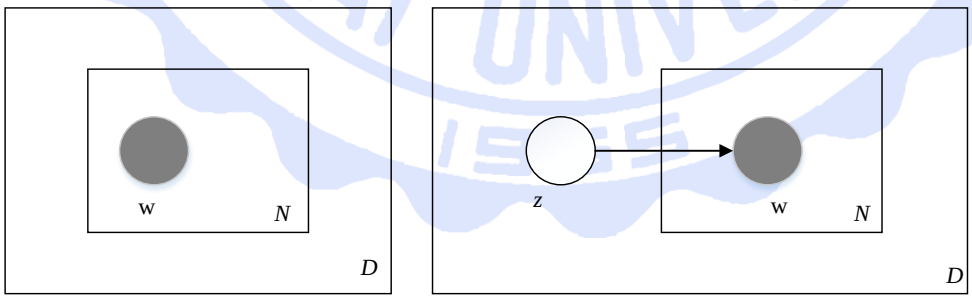
2.1.4 機率模型(Probability Model)

典型機率模型於資訊檢索的應用中，最早是由 Robertson 與 Sparck Jones 在 1976 年所提出[11]，將所要檢索項目與文檔間的相似度，由機率百分比的模型呈現。推演至今，其最為基礎的機率模型為 Unigram model，此模型中文檔由字彙所組成，因此每個字彙在文檔中的機率形成一多項分配，而此一文檔中的所有字彙其機率總合為 1。

$$p(\mathbf{w}) = \prod_{n=1}^{N} p(w_n)$$

Mixture of Unigram 是將原有的 Unigram model 增加一個隨機變量 Z 而成，在 Mixture of Unigram 中，此隨機變量 Z 為一主題。而每份文檔由主題 Z 所產生，接著依照多項分配 p(w|z) 產生 N 個字彙，其文檔機率表示如下：

$$p(\mathbf{w}) = \sum_{z} p(z) \prod_{n=1}^{N} p(w_n|z)$$



(a)unigram (b) mixture of unigrams

圖 2.4 UNIGRAM MODEL&MIXTURE OF UNIGRAM 貝氏圖

2.2 Latent Dirichlet Allocation

Latent Dirichlet Allocation(LDA)，是 Blei 於 2003 年[12]利用 Probabilistic Latent Semantic Indexing (pLSI)模型為基礎所提出，目的在解決 pLSI 所產生 over fitting 的問題。LDA 模型最重要的兩個核心概念為:每一個文檔皆由許多隨機隱含的主題建構而成，而這些主題則是透過字彙所形成的分布組成。

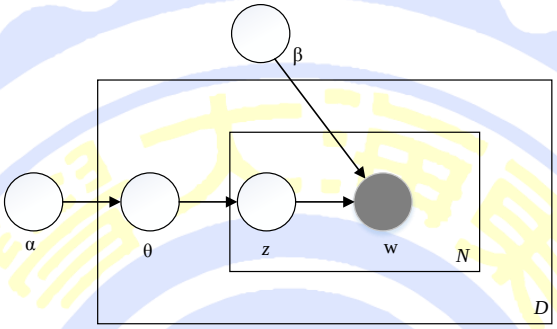


圖 2.5 LDA 貝氏圖

LDA 為一貝氏模型，其中 D 為文檔數目，N 為在某一個文檔中字彙的總數，方形為重複次數。w 為文檔中的字彙，是一個可觀察的已知變量；z 為文檔中的主題，θ為文檔中主題的分布，這兩個變量皆為未知的隱含變量。α為每個文檔下主題的多項分佈中的 Dirichlet 先驗參數，β則是每個主題下字彙的多項分佈中的 Dirichlet 先驗參數。

2.3 Gibbs sampling

Gibbs sampling 為 MCMC(Markov-chain Monte Carlo)算法的一種特殊情況，經常用於處理高維模型的近似推斷。在 Blei 的原文中 ϑ &φ是利用 EM algorithm 算法求解，但 EM algorithm 的方法可能會陷入局部最佳解的問題，因此 Gregor Heinrich 的論文中[13]提出利用 Gibbs sampling 的方法求解，而往後的相關論文在利用 LDA 時也都使用 Gibbs sampling 求解。

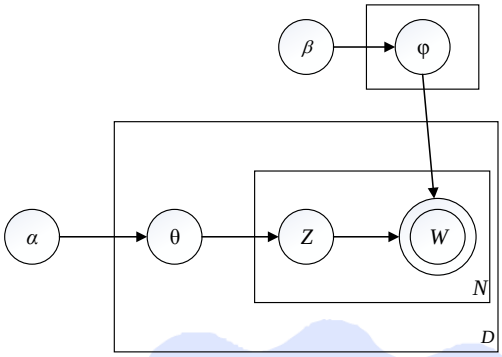


圖 2.6 LDA 貝氏圖

將圖 2.6 LDA 貝氏圖寫成 complete-data 的聯合分佈為:

$$\mathrm{p}(w_m, z_m, \theta_m, \Phi|\alpha, \beta) = \left\{\left[\prod_{n=1}^{N_m} p(w_{m,n}|\varphi_{z_{m,n}}) p(Z_{m,n}|\theta_m)\right] \cdot p(\theta_m|\alpha)\right\} \cdot p(\Phi|\beta)$$

但這個邊緣分佈是無法求解的，因為 $z_{m,n}$ 是隱藏變量，每個詞都跟θ_m和Φ都跟$Z_{m,n}$有關，而連乘又是非常難用積分得到的，這個就是耦合現象。因此我們利用 Gibbs Sample 來求解。

將上面這個聯合分佈利用積分去掉了參數θ_m（doc-topic 分佈）和Φ（topic-word 分佈）得到:

$$\mathrm{p}(w, z|\alpha, \beta) = \mathrm{p}(w|z, \beta)\mathrm{p}(z|\alpha)$$

透過 Gibbs Sample 我們所需要採樣的部分如下:

$$\mathrm{p}(Z_i|Z_{\neg i}, \mathrm{w}) = \frac{\mathrm{p}(\mathrm{z}, \mathrm{w})}{p(Z_{\neg i}, \mathrm{w})} = \frac{p(w|z, \beta)}{p(w|Z_{\neg i}, \beta)} \cdot \frac{p(z|\alpha)}{p(Z_{\neg i}|\alpha)}$$

$$= \frac{\Delta(n_z+\beta)\cdot\Delta(n_m+\alpha)}{\Delta(n_{z,\neg i}+\beta)\cdot\Delta(n_{m,\neg i}+\alpha)} = \frac{\frac{\Gamma(n_z^{(t)}+\beta_t)}{\Gamma(\sum_{v=1}^{V} n_z^{(v)})} \cdot \frac{\Gamma(n_m^{(z)}+\alpha_z)}{\Gamma(\sum_{z=1}^{K} n_m^{(z)}+\alpha_z)}}{\frac{\Gamma(n_z^{(t)}-1+\beta_t)}{\Gamma\left(\left[\sum_{z=1}^{k} n_m^{(z)}+\alpha_z\right]-1\right)}}$$

$$= \frac{n_{\mathrm{z},\neg i}^{(t)} + \beta_t}{\left[\sum_{v=1}^{V} n_z^{(v)} + \beta_v\right] - 1} \cdot \frac{n_{m,\neg i}^{(z)} + \alpha_z}{\left[\sum_{z=1}^{K} n_m^{(z)} + \alpha_z\right] - 1}$$

$$\propto \frac{n_{z,\neg i}^{(t)} + \beta_t}{\left[\sum_{v=1}^{V} n_z^{(v)} + \beta_v\right] - 1} \cdot (n_{m,\neg i}^{(z)} + \alpha_z)$$

因此 LDA 透過 Gibbs Sample 的採樣過程如下:

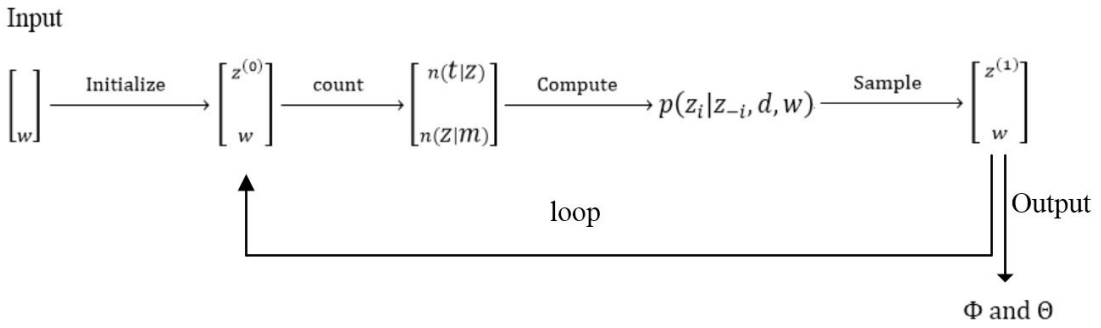


圖 2.7LDA 透過 GIBBS SAMPLE 的採樣過程

Initialize：初始時隨機給文檔中的每個單詞分配主題。

Count：計算每個在每篇文檔屬於第 k 個主題的單詞總數，以及第 k 個主題在所有文檔中生成第 t 個單詞的總數。

Compute：去除當前單詞的主題，根據該文檔中其他單詞的主題評估當前單詞被賦予各個主題的機率。

Sample：在得到當前單詞的主題機率分佈後，根據這個機率分佈為該詞採樣一個新的主題。

Loop：從 Count 步驟重新開始，不斷更新下一個單詞的主題，直到和收斂。

2.4 ThemeRiver

文本可視化技術綜合了文本分析、數據挖掘、數據可視化、計算機圖形學、人機交互、認知科學等學科的理論和方法，為人們理解複雜的文本內容、結構和內在的規律等信息的有效手段。

本文利用主題河(ThemeRiver)[13]這個可視化方法來呈現主題的樣貌，透過 LDA 生成的主題類別於時間上的強弱變化，其每個主題類別的強弱程度來自於每個主題對應到文檔的數量，主題河(ThemeRiver)能讓研究者輕鬆的識別主題的變化。

