

文字探勘與紮根理論在技術現況分析的應用:以太陽能電池為例

洪朝富
真理大學資訊管理學系
au4076@au.edu.tw

邱賜福
真理大學工業管理與經營資訊學系
chiu@mail.au.edu.tw

吳宗錡
真理大學資訊管理學系
julh1991@hotmail.com

鄧祺弘
真理大學資訊管理學系
harry60078@hotmail.com

陳柏呈
真理大學資訊管理學系
peteras444@gmail.com

李侑璋
真理大學資訊管理學系
tonk120784@yahoo.com.tw

廖健名
真理大學資訊管理學系
z8051319@yahoo.com.tw

摘要

本研究以紮根理論分析步驟配合文字探勘(text mining)的質性機會發現(QCD, qualitative chance discovery)模型，不僅能夠在開放性編碼時看到全面性資料內容全貌，還能進一步經由紮根理論的主軸編碼，聚焦在某一主題內容的細部資訊，同時還可以經由紮根理論選擇性編碼進行資訊整合。本研究運用此一方法論分析薄膜太陽能電池專利，除了能夠瞭解太陽能電池的技術現況之外，還可以細部的觀察出新技術的浮現，與整合出主流與新技術間的關係，可以幫助研究者瞭解未來太陽能電池的可能發展方向。

關鍵字：紮根理論、文字探勘，質性機會探索模型，太陽能電池，專利分析

文字探勘與紮根理論在技術現況分析的應用:以太陽能電池為例

壹、前言

基於工業發展與科技之進步，對能源的需求不減反增，導致地球能源逐漸銳減，因而許多先進國家越來越重視環保議題，使環保意識逐漸抬頭，近年先進國家紛紛加入再生能源市場研究，固再生能源市場發展越來越熱絡，其特性為取之不盡，用之不竭，如太陽能發電、水力發電、風力發電、潮汐發電，但水力受限氣候、風力受限地理、潮汐發電不穩定等影響，反而太陽能發電不受地理影響，較無公害問題，是近年來備受注視的新能源。

因此為了解太陽能電池的發展現況，本研究蒐集薄膜太陽能電池專利相關資料，並發展以紮根理論分析步驟的文字探勘分析方法，進行專利資料分析，繪出薄膜太陽能電池發展現況，幫助研究者評估其技術走向和太陽能電池發展策略。

貳、文獻探討

由於專利資料屬於非結構性文字資料，以下我們將探討紮根理論與相關的分析技術，對解析文字性資料走向的現況。

2.1 紮根理論(Grounded theory)

質性研究(Qualitative research)又稱為質的研究，研究者為了深入探討某個問題，在自然情境下透過觀察、訪談或分析文件等等，廣泛的蒐集資料，分析資料內容與價值觀等(藍毓仁 2008)。紮根理論 Strauss & Corbin(2001)是質性研究方法的一種，在紮根理論(Grounded theory)是指將所收集到的資料打散、賦予概念(Conceptualized)，及再以新的方式將資料重新放在一起的操作化過程，幫助研究者形塑理論的分析程序是由開放性編碼對資料進行逐字逐句的登錄，在進行資料登錄時，研究者從資料中指認重要的事物等抽象表徵，並對這些表徵重新命名(類似文字探勘的同義詞處理)；主軸編碼的目的是為了將開放編碼中被分割的資料再加以類聚起來；選擇編碼是一個統整與精鍊類別的歷程，也就是發現核心類別和選取核心類別的準則((Strauss & Corbin 1990))。

因此紮根理論將生活中大量的研究資料及分析報告資料進行步驟程序化的分析，並發展與假設以經驗證過的理論，在諸多日常生活領域具有極高度的應用。但是紮根理論基本上是由人一字一句整理資料，所以必須耗費很多人力成本，因此發展合適的資訊化工具，與其配合使用是減少人力負擔的可行方法。

2.2 資料探勘(data mining)

資料探勘是一種處理結構化資料的技術，從已存在的資料庫中挖掘潛在、明確、實用資料的過程，除了能了解過去的資訊，並能挖掘未知的潛在價值(曾元顯 2004)。(以下為本研究對資料探勘的簡易流程圖示。)



圖 1 資料探勘流程图(蔡玉娟, 2003)

首先透過群集分析從資料庫中選取所要研究的目標專利，進行資料前處理去除無意義字詞，取出有意義字詞後進行資料倉儲建立，將儲存的資料進行探勘分析進一步得出有價值性的資訊，再將其資訊進行分類，最後運用關聯分析從分類中的大量相似資訊探討出未知的潛在價值、年度技術的發展性、預測未來技術方向的趨勢(蔡玉娟 2003)。然而面對文字形資料，則一般的文字探勘有其侷限性。

2.3 文字探勘

文字探勘(Text Mining)是資料探勘的延伸應用，其囊括資訊檢索萃取、計算語言學、自然語言處理、資料探勘、機器學習等跨領域知識，特別強調從非結構(un-structured)或半結構(semi-structured)的文字中發掘出未知、隱含且有用的資訊。Sullivan (2001)定義文字探勘為「一種編輯、組織及分析大量文件的過程，主要提供分析人員或決策者等特定使用者對特定資訊(如摘要、關鍵字)，發現資訊特徵及其間的關聯性」。與資料探勘相較之下，文字探勘需要額外的資料選擇處理程序及複雜的資料特徵萃取步驟，其後端通常以大量的文字資料庫為基礎，採用與詞彙出現的頻率及數量等相關數據進行分析，可應用於趨勢預測、犯罪分析、知識萃取、知識管理、決策輔助等領域(葉乃菁、王玳琪、張嘉珍、吳騏、賴志遠，2009)。

參、研究方法

如文獻探討如能在紮根理論的各分析步驟，導入文字探勘技術，不儘可保持紮根的優點，也能結合文字探勘的優點，如同Hong (2009)提出 Qualitative chance discovery model (QCD)。首先，專家從初始的開放式編碼分析得到的關聯圖中，發現有趣的主題字詞，與和主題相關的屬性字詞，用主題字詞和相關的屬性字詞萃取有關的文件資料，然後進一步的聚焦有趣單一價值主題，去除雜訊字詞讓浮現較為單純與容易理解的單一價值主題的字詞結構，重複上述的步驟直到所有的單一價值主題結構都被萃取出來。最後，分析者根據價值的情境統合相關的單一價值主題結構，只要結構中的點重複，就可以組合出多結構的價值情境，如同二分關聯網路，這種價值較為聚焦的多價值主題結構，可以幫助分析者發現單一價值與單一價值間短路徑的弱關聯連結點，就像機會點一樣，是發展未來情境的重要觸發點。

依據上述概念模型，本研究建立如下的研究流程與主要步驟說明如下：