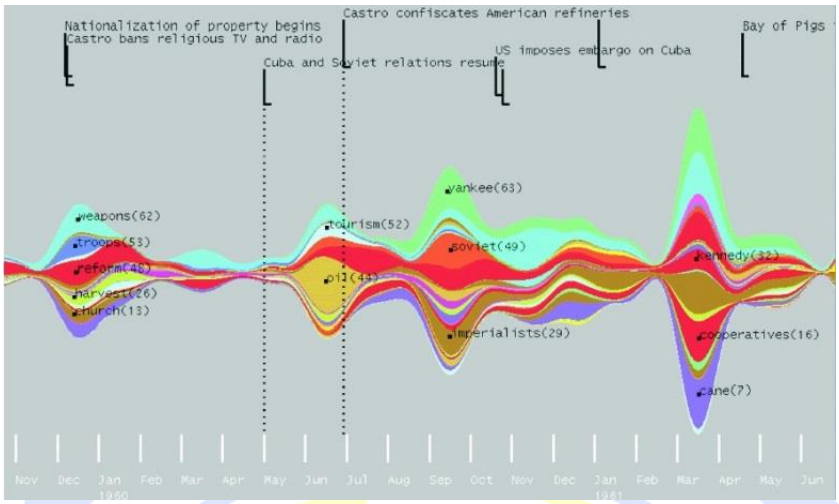


圖 2.8 THEMERIVER 呈現樣貌

第三章研究方法

3.1 方法流程圖

此章節將闡述本論文的方法及流程。而由圖 3.1 可以觀覽本論文方法全貌的運作狀況，下面的小節也會以此圖做輔助說明。

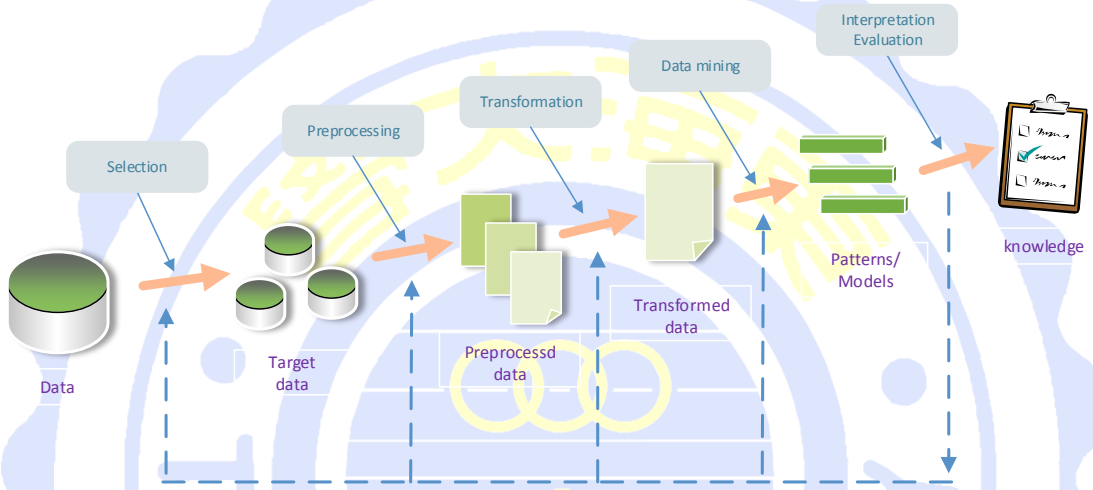


圖 3.1 方法流程圖

3.1.1 流程步驟

根據圖 3.1 所示，為本論文的流程圖。將資料透過資料選擇、資料處理、資料轉換、資料探勘及討論分析等基本步驟達到我們要得到的分類結果。

3.1.2 資料選擇

本論文使用的 Data set 為 IEEE Xplore Digital Library 中的 Browse Journals & Magazines，而我們所選擇的對象為 Browse Journals & Magazines 中，西元 2000 年至西元 2013 年的期刊論文，利用 HTML Parser 來取得我們需要的資料集。

如圖 3.2 所示利用 HTML Parser 抓取 Browse Journals & Magazines 中從西元 2000 年至西元 2013 年所有期刊論文的標題及摘要的部分，總共 346543 篇期刊

論文，其檔案輸出的格式為一篇論文一個 txt 檔，以利後續進行資料處理，資料轉換等步驟 。

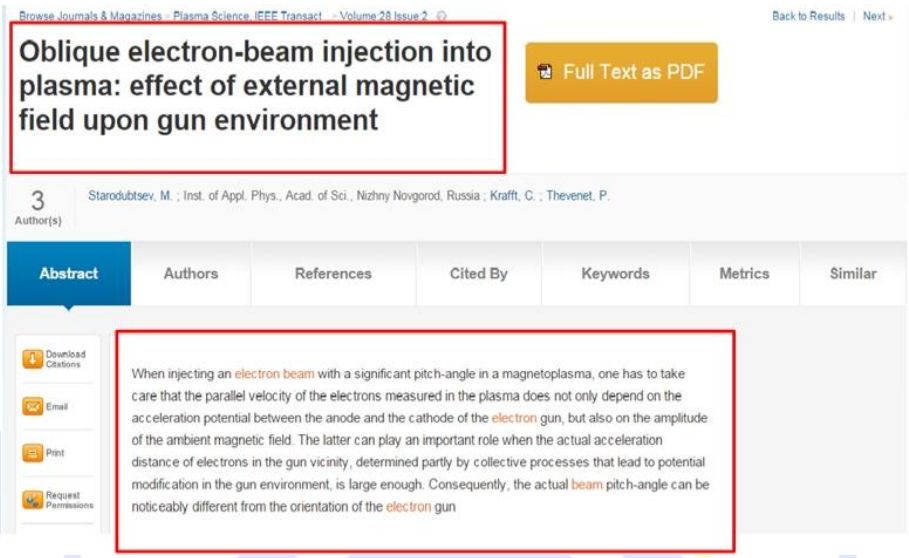


圖 3.2 HTML PARSER 抓取目標

3.1.3 資料處理

抓取下來的文檔，我們將進行預處理，而我們將針對兩種狀況進行預處理:

(一) 在文檔中所有的連接詞及停止詞，我們將於已刪除，因為這些詞類在主題探勘上無法給予我們幫助，往往被認為是數據處理的障礙，通常存在於各個文檔之中，並且擁有極高的詞頻。

(二) 我們使用 Stanford Log-linear Part-Of-Speech Tagger(POS)這項工具將我們所有文檔中的詞進行詞性標記，詞性標記後的結果我們將只保留名詞及專有名詞於我們的文檔中，因為此兩種詞類在自然語處理上有最佳的代表性。

圖 3.3 STANFORD POS 後的文檔

3.1.4 資料轉換

將預處理後的的文檔集，轉換成 LDA 方法所需之格式，而第一行為所有文件的篇數，一篇文件為一行然後依序輸入，LDA 所需文檔格式如下:

[M]

[$document_1$]

[$document_2$]

...

[$document_M$]

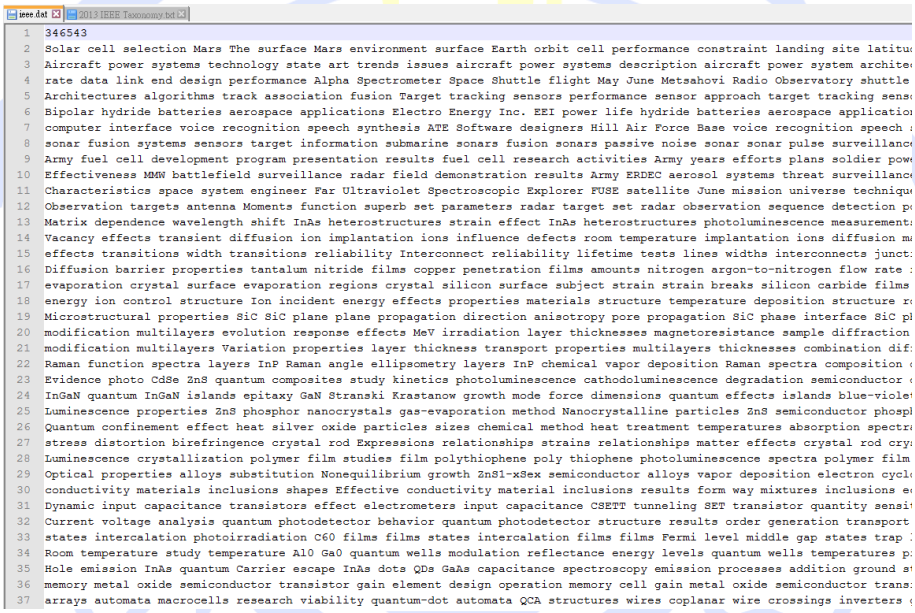


圖 3.4 轉換後的文檔集

3.1.5 資料探勘及討論分析

最後的兩個步驟為資料探勘及討論分析，資料探勘所使用之方法為 Topic model 中的 LDA，LDA 透過 Document-Topic & Topic-word 的關係進行主題萃取，而 LDA 的運行有參數需要調配，我們留於下一節進行實驗及討論，而整個流程最後的結果的討論及分析我們留於下一個章做討論。

3.2 LDA 參數調配

在前一節方法架構的部分，我們有提到 LDA 的參數調配，其中 LDA 有三

個最主要的可變參數所控制，第一個為主題數=*K*，第二個為文檔-主題 Dirichlet 先驗參數α，接著是主題-詞 Dirichlet 先驗參數β。

3.2.1 Perplexity 曲線

在自然語言處理中，Perplexity 常用來度量語言模型的質量，值越小，模型質量越好[14]。因此 Perplexity 可以用來決定主題數 *K* 的數量。

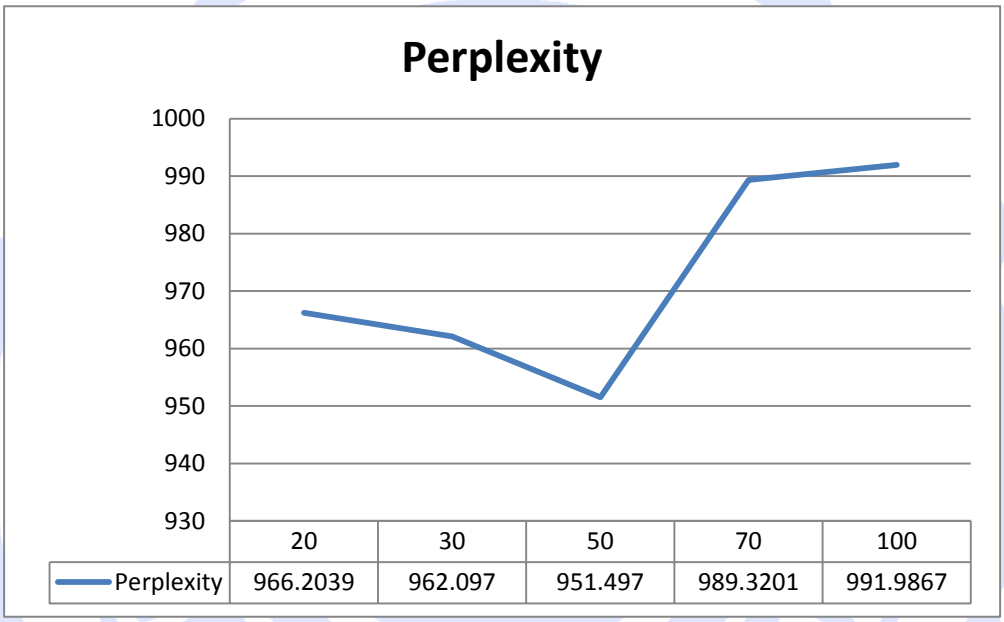


圖 3.5 PERPLEXITY 折線圖

由圖 3.5 可以明確地看出在主題數 *K*=50 的時候 Perplexity 有最小值，因此 *K*=50 為最佳的主題參數。

3.2.2 Dirichlet 先驗參數

Dirichlet 先驗參數α & β在許多文獻當中都表示，在α=0.5及β=0.1時有最佳效果，我們利用 500 篇 IEEE 論文作為參數測試集，其中已知 30 篇來自指定主題 Artificial intelligence，我們取α=0.1, 0.5, 2.5 與β=0.02, 0.1, 0.5 共 9 組進行實驗，計算其 Recall 與 Precision 來驗證哪一組的α與β數據為佳。

在信息檢索、分類、識別等領域中兩個最基本指標是召回率(Recall) 和準確率(Precision)，召回率也叫查全率，準確率也叫查准率。另外 f1 score 是 Precision 和 Recall 加權調和平均。

表 3.1 檢索相關表

	Relevant	Nonrelevant
Retrieved	true positives (tp)	false positives (fp)
Not retrieved	false negatives (fn)	true negatives (tn)

$$R = tp/(tp + fn)$$
$$P = tp/(tp + fp)$$

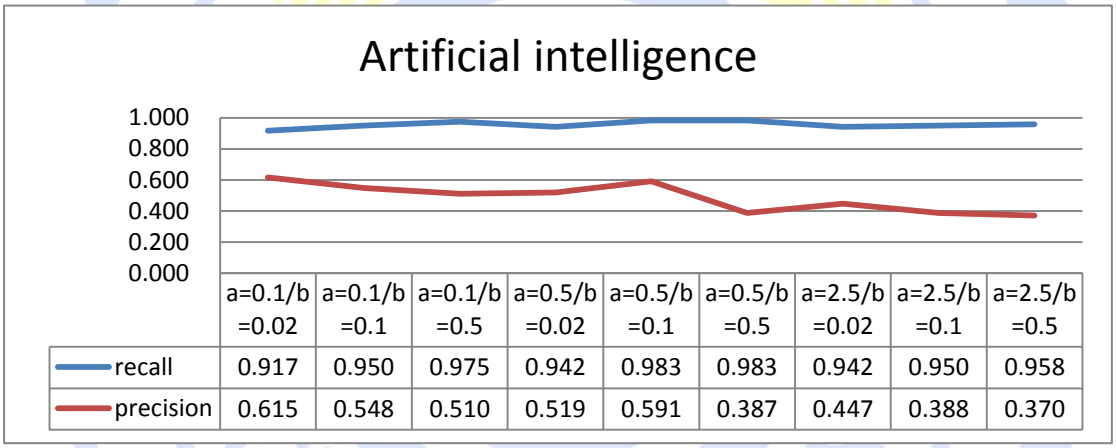


圖 3.6 RECALL&PRECISION 值折線圖

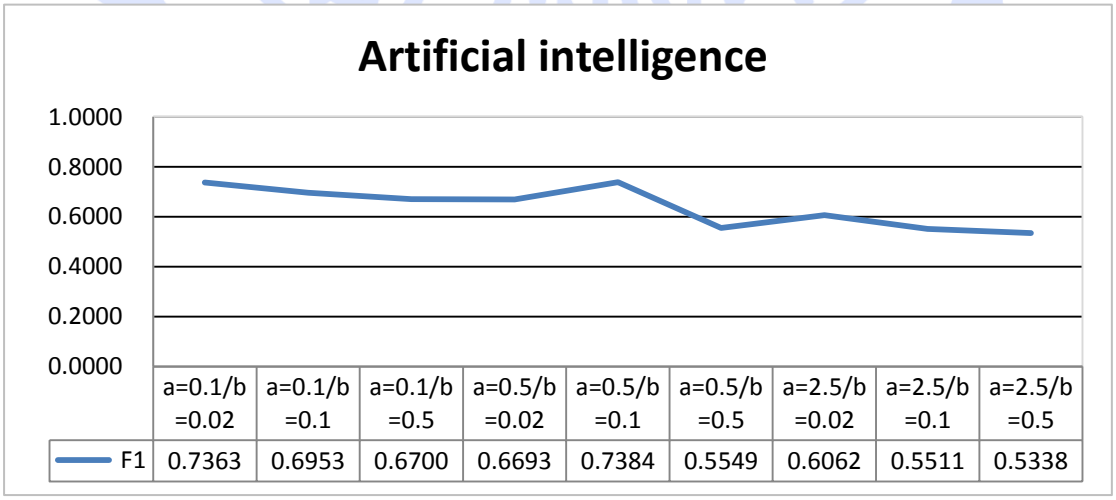


圖 3.7 F1 SCORE 值折線圖

由圖 3.6 中可以看出在α=0.5及β=0.1時有相當優秀的Recall 值與Precision 值，並且在圖 3.7 時也有最佳的 f1 score 值，因此我們在α及β的參數值選取α=0.5及β=0.1。

第四章結果討論分析

這個章節將展示本篇論文之實作成果，並且與 2013 IEEE taxonomy 進行比較，並且透過本論文輸出之結果進行討論與建議分類方法。

4.1 IEEE taxonomy

以 2013 IEEE taxonomy 為例，我們可以將 IEEE taxonomy 的分類看成是一個四層的 Tree，其中最上層總共有 51 大類，第二層總共有 675 類，第三層有 2467 類，第四層有 2999 類。

4.2 LDA 分類結果

執行 LDA 輸出結果會生成幾個檔案，首先是描述文檔與主題之間關係的 theta 檔，其次為描述主題與詞之間關係的 phi 檔，以及最主要的 twords 檔為 50 個主題及其主題下詞的結果，我們將完整主題輸出結果置於附錄。

LDA 輸出之結果後將其命名，並且與 2013 IEEE taxonomy 進行比較，如圖 4.1 所示，我們將比較結果以四種顏色做為區別，黑色字體的類別為在 2013 IEEE taxonomy 與 LDA 皆是相同的分類結果，紅色字體的類別為在 2013 IEEE taxonomy 有這個類別，但透過 LDA 分類後的結果並不明顯，因此在 LDA 中無此類別，藍色字體的類別為 2013 IEEE taxonomy 有此類別，在 LDA 的分類結果中也有這個類別結果，並且不只一個輸出主題，意思也就是在 2013 IEEE taxonomy 的一個類別結果，在 LDA 的輸出中不止為一個類別，可能為兩個甚至三個或四個，而這兩三個類別的結果往往可以在 2013 IEEE taxonomy 的下一層中找到相對應的子類別，最後綠色字體的類別結果為，LDA 輸出結果有此類別，但在 2013 IEEE taxonomy 的類別表中無法找到非常相對應之類別，可以建議此類別為一個新類別，topic 欄位為 LDA 輸出時給予的主題編號。

2013 IEEE Taxonomy	LDA	topic
Aerospace and electronic systems	Aerospace and electronic systems	47
Antennas and propagation	Antennas and propagation	24
Broadcast technology		
Circuits and systems	Circuits and systems(Silicon On Insulator)	21
	Circuits and systems(Circuits)	31
Communications technology	Communications technology(Modulation)	6
	Communications technology(MIMO)	41
	Communications technology(Communication systems)	44
Components, packaging, and manufacturing technology	Components, packaging, and manufacturing technology(Semiconductor growth)	38
Computational and artificial intelligence		
Computers and information processing	Computers and information processing	10
Consumer electronics		
Control systems	Control systems	49
Dielectrics and electrical insulation	Dielectrics and electrical insulation	32
Education		
Electromagnetic compatibility and interference		
Electron devices	Electron devices(Semiconductor devices)	18
	Electron devices(Microelectromechanical systems)	29
Electronic design automation and methodology		
Engineering - general		
Engineering in medicine and biology	Engineering in medicine and biology	11
Engineering management	Engineering management((Research and development))	35
Geoscience and remote sensing	Geoscience and remote sensing	22
IEEE organizational topics		
Imaging	Imaging	4
	Imaging(Motion pictures)	5
	Imaging(Biomedical imaging)	13
Industrial electronics		
Industry applications	Industry applications(Security)	8
	Industry applications(Chemical technology)	25
	Industry applications(Machinery)	42
Information theory		
Instrumentation and measurement	Instrumentation and measurement(Pulse& Time measurements)	1
	Instrumentation and measurement(Fluid flow measurement)	2
	Instrumentation and measurement(Microscopy)	14
Intelligent transportation systems	Intelligent transportation systems	9
Lasers and electrooptics	Lasers and electrooptics(Lasers)	16
	Lasers and electrooptics(Semiconductor lasers)	20
Magnetics	Magnetics	28
Materials, elements, and compounds	Materials, elements, and compounds(Media & hysteresis)	33
	Materials, elements, and compounds(Compounds)	36
	Materials, elements, and compounds(Materials)	40
Mathematics	Mathematics(Optimization)	3
	Mathematics(Accuracy)	12
	Mathematics(Sequence)	23
	Mathematics(Statistics)	39
Microwave theory and techniques		
Nanotechnology	Nanotechnology	43
Nuclear and plasma sciences	Nuclear and plasma sciences(Radiation effects)	30
	Nuclear and plasma sciences(Plasma)	48
Oceanic engineering and marine technology		
Power electronics		
Power engineering and energy		
Product safety engineering		
Professional communication		
Reliability	Reliability	7
Resonance		
Robotics and automation		
Science - general	Science - general (Physics)	26
Sensors	Sensors	17
Signal processing	Signal processing (Noise)	19
	Signal processing (Amplifiers)	34
	Signal processing(Filters)	0
Social implications of technology		
Solid state circuits		
Superconductivity	Superconductivity	27
Systems engineering and theory	Systems engineering and theory	45
Systems, man, and cybernetics		
Ultrasonics, ferroelectrics, and frequency control	Ultrasonics, ferroelectrics, and frequency control	46
Vehicular and wireless technologies	Vehicular and wireless technologies	15
	Graphene	37

圖 4.1 2013 IEEE TAXONOMY 與 LDA 結果之比較

4.3 LDA 分類結果與 IEEE taxonomy 比較討論

我們將針對上一節中的幾種情形進行更仔細的討論，首先由於黑色字體的類別在 LDA 與 2013 IEEE taxonomy 的分類結果皆相同，因此我們進而討論其他種分類上相異的情形，先將針對藍色字體的類別情形進行討論，再對於紅色字體類別與綠色字體類別進行討論。

圖 4.2 的類別為 Circuits and systems，在 LDA 的分類結果輸出為 Silicon On Insulator 與 Circuits 兩個子類別，其中 Silicon On Insulator 縮寫為 SOI，此技術發明者為 IBM，大量被 IBM，以及一些半導體公司，例如 AMD 和 NVIDIA 所應用，因此相關論文的研究甚豐。

(Circuits and systems)			(Circuits and systems)		
(Silicon On Insulator)			(Circuits)		
Topic 21		202640.8495	Topic 31		239244.7415
	gate	12269.58548		circuit	13170.43043
	transistors	8347.621915		converter	8948.500893
	device	6149.737829		output	8419.862549
	devices	5316.811478		input	5698.281721
	transistor	4995.531763		switch	4072.029719
	channel	4235.673733		cmos	3945.197213
	leakage	4123.218853		circuits	3936.007888
	characteristic	3981.845919		converters	3820.206385
	oxide	3650.336676		amplifier	3766.165396
	mobility	2651.542122		supply	3099.463729
	drain	2649.61412		capacitor	2816.598263
	degradation	2534.149853		inverter	2748.896865
	threshold	2514.306217		operation	2747.492771
	capacitance	2351.348681		range	2377.228695
	MOSFETs	2288.141428		efficiency	2287.938737
	bias	2198.842003		applications	2267.555978
	SOI	1662.607152		switches	2255.527314
	breakdown	1609.192479		load	2168.233779
	stress	1579.15083		topology	2110.92592
	capacitors	1573.110898		prototype	1881.813913

圖 4.2 CIRCUITS AND SYSTEMS

圖 4.3 為 Communications technology，在 LDA 的分類中又將其分成 Modulation、MIMO 以及 Communication systems，其中「MIMO」的全名為「Multiple Input and Multiple Output」代表「多重輸入與多重輸出技術」，有別於傳統無線基地台是單線運作的設計，MIMO 是多重天線同時運作，使無線網路能以「多徑傳輸」，除了可加大資料傳輸量外，還能延長訊號的距離。由於 3G 及 4G 網路在近

幾年熱烈發展並且應用此技術，因此 MIMO 相關討論數量甚多。

(Communications technology)			(Communications technology)			(Communications technology)		
(Modulation)			(MIMO)			(Communication systems)		
Topic 06		162326.9686	Topic 41		212916.0684	Topic 44		250460.9216
	signal	13440.12146		channel	18336.83023		networks	19230.65598
	modulation	8628.406665		channels	9010.694058		traffic	5920.670116
	signals	5956.608542		interference	6059.422134		nodes	5096.924337
	transmission	5478.384426		capacity	5069.964573		protocol	4364.783174
	fiber	3519.358982		receiver	4996.29536		packet	4193.217263
	dispersion	3392.405712		rate	4009.166229		communicati	3235.529353
	modulator	2652.866739		MIMO	3595.945900471127		scheme	3235.312738
	bandwidth	1761.069588		diversity	3153.053961		node	2938.349426
	ratio	1737.043498		communicati	3144.161884		access	2826.929678
	crosstalk	1697.13397		scheme	2987.212804		delay	2653.164608
	conversion	1647.549708		radio	2983.565619		path	2496.077008
	wavelength	1584.518894		transmission	2703.387336		link	2466.143351
	amplitude	1572.069309		spectrum	2524.165922		bandwidth	2435.424395
	optical	1536.03311		information	2432.044063		throughput	2342.911083
	receiver	1397.304306		users	2319.116045		number	2338.33078
	scheme	1283.526466		error	2169.648929		protocols	2307.732706
	WDM	1281.697646		receivers	1951.815347		scheduling	2154.19423
	amplifier	1218.278619		relay	1860.683529		transmission	1931.752431
	carrier	1207.93166		transmitter	1820.725277		allocation	1920.640122
	modulators	1133.680085		SNR	1799.075039		resource	1905.278376

圖 4.3 COMMUNICATIONS TECHNOLOGY

圖 4.4 為 Electron devices，可分為 Semiconductor devices 與 Microelectromechanical systems，由於半導體的應用廣泛，使得這兩個類別得以凸顯。

(Electron devices)			(Electron devices)		
(Semiconductor devices)			(Microelectromechanical systems)		
Topic 18		160829.5559	Topic 29		166465.5412
	efficiency	10313.24422		silicon	10053.1161
	cell	8849.301088		process	9332.992032
	device	8482.332408		fabrication	5550.812204
	cells	8251.884833		devices	5492.56108
	devices	6338.919008		structures	4423.392369
	diodes	5237.690235		nanowires	3556.740942
	diode	3222.942857		technology	3271.486971
	light	2129.994988		device	2605.163361
	leds	1848.959038		nanowire	2412.378814
	heterojunctio	1834.424332		integration	2332.126348
	conversion	1768.115508		substrate	2323.684035
	bias	1596.354392		MEMS	2265.113322
	emitter	1512.291856		applications	2252.946014
	base	1419.042245		wafer	2203.384512
	characteristic	1402.694497		semiconduct	1918.237528
	junction	1371.89289		chip	1755.003421
	quantum	1195.547523		lithography	1583.17772
	led	1194.920858		etching	1362.08619
	enhancement	1134.3083		bonding	1205.373956
	region	1119.831935		interconnect	1193.294472

圖 4.4 ELECTRON DEVICES

圖 4.5 屬於 Imaging，但仔細觀察這三個類別分別屬於 Imaging、Motion pictures 與 Biomedical imaging。近幾年影像處理的技術相當熱門，不止原有 Imaging 這個類別的研究，在 911 事件後 Motion pictures 技術對於安全監視相關的方面有重大的需求，然後在健康意識抬頭及老人化的影響 Biomedical imaging 的相關研究論相當的多。

(Imaging)			**(Imaging)**			**(Imaging)**		
Topic 04		250258.5174	(Motion pictures)			(Biomedical imaging)		
	image	13857.37615	Topic 05		145094.2315	Topic 13		126448.4356
	images	8295.695905		video	8896.602923		imaging	9779.264278
	classification	6053.408578		scheme	7740.705963		images	4777.802534
	features	5113.572309		quality	5895.723252		image	4441.576431
	feature	4564.679013		coding	3718.160806		reconstructio	3797.897833
	information	3811.958708		motion	3582.860355		resolution	2395.548494
	recognition	3719.306906		wavelet	3426.669953		tomography	1916.05575
	object	3554.522716		compression	3411.851292		registration	1609.096421
	color	3268.907392		frame	2622.781509		tissue	1551.884305
	segmentatior	3258.947688		image	2445.708435		contrast	1463.74986
	methods	3070.548709		distortion	2363.219773		acquisition	1175.719358
	objects	2950.620293		rate	2245.275831		volume	1082.779856
	space	2295.667496		prediction	2100.54956		pet	1022.566519
	framework	2038.570329		quantization	1863.245201		coherence	963.820616
	face	2017.695542		schemes	1725.024578		resonance	837.179648
	extraction	1929.707383		frames	1475.633871		methods	800.4485924
	set	1910.169497		block	1400.918265		blood	782.3341007
	training	1764.085741		complexity	1365.04648		scanner	771.1032714
	representatio	1763.743245		bit	1341.778274		ultrasound	768.9595819
	accuracy	1659.279176		error	1280.344761		correction	759.434108
				coefficients	1201.527858		cancer	742.8997602

圖 4.5 IMAGING

圖 4.6 中 Industry applications 包含了許多類別，領域相當廣泛，但在 Security 與 Chemical technology 以及 Machinery 這三個領域的研究最為豐碩。

(Industry applications)			(Industry applications)			(Industry applications)		
(Security)			(Chemical technology)			(Machinery)		
Topic 08		243398.1232	Topic 25		145779.1713	Topic 42		143506.418
	management	4980.681633		pressure	8625.634133		motor	6912.094767
	security	4495.133808		gas	8419.243782		flux	4384.430273
	services	4400.325176		hydrogen	4291.882033		machine	4248.608053
	information	4207.240689		plasma	3900.086144		induction	3969.767628
	service	4102.518537		diamond	2731.296147		speed	3906.6445
	users	3303.25068		chemical	2609.842114		drive	3628.101065
	user	3200.761185		carbon	2216.429139		rotor	3045.982674
	web	3125.918701		treatment	2059.764319		magnet	2842.40221
	applications	2938.029277		air	1945.924027		torque	2831.09498
	internet	2535.440445		vapor	1869.13349		machines	2633.210624
	market	2388.256159		oxygen	1725.008365		stator	2328.989894
	framework	1965.535205		rate	1646.912889		coil	2309.502623
	access	1875.623098		water	1593.98114		motors	2270.429951
	application	1725.513966		reaction	1593.890073		generator	2194.905735
	attacks	1598.96959		nitrogen	1592.51154		core	1743.464361
	infrastructure	1594.146472		exposure	1578.625943		losses	1674.996462
	devices	1553.417041		species	1558.908499		drives	1549.876353
	computing	1540.1444		pressures	1370.956381		position	1521.894426
	architecture	1526.400135		reactor	1196.765485		currents	1344.587756
	environment	1505.759961		flow	1144.356195		force	1267.305496

圖 4.6 INDUSTRY APPLICATIONS

圖 4.7 可以看出這三類在 IEEE taxonomy 的分類中，皆屬於 Instrumentation and measurement 的類別，但仔細觀察這三個類別分別屬於其類別的 Pulse& Time measurements、Fluid flow measurement、Microscopy。

(Instrumentation and measurement)			(Instrumentation and measurement)			(Instrumentation and measurement)		
(Pulse& Time measurements)			(Fluid flow measurement)			(Microscopy)		
Topic 01		122710.4712	Topic 02		140372.3403	Topic 14		116414.5402
	pulse	10792.60837		stress	7573.011632		force	7551.340736
	response	6649.048429		flow	6178.793182		polymer	5050.992653
	pulses	4579.85322		velocity	3968.473149		probe	4601.166499
	delay	4264.935527		deformation	2631.316224		electrodes	3100.147696
	generation	3240.462292		strain	2257.19246		electrode	2900.987183
	amplitude	2687.273947		disk	2185.395313		tip	2805.381223
	oscillator	2213.422537		vibration	1923.779751		microscope	2704.110831
	recovery	2204.766077		actuator	1721.145125		sample	2445.880009
	clock	2163.410234		shock	1620.807921		contact	2395.91377
	synchronizat	2101.514533		fluid	1545.257912		poly	2222.37651
	timing	2043.864941		modulus	1455.061981		microscopy	2217.112482
	duration	1772.273668		mass	1332.037558		surfaces	2113.014844
	waveform	1747.376541		shear	1297.816142		cantilever	1657.773513
	rate	1730.571644		strength	1254.870621		AFM	1341.825758
	cycle	1566.201564		force	1203.910816		molecules	1244.447488
	transient	1494.5697		displacement	1149.14828		probes	1013.41728
	jitter	1487.982787		experiments	1138.975457		film	1003.986153
	times	1474.664539		friction	1132.196515		polymers	964.7007537
	impulse	1328.2036		actuators	1124.598083		resolution	927.2540257
	waveforms	1324.551275		forces	1089.193112		droplet	768.6731555

圖 4.7 INSTRUMENTATION AND MEASUREMENT

圖 4.8Lasers and electrooptics 相關的研究中，Semiconductor lasers 為現今科技實際應用的重要光電元件，像是在光纖通信、雷射列印、高密度光儲存、條碼掃瞄等皆需要應用到半導體雷射。

(Lasers and electrooptics)			(Lasers and electrooptics)		
(Lasers)			(Semiconductor lasers)		
Topic 16		153074.9514	Topic 20		221249.7383
	laser	25813.43351		quantum	14552.61669
	lasers	6556.91955		emission	11107.02974
	gain	5609.531484		band	10563.50429
	wavelength	5170.537641		spectra	6830.685037
	fiber	3191.895992		absorption	6597.930612
	threshold	3029.931688		photolumines	5115.893693
	output	3005.122814		excitation	4276.829862
	pump	2891.627291		dots	4147.083091
	semiconduct	2354.23628		gap	3992.209187
	cavity	2147.035812		intensity	3704.530449
	operation	2135.524846		peak	3589.098454
	range	1624.810931		states	3251.154437
	diode	1579.433516		wells	2745.570442
	mirror	1460.912043		luminescence	2602.860099
	emission	1445.805407		QDs	2423.935389
	raman	1390.714668		spectrum	2303.638482
	cascade	1377.226016		transitions	2221.135562
	ablation	1322.02753		peaks	1995.955991
	amplifier	1260.441524		spectroscopy	1945.933589
	wavelengths	1179.640458		raman	1906.263534

圖 4.8 LASERS AND ELECTROOPTICS

圖 4.9 為 Mathematics，而 Mathematics 在科學的應用面更加的廣泛，在 LDA 的分類中產生 Optimization、Accuracy、Sequence 與 Statistics 等四個子類別。這四個類別所能處理的。

(Mathematics)			(Mathematics)			(Mathematics)			(Mathematics)		
(Optimization)			(Accuracy)			(sequence)			(Statistics)		
Topic 03		223113.4282	Topic 12		208984.2725	Topic 23		137777.0991	Topic 39		218854.1769
	problem	16160.65879		functions	5855.544244		codes	8708.90505		distribution	14786.70423
	optimization	11388.97441		equations	5094.769856		code	5996.619269		function	11075.5358
	algorithms	8068.45085		equation	5078.689634		error	5760.216087		parameters	7972.39139
	solution	5680.215113		solution	4154.480396		sequence	2846.675623		values	7287.011826
	problems	5172.603775		boundary	4130.935887		sequences	2238.736727		density	6645.705218
	constraints	4730.782169		function	3937.628552		decoding	2216.835873		value	5964.290248
	search	3587.421239		fields	3170.303847		distance	2148.508381		parameter	4148.057034
	cost	3489.62517		approximatio	3160.742785		complexity	1807.911006		ratio	4092.066918
	solutions	3012.503566		domain	2768.734616		bounds	1738.165831		factor	4077.880818
	number	2983.881037		formulation	2573.211266		class	1732.891145		correlation	4040.766632
	set	2976.309568		basis	2509.111903		block	1715.118449		distributions	3886.639074
	selection	2436.941084		calculation	2347.212346		number	1700.394137		rate	3718.052891
	strategy	2301.897791		matrix	2306.631319		length	1529.353604		length	3680.333987
	programming	2252.588192		accuracy	2291.964485		decoder	1447.661075		number	3659.081049
	tree	2045.775121		problems	1872.380605		errors	1399.53525		variation	3471.110617
	methods	2026.736704		expansion	1844.324854		rate	1370.581548		size	3305.769412
	framework	1968.566651		transfer	1812.037127		construction	1308.022921		dependence	3045.874069
	decision	1882.459183		FDTD	1800.489787		probability	1271.037778		probability	3029.626921
	constraint	1831.560926		computation	1796.613639		information	1157.212767		carlo	2965.45167
	strategies	1761.124198		solutions	1776.448143		LDPC	1081.993386		monte	2841.831169

圖 4.9 MATHEMATICS

圖 4.10 為 Signal processing 的類別，這三個類別分別屬於 Signal processing 下第二層類別的 Filters、Noise、Amplifiers。可以看出在 Signal processing 相關的分類中 Filters、Noise、Amplifiers 有相當的討論，足以成為上層的類別。

(Signal processing)			(Signal processing)			(Signal processing)		
(Filters)			(Noise)			(Amplifiers)		
Topic 00		217615.9419	Topic 19		77772.08496	Topic 34		169340.1753
	estimation	14058.42627		noise	18206.79459		mode	12722.92571
	filter	10335.53674		source	6988.524721		modes	8444.510596
	matrix	5767.955446		sources	4449.627369		waveguide	7530.730554
	signal	4830.932027		page	1454.782648		fiber	7172.523999
	filters	4828.436386		first	1386.322589		coupling	5339.732284
	signals	3491.887374		battery	1320.784654		loss	4458.368173
	error	3271.343339		vol	1242.284413		waveguides	3891.863296
	algorithms	2855.001281		figure	1194.341172		index	3817.857406
	estimator	2833.039789		fuel	1094.316547		cavity	3089.571258
	methods	2601.482635		low-frequenc	916.1155897		grating	2964.915708
	noise	2072.569933		appl	837.4662841		fibers	2863.986466
	estimates	2057.113775		authors	746.0453486		resonance	2847.492028
	problem	2009.811413		fluctuations	706.6196522		bragg	2586.881372
	vector	1895.281257		EMI	693.3004353		wavelength	2482.371156
	covariance	1678.863841		merit	663.3957389		gratings	2326.083295
	number	1647.112377		engine	642.5000305		resonator	2101.60677
	estimate	1570.277851		comments	585.3721832		resonators	1767.162901
	identification	1570.26738		author	555.9619979		crystal	1607.8707
	matrices	1568.002869					transmission	1534.086822
	parameters	1562.252992					core	1522.183579

圖 4.10 SIGNAL PROCESSING

另外紅色字體的類別為 IEEE taxonomy 有的類別但透過 LDA 的分類卻沒有相對應，可能原因在於 LDA 僅輸出 50 個類別，且在 LDA 的輸出中不止一個類別，可能為兩個甚至三個或四個對應到 IEEE taxonomy 中的同一個類別，造成類別數的不足，無給予足夠的類別對應到這些紅色字體的類別。

綠色字體為 LDA 輸出的類別，但在 IEEE taxonomy 中無法準確對應到的類別，如圖 4.11 所示 Graphene 為一種是世上最薄卻也是最堅硬的奈米材料，也是目前世上電阻率最小的材料，因此有許多論文再對此類別做相關研究。也是相當新興的研究項目。

(Graphene)		
Topic 37		128342.2464
	array	11612.21345
	arrays	6681.113146
	pattern	4478.382549
	graphene	4194.887802
	elements	3870.818504
	patterns	3653.653022
	shape	3477.724679
	wire	3219.602666
	edge	3042.855621
	cross	2847.119236
	direction	2768.477529
	section	2308.607447
	element	2254.602748
	wires	2149.668537
	plane	2006.276014
	ring	1971.424259
	planar	1696.864104
	configuration	1672.205674
	width	1489.913414
	geometry	1404.537134

圖 4.11 GRAPHENE

4.4 ThemeRiver

從可視化工具的結可以幫我們更容易看出這些類別在時間軸下的強弱狀況，圖 4.11 為類別強度前五強者，可以很明顯看到這五個主題的論文量持續增加穩定成長。這五類分別為 Security、Antennas and propagation、Circuits、Research and development、Systems engineering and theory。

而末五位則可以看出較大的起伏狀態，分別為 Biomedical imaging、Noise，Geoscience and remote sensing、Physics、Aerospace and electronic systems。其中物理類在 2005 年強度大幅下降，可能原因為當年度的文檔數量不足被稀釋了。

透過這樣的可視化呈現我們也能看出這些類別的脈絡，從中看出各個主題的生命週期。

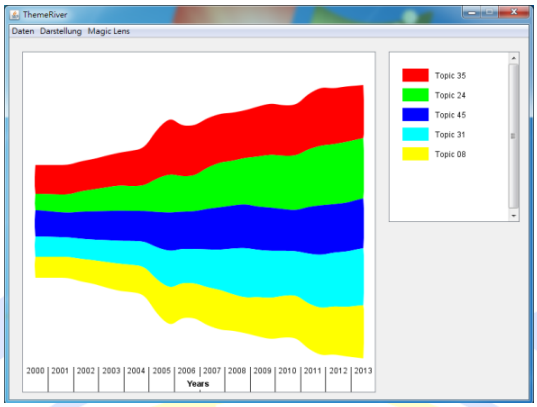


圖 4.11 主題強度前五位

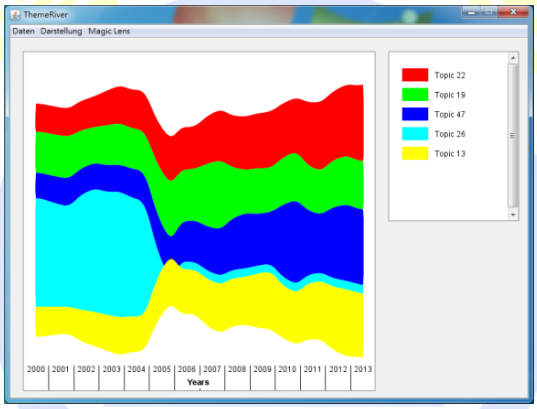


圖 4.12 主題強度末五位

第五章結論與未來展望

本研究致力於分類文題的探討，利用真實世界大量的期刊論文進行分類，並利用主題模型的方法進行分類其分類結果在與知名期刊的分類進行比較討論，並且從實驗結果透過各個主題的強度可以看出許多主題明顯被低估，而某些主題則被明顯高估，因此可以透過本研究的實驗結果進行分類的建議。

然而一般的分類表往往為一個階層的分類表，但 LDA 的分類結果只能給予最上層類別的建議，因此如果透過 LDA 輸出結果中的 theta 檔計算文檔之間的相似度關係，再透過其連結的權重再配合文檔間詞頻的關係建構階層的上下關係，相信能夠有新的發現及展望。

同時利用 LDA 分類算法運用在期刊論文上，獲得其結果，相信此方法流程能夠廣泛運用在其他專業的文檔的分類上，例如企業文件、醫療文件...等，相信能夠從輸出結果中發現一些未知或者不夠被重視的隱含主題。

參考文獻

[1] Alan G Fraser, Frank D Dunstan (2010) ,“ On the impossibility of being expert”.

[2] Margaret H. Dunham (2003), “Data Mining: Introductory and Advanced Topics”, Prentice Hall.

[3] Joachims, T. (1998), “Text categorization with support vector machines: learning with many relevant features”. In Proceedings of ECML-98, 10th European Conference on Machine Learning (Chemnitz, DE), pp. 137–142 .

[4] Gongde Guo, Hui Wang, David Bell, Yaxin Bi and Kieran Greer (2003), “KNN Model-Based Approach in Classification”, Proc. ODBASE pp- 986 – 996.

[5] Yiming Yang And Christopher G. Chute Mayo Cllnic (1994),“An Example-Based Mapping Method For Text Categorization And Retrieval” ACM Transactions On Information Systems, Vol. 12, No 3, Pages 252-277.

[6] Mnish Mehta, Rakesh agrwal (1996),” SLIQ: A Fast Scalable Classifier for Data Mining” .

[7] SHI Yong-feng, ZHAO (2004), “Comparison of text categorization algorithm”, Wuhan university Journal of natural sciences.

[8] Coulter, Neal (chair); French, James; Glinert, Ephraim; Horton, Thomas; Mead, Nancy; Ralston, Anthony; Rada, Roy; Rodkin, Craig; Rous, Bernard; Tucker, Allen; Wegner, Peter; Weiss, Eric; Wierzbicki, Carol (January 21, 1998), "Computing Classification System 1998: Current Status and Future Maintenance Report of the CCS Update Committee" , Computing Reviews (New York, NY, USA: ACM): 1–5.

[9] Anne Kao (2001), “Re: Reuters Corpus problems,” trecfiltering@list.research.microsoft.com .

[10] G. Salton, A. Wong, and C. S. Yang (1975), “A Vector Space Model for

Automatic Indexing , ” Communications of the ACM, vol. 18, nr. 11, pages 613–620.

[11] Robertson, S. E. & Sparck Jones, K. (1976), “ Relevance weighting of search terms. ” Journal of the American Society for Information Science, 27, 129-146.

[12] Blei, David M.; Ng, Andrew Y.; Jordan, Michael I. Lafferty, John, ed (2003), “Latent Dirichlet allocation” Journal of Machine Learning Research 3 (4–5): pp. 993–1022. doi:10.1162/jmlr.2003.3.4-5.993

[13] G. Heinrich(2005), “Parameter estimation for text analysis” Technical report.

[14]Havre, S. Hetzler, B. Nowell, L (2000), “ThemeRiver: visualizing theme changes over time” Information Visualization, 2000. InfoVis 2000. IEEE Symposium on.

附錄

(Signal processing)			(Instrumentation and measurement)			(Instrumentation and measurement)		
(Filters)			(Pulse& Time measurements)			(Fluid flow measurement)		
Topic 00		217615.9419	Topic 01		122710.4712	Topic 02		140372.3403
	estimation	14058.42627		pulse	10792.60837		stress	7573.011632
	filter	10335.53674		response	6649.048429		flow	6178.793182
	matrix	5767.955446		pulses	4579.85322		velocity	3968.473149
	signal	4830.932027		delay	4264.935527		deformation	2631.316224
	filters	4828.436386		generation	3240.462292		strain	2257.19246
	signals	3491.887374		amplitude	2687.273947		disk	2185.395313
	error	3271.343339		oscillator	2213.422537		vibration	1923.779751
	algorithms	2855.001281		recovery	2204.766077		actuator	1721.145125
	estimator	2833.039789		clock	2163.410234		shock	1620.807921
	methods	2601.482635		synchronizatior	2101.514533		fluid	1545.257912
	noise	2072.569933		timing	2043.864941		modulus	1455.061981
	estimates	2057.113775		duration	1772.273668		mass	1332.037558
	problem	2009.811413		waveform	1747.376541		shear	1297.816142
	vector	1895.281257		rate	1730.571644		strength	1254.870621
	covariance	1678.863841		cycle	1566.201564		force	1203.910816
	number	1647.112377		transient	1494.5697		displacement	1149.14828
	estimate	1570.277851		jitter	1487.982787		experiments	1138.975457
	identification	1570.26738		times	1474.664539		friction	1132.196515
	matrices	1568.002869		impulse	1328.2036		actuators	1124.598083
	parameters	1562.252992		waveforms	1324.551275		forces	1089.193112
(Mathematics)			**(Imaging)**			**(Imaging)**		
(Optimization)			Topic 04		250258.5174	(Motion pictures)		
Topic 03		223113.4282		image	13857.37615	Topic 05		145094.2315
	problem	16160.65879		images	8295.695905		video	8896.602923
	optimization	11388.97441		classification	6053.408578		scheme	7740.705963
	algorithms	8068.45085		features	5113.572309		quality	5895.723252
	solution	5680.215113		feature	4564.679013		coding	3718.160806
	problems	5172.603775		information	3811.958708		motion	3582.860355
	constraints	4730.782169		recognition	3719.306906		wavelet	3426.669953
	search	3587.421239		object	3554.522716		compression	3411.851292
	cost	3489.62517		color	3268.907392		frame	2622.781509
	solutions	3012.503566		segmentation	3258.947688		image	2445.708435
	number	2983.881037		methods	3070.548709		distortion	2363.219773
	set	2976.309568		objects	2950.620293		rate	2245.275831
	selection	2436.941084		space	2295.667496		prediction	2100.54956
	strategy	2301.897791		framework	2038.570329		quantization	1863.245201
	programming	2252.588192		face	2017.695542		schemes	1725.024578
	tree	2045.775121		extraction	1929.707383		frames	1475.633871
	methods	2026.736704		set	1910.169497		block	1400.918265
	framework	1968.566651		training	1764.085741		complexity	1365.04648
	decision	1882.459183		representation	1763.743245		bit	1341.778274
	constraint	1831.560926		accuracy	1659.279176		error	1280.344761
	strategies	1761.124198					coefficients	1201.527858
(Communications technology)			**(Reliability)**			**(Industry applications)**		
(Modulation)			Topic 07		212486.2958	(Security)		
Topic 06		162326.9686		test	8355.123021	Topic 08		243398.1232
	signal	13440.12146		fault	5849.845413		management	4980.681633
	modulation	8628.406665		reliability	5580.234394		security	4495.133808
	signals	5956.608542		load	4718.715782		services	4400.325176
	transmission	5478.384426		line	4295.89755		information	4207.240689
	fiber	3519.358982		transmission	3863.663387		service	4102.518537
	dispersion	3392.405712		grid	3643.829575		users	3303.25068
	modulator	2652.866739		failure	3414.672341		user	3200.761185
	bandwidth	1761.069588		distribution	3330.06912		web	3125.918701
	ratio	1737.043498		generation	3224.579849		applications	2938.029277
	crosstalk	1697.13397		protection	3024.743227		internet	2535.440445
	conversion	1647.549708		lines	3000.820193		market	2388.256159
	wavelength	1584.518894		faults	2946.685779		framework	1965.535205
	amplitude	1572.069309		wind	2708.737771		access	1875.623098
	optical	1536.03311		testing	2626.214912		application	1725.513966
	receiver	1397.304306		operation	2311.638327		attacks	1598.96959
	scheme	1283.526466		transformer	2282.840879		infrastructure	1594.146472
	WDM	1281.697646		tests	2203.845204		devices	1553.417041
	amplifier	1218.278619		equipment	1747.361864		computing	1540.1444
	carrier	1207.93166		cable	1674.344149		architecture	1526.400135
	modulators	1133.680085					environment	1505.759961

(Intelligent transportation systems)			**(Computers and information processing)**			**(Engineering in medicine and biology)**		
Topic 09		207538.1294	Topic 10		213957.5538	Topic 11		158403.4841
	charge	10063.99542		memory	9302.470549		activity	3430.921768
	transport	9592.264645		architecture	6285.910597		processing	2605.500484
	electron	7979.480419		hardware	4754.345		task	2455.991531
	interface	7375.815562		implementation	4129.114881		brain	2036.557069
	carrier	6359.352934		logic	3484.494717		speech	1709.873108
	barrier	5877.990682		storage	3223.188648		information	1653.923976
	mobility	4186.330668		applications	3057.973431		cortex	1558.977305
	states	4131.817741		processing	2682.363033		study	1506.431361
	density	3862.50927		processor	2375.676804		group	1325.319732
	hole	3762.957178		software	2260.918804		memory	1314.473251
	contact	3706.141046		consumption	2133.764536		responses	1286.643854
	resistance	3457.314406		architectures	1965.639085		activation	1238.579965
	conduction	3256.217873		circuits	1952.043939		response	1234.660401
	injection	3078.854076		area	1854.05192		subjects	1233.731674
	electrons	2819.960978		operations	1836.144777		regions	1216.725557
	contacts	2740.471619		number	1601.416105		stimuli	1212.927067
	schottky	2688.250748		techniques	1568.176751		stimulation	1206.803493
	level	2252.099893		operation	1450.544295		neurons	1185.958031
	conductance	2184.267072		designs	1446.423654		action	1146.140158
	characteristics	2141.425669		module	1443.714019		attention	1077.417772
(Mathematics)			**(Imaging)**			**(Instrumentation and measurement)**		
(Accuracy)			(Biomedical imaging)			(Microscopy)		
Topic 12		208984.2725	Topic 13		126448.4356	Topic 14		116414.5402
	functions	5855.544244		imaging	9779.264278		force	7551.340736
	equations	5094.769856		images	4777.802534		polymer	5050.992653
	equation	5078.689634		image	4441.576431		probe	4601.166499
	solution	4154.480396		reconstruction	3797.897833		electrodes	3100.147696
	boundary	4130.935887		resolution	2395.548494		electrode	2900.987183
	function	3937.628552		tomography	1916.05575		tip	2805.381223
	fields	3170.303847		registration	1609.096421		microscope	2704.110831
	approximation	3160.742785		tissue	1551.884305		sample	2445.880009
	domain	2768.734616		contrast	1463.74986		contact	2395.91377
	formulation	2573.211266		acquisition	1175.719358		poly	2222.37651
	basis	2509.111903		volume	1082.779856		microscopy	2217.112482
	calculation	2347.212346		pet	1022.566519		surfaces	2113.014844
	matrix	2306.631319		coherence	963.820616		cantilever	1657.773513
	accuracy	2291.964485		resonance	837.179648		AFM	1341.825758
	problems	1872.380605		methods	800.4485924		molecules	1244.447488
	expansion	1844.324854		blood	782.3341007		probes	1013.41728
	transfer	1812.037127		scanner	771.1032714		film	1003.986153
	FDTD	1800.489787		ultrasound	768.9595819		polymers	964.7007537
	computation	1796.613639		correction	759.434108		resolution	927.2540257
	solutions	1776.448143		cancer	742.8997602		droplet	768.6731555
(Vehicular and wireless technologies)			**(Lasers and electrooptics)**			**(Sensors)**		
Topic 15		140049.6201	(Lasers)			Topic 17		165076.2999
	motion	4331.441549	Topic 16		153074.9514		sensor	17596.63209
	vehicle	3056.717224		laser	25813.43351		measurement	16521.91084
	robot	2713.881935		lasers	6556.91955		sensors	8719.482168
	environment	2238.615712		gain	5609.531484		sensitivity	7316.159429
	position	2071.988477		wavelength	5170.537641		calibration	3933.120476
	display	1878.178914		fiber	3191.895992		range	3683.686264
	vehicles	1848.344098		threshold	3029.931688		monitoring	2839.266894
	body	1612.513647		output	3005.122814		accuracy	2686.736712
	positioning	1515.882917		pump	2891.627291		characterization	2135.697793
	environments	1484.73945		semiconductor	2354.23628		resolution	1978.173982
	navigation	1434.949361		cavity	2147.035812		response	1914.052134
	robots	1305.962753		operation	2135.524846		instrument	1655.47162
	movement	1250.05193		range	1624.810931		applications	1648.15303
	trajectory	923.3689637		diode	1579.433516		reference	1634.816227
	aircraft	913.0760502		mirror	1460.912043		setup	1554.267368
	hand	906.6725246		emission	1445.805407		uncertainty	1486.043455
	road	878.6977703		raman	1390.714668		device	1462.477728
	gps	819.7824439		cascade	1377.226016		signal	1406.0826
	panel	788.9729117		ablation	1322.02753		precision	1321.201791
	safety	758.2005886		amplifier	1260.441524		techniques	1265.0388
				wavelengths	1179.640458			

(Electron devices)			**(Signal processing)**			**(Lasers and electrooptics)**		
(Semiconductor devices)			(Noise)			(Semiconductor lasers)		
Topic 18		160829.5559	Topic 19		77772.08496	Topic 20		221249.7383
	efficiency	10313.24422		noise	18206.79459		quantum	14552.61669
	cell	8849.301088		source	6988.524721		emission	11107.02974
	device	8482.332408		sources	4449.627369		band	10563.50429
	cells	8251.884833		page	1454.782648		spectra	6830.685037
	devices	6338.919008		first	1386.322589		absorption	6597.930612
	diodes	5237.690235		battery	1320.784654		photoluminesce	5115.893693
	diode	3222.942857		vol	1242.284413		excitation	4276.829862
	light	2129.994988		figure	1194.341172		dots	4147.083091
	leds	1848.959038		fuel	1094.316547		gap	3992.209187
	heterojunction	1834.424332		low-frequency	916.1155897		intensity	3704.530449
	conversion	1768.115508		appl	837.4662841		peak	3589.098454
	bias	1596.354392		authors	746.0453486		states	3251.154437
	emitter	1512.291856		fluctuations	706.6196522		wells	2745.570442
	base	1419.042245		EMI	693.3004353		luminescence	2602.860099
	characteristics	1402.694497		merit	663.3957389		QDs	2423.935389
	junction	1371.89289		engine	642.5000305		spectrum	2303.638482
	quantum	1195.547523		comments	585.3721832		transitions	2221.135562
	led	1194.920858		author	555.9619979		peaks	1995.955991
	enhancement	1134.3083					spectroscopy	1945.933589
	region	1119.831935					raman	1906.263534
(Circuits and systems)			**(Geoscience and remote sensing)**			**(Mathematics)**		
(Silicon On Insulator)			Topic 22		149796.4267	(sequence)		
Topic 21		202640.8495		SAR	3311.901573	Topic 23		137777.0991
	gate	12269.58548		satellite	3011.832719		codes	8708.90505
	transistors	8347.621915		radar	2257.796491		code	5996.619269
	device	6149.737829		soil	1776.949488		error	5760.216087
	devices	5316.811478		water	1527.654708		sequence	2846.675623
	transistor	4995.531763		aperture	1424.186123		sequences	2238.736727
	channel	4235.673733		earth	1350.504313		decoding	2216.835873
	leakage	4123.218853		microwave	1300.332005		distance	2148.508381
	characteristics	3981.845919		moisture	1286.743742		complexity	1807.911006
	oxide	3650.336676		sea	1259.787905		bounds	1738.165831
	mobility	2651.542122		ice	1131.894338		class	1732.891145
	drain	2649.61412		ocean	1121.940658		block	1715.118449
	degradation	2534.149853		land	1111.902399		number	1700.394137
	threshold	2514.306217		resolution	1055.46447		length	1529.353604
	capacitance	2351.348681		observations	1053.343591		decoder	1447.661075
	MOSFETs	2288.141428		space	1049.707688		errors	1399.53525
	bias	2198.842003		wind	1005.395591		rate	1370.581548
	SOI	1662.607152		areas	1004.225967		construction	1308.022921
	breakdown	1609.192479		ground	951.1685586		probability	1271.037778
	stress	1579.15083		retrieval	910.5635313		information	1157.212767
	capacitors	1573.110898					LDPC	1081.993386
(Antennas and propagation)			**(Industry applications)**			**(Science - general)**		
Topic 24		242875.526	(Chemical technology)			(Physics)		
	antenna	18652.9535	Topic 25		145779.1713	Topic 26		100510.4644
	ghz	10928.3396		pressure	8625.634133		physics	21871.70714
	bandwidth	6437.55423		gas	8419.243782		institute	20602.20956
	microwave	6410.748565		hydrogen	4291.882033		American	20026.8782
	antennas	5431.856617		plasma	3900.086144		relaxation	3324.637197
	impedance	4645.721105		diamond	2731.296147		resonance	2589.825357
	radiation	4150.703944		chemical	2609.842114		dependence	1026.089255
	filter	4126.316909		carbon	2216.429139		sample	962.2409118
	microstrip	3491.849569		treatment	2059.764319		electron	952.3846438
	band	3325.358354		air	1945.924027		range	806.4779413
	loss	2831.92041		vapor	1869.13349		temperatures	802.8185027
	line	2825.248353		oxygen	1725.008365		samples	795.4466877
	patch	2630.546133		rate	1646.912889		article	769.1302837
	frequencies	2615.729639		water	1593.98114		experiments	455.6919364
	ground	2479.201608		reaction	1593.890073		spectra	423.6039746
	slot	2436.054607		nitrogen	1592.51154		squid	408.5162985
	mhz	2357.891613		exposure	1578.625943		nmr	380.9154853
	gain	2349.646993		species	1558.908499		fields	371.4259596
	applications	2257.077946		pressures	1370.956381		materials	366.020652
	transmission	2234.593294		reactor	1196.765485		interference	356.187736
				flow	1144.356195		order	346.8380046

(Superconductivity)			**(Magnetics)**			**(Electron devices)**		
Topic 27		229910.4342	Topic 28		158512.3031	(Microelectromechanical systems)		
	film	17705.10811		domain	7494.549441	Topic 29		166465.5412
	deposition	8940.521856		spin	6949.673103		silicon	10053.1161
	oxide	7024.993149		bias	4710.656332		process	9332.992032
	thickness	6152.179111		magnetization	4632.489135		fabrication	5550.812204
	growth	6130.645935		exchange	4384.971739		devices	5492.56108
	zno	5888.829047		junctions	3508.002843		structures	4423.392369
	electron	4980.521434		tunnel	3473.036651		nanowires	3556.740942
	interface	3780.851535		coupling	3457.600382		technology	3271.486971
	substrate	3434.182025		magnetoresistai	3110.568366		device	2605.163361
	grain	3353.686093		switching	2866.993389		nanowire	2412.378814
	layers	3088.682165		wall	2857.800756		integration	2332.126348
	substrates	3080.779238		fields	2822.36394		substrate	2323.684035
	formation	2906.80799		reversal	2421.936266		MEMS	2265.113322
	microscopy	2887.401984		layers	2314.978116		applications	2252.946014
	transmission	2697.511655		junction	2245.253127		wafer	2203.384512
	silicon	2620.020214		thickness	2058.554178		semiconductor	1918.237528
	chemical	2608.397369		dependence	1836.3007		chip	1755.003421
	diffraction	2575.990665		domains	1791.994481		lithography	1583.17772
	oxidation	2462.145613		state	1465.768694		etching	1362.08619
	annealing	2160.355504		vortex	1328.881869		bonding	1205.373956
							interconnects	1193.294472
(Nuclear and plasma sciences)			**(Circuits and systems)**			**(Dielectrics and electrical insulation)**		
(Radiation effects)			(Circuits)			Topic 32		208822.8587
Topic 30		145830.7741	Topic 31		239244.7415		transition	9700.946027
	detector	7321.815533		circuit	13170.43043		lattice	3492.669969
	radiation	5253.202208		converter	8948.500893		behavior	3216.286876
	detectors	4282.691443		output	8419.862549		alloys	3015.650466
	resolution	4188.394732		input	5698.281721		samples	2923.058669
	neutron	2452.693832		switch	4072.029719		change	2619.088606
	x-ray	1919.524087		cmos	3945.197213		alloy	2522.643072
	gamma	1595.07454		circuits	3936.007888		bulk	2410.83149
	spectrometer	1589.874031		converters	3820.206385		temperatures	2409.110758
	pixel	1571.094076		amplifier	3766.165396		range	2341.300586
	readout	1421.779867		supply	3099.463729		diffraction	2265.843167
	event	1369.247375		capacitor	2816.598263		room	2252.506904
	photon	1359.382725		inverter	2748.896865		state	2239.053195
	times	1248.309509		operation	2747.492771		composition	2141.720213
	kev	1223.674884		range	2377.228695		phases	2094.083484
	dose	1137.61345		efficiency	2287.938737		ceramics	1903.019799
	source	985.5580834		applications	2267.555978		crystal	1648.808267
	camera	952.9665742		switches	2255.527314		resistivity	1615.991243
	events	936.6007059		load	2168.233779		dependence	1591.908501
	range	925.3435603		topology	2110.92592		transformation	1554.838376
	rate	901.6080296		prototype	1881.813913			
(Materials, elements, and compounds)			**(Signal processing)**			**(Engineering management)**		
(Media & Hysteresis)			(Amplifiers)			(Research and development)		
Topic 33		121178.3108	Topic 34		169340.1753	Topic 35		241420.6783
	media	5383.598024		mode	12722.92571		research	6276.605937
	hysteresis	3864.073302		modes	8444.510596		technology	4643.631877
	magnetization	3650.663808		waveguide	7530.730554		development	4434.812926
	recording	3580.340602		fiber	7172.523999		engineering	4237.927463
	coercivity	2806.404109		coupling	5339.732284		article	3067.220524
	anisotropy	2488.794539		loss	4458.368173		years	2463.683296
	saturation	2298.594513		waveguides	3891.863296		technologies	2098.00732
	loops	2122.199082		index	3817.857406		project	2066.016527
	head	2079.038289		cavity	3089.571258		author	1921.378545
	grain	1774.628934		grating	2964.915708		software	1820.595434
	loop	1680.20255		fibers	2863.986466		students	1813.980651
	size	1505.061431		resonance	2847.492028		computer	1748.906697
	FePt	1468.973966		bragg	2586.881372		university	1739.032806
	deg	1356.588602		wavelength	2482.371156		industry	1565.64976
	permeability	1246.353861		gratings	2326.083295		issues	1530.804782
	magnets	1117.714086		resonator	2101.60677		applications	1481.209501
	medium	1088.443223		resonators	1767.162901		science	1415.616242
	kOe	1072.704038		crystal	1607.8707		review	1334.416858
	density	1055.121737		transmission	1534.086822		challenges	1319.42283
	grains	1016.41745		core	1522.183579		work	1285.562574

(Materials, elements, and compounds)			**(Graphene)**			**(Components, packaging, and manufacturing**		
(Compounds)			Topic 37		128342.2464	**technology)** (Semiconductor growth)		
Topic 36		171355.2934		array	11612.21345	Topic 38		163600.4365
	diffusion	6155.385462		arrays	6681.113146		growth	8022.874205
	defects	5793.159894		pattern	4478.382549		GaN	7456.82406
	defect	4469.967849		graphene	4194.887802		carbon	6738.030635
	concentration	4418.200633		elements	3870.818504		strain	5592.985607
	silicon	4343.471094		patterns	3653.653022		layers	5356.648607
	formation	4162.632509		shape	3477.724679		gaas	4678.584058
	annealing	3681.394867		wire	3219.602666		epitaxy	4045.382306
	irradiation	3475.493473		edge	3042.855621		nanotubes	3335.932503
	atoms	3310.269917		cross	2847.119236		nanotube	2731.919788
	ion	3019.633506		direction	2768.477529		electron	2612.283418
	oxygen	2805.986023		section	2308.607447		Ga0	2074.373172
	clusters	2369.087114		element	2254.602748		vapor	1830.659955
	implantation	2354.349684		wires	2149.668537		buffer	1808.26182
	activation	2120.178782		plane	2006.276014		beam	1790.840965
	samples	2054.694593		ring	1971.424259		substrates	1716.747229
	SiC	2025.404283		planar	1696.864104		aln	1655.822939
	damage	1947.249454		configuration	1672.205674		islands	1634.038938
	ions	1860.809872		width	1489.913414		dislocations	1616.0602
	cluster	1782.783403		geometry	1404.537134		sapphire	1601.333212
	lifetime	1677.438252					density	1586.746522
(Mathematics)			**(Materials, elements, and compounds)**			**(Communications technology)**		
(Statistics)			(Materials)			(MIMO)		
Topic 39		218854.1769	Topic 40		144659.8518	Topic 41		212916.0684
	distribution	14786.70423		materials	7648.013718		channel	18336.83023
	function	11075.5358		material	7071.589469		channels	9010.694058
	parameters	7972.39139		heat	4410.523452		interference	6059.422134
	values	7287.011826		conductivity	4031.096418		capacity	5069.964573
	density	6645.705218		breakdown	2729.332855		receiver	4996.29536
	value	5964.290248		resistance	2636.023777		rate	4009.166229
	parameter	4148.057034		heating	2499.70469		MIMO 3595.945900471127	
	ratio	4092.066918		gap	2467.832034		diversity	3153.053961
	factor	4077.880818		air	2270.809847		communication	3144.161884
	correlation	4040.766632		permittivity	1915.276145		scheme	2987.212804
	distributions	3886.639074		copper	1796.182187		radio	2983.565619
	rate	3718.052891		metal	1769.4694		transmission	2703.387336
	length	3680.333987		insulation	1768.494698		spectrum	2524.165922
	number	3659.081049		characteristics	1648.823079		information	2432.044063
	variation	3471.110617		samples	1580.036123		users	2319.116045
	size	3305.769412		water	1510.225443		error	2169.648929
	dependence	3045.874069		oil	1488.721438		receivers	1951.815347
	probability	3029.626921		strength	1448.325864		relay	1860.683529
	carlo	2965.45167		steel	1308.299224		transmitter	1820.725277
	monte	2841.831169		sample	1241.693712		SNR	1799.075039
(Industry applications)			**(Nanotechnology)**			**(Communications technology)**		
(Machinery)			Topic 43		139100.4935	(Communication systems)		
Topic 42		143506.418		nanoparticles	6033.936305	Topic 44		250460.9216
	motor	6912.094767		particles	5717.804488		networks	19230.65598
	flux	4384.430273		particle	4640.154969		traffic	5920.670116
	machine	4248.608053		size	4549.849836		nodes	5096.924337
	induction	3969.767628		glass	3910.705042		protocol	4364.783174
	speed	3906.6445		composites	1907.874027		packet	4193.217263
	drive	3628.101065		fluorescence	1695.550419		communication	3235.529353
	rotor	3045.982674		synthesis	1510.116429		scheme	3235.312738
	magnet	2842.40221		gold	1502.38929		node	2938.349426
	torque	2831.09498		nanoparticle	1446.910107		access	2826.929678
	machines	2633.210624		silica	1399.207593		delay	2653.164608
	stator	2328.989894		matrix	1376.16692		path	2496.077008
	coil	2309.502623		DNA	1287.92386		link	2466.143351
	motors	2270.429951		diameter	1252.515499		bandwidth	2435.424395
	generator	2194.905735		silver	1247.636143		throughput	2342.911083
	core	1743.464361		concentration	1219.195595		number	2338.33078
	losses	1674.996462		glasses	1150.554079		protocols	2307.732706
	drives	1549.876353		shell	1092.09685		scheduling	2154.19423
	position	1521.894426		solution	1065.213951		transmission	1931.752431
	currents	1344.587756		sizes	981.5867048		allocation	1920.640122
	force	1267.305496					resource	1905.278376

(Systems engineering and theory)			(Ultrasonics, ferroelectrics, and frequency control)			(Aerospace and electronic systems)		
Topic 45		257169.2013				Topic 47		118044.1841
	models	14218.55902	Topic 46		161679.0785		detection	18234.42655
	modeling	10938.10442		wave	8920.298493		target	7273.220337
	simulation	10733.87195		crystal	7278.176092		radar	4398.819944
	effects	8711.423614		polarization	6728.639059		tracking	2987.450086
	behavior	8439.892664		angle	5072.460057		targets	2874.890988
	parameters	8196.338282		waves	4652.462333		location	2364.601818
	simulations	6665.810337		propagation	4362.331175		signal	2328.279496
	theory	5911.711363		crystals	4242.559752		localization	2180.593812
	study	5759.205737		terahertz	2790.71129		fusion	1881.66303
	dynamics	3614.681663		reflection	2616.969428		processing	1594.871484
	methods	3205.5878		incident	2368.540032		range	1354.836159
	case	2987.335892		thz	2004.270155		doppler	1313.088608
	techniques	2838.674906		lens	1972.30736		clutter	1274.967681
	tool	2717.852475		diffraction	1938.082506		signature	1198.031134
	conditions	2682.697977		direction	1786.412942		signals	1186.631335
	characteristics	2596.501671		rotation	1757.878712		background	1052.043635
	agreement	2543.293298		liquid	1750.336961		detector	1051.574612
	experiments	2507.622249		scattering	1673.440698		information	910.9713978
	methodology	2466.289641		light	1668.028698		presence	809.8112434
	account	2394.129298		angles	1580.18093		test	776.3903273
				transmission	1573.056247			
(Nuclear and plasma sciences)			(Control systems)					
(Pasma)			Topic 49		158625.7152			
Topic 48		186699.5761		stability	8622.617361			
	plasma	14079.74299		controller	8192.403941			
	ion	10641.09064		state	5793.350806			
	beam	10146.44959		feedback	5295.158734			
	electron	8205.794812		compensation	2862.510879			
	source	5135.995988		scheme	2527.943004			
	density	3581.987389		conditions	2204.15595			
	ions	3108.131641		tracking	2184.945758			
	cathode	2349.055975		controllers	2097.04897			
	beams	2210.530789		effectiveness	1994.819407			
	emission	2159.644045		dynamics	1874.825537			
	plasmas	1923.268141		output	1806.9994			
	vacuum	1884.860999		disturbance	1564.257425			
	electrons	1612.975996		loop	1523.556602			
	arc	1376.745907		strategy	1522.785881			
	ionization	1335.159184		input	1464.540669			
	cyclotron	1163.746687		parameters	1401.567186			
	sources	1052.112276		reference	1377.627023			
	anode	1015.248272		disturbances	1371.121047			
	production	1000.459373		uncertainties	1327.314917			
	jet	997.2475609						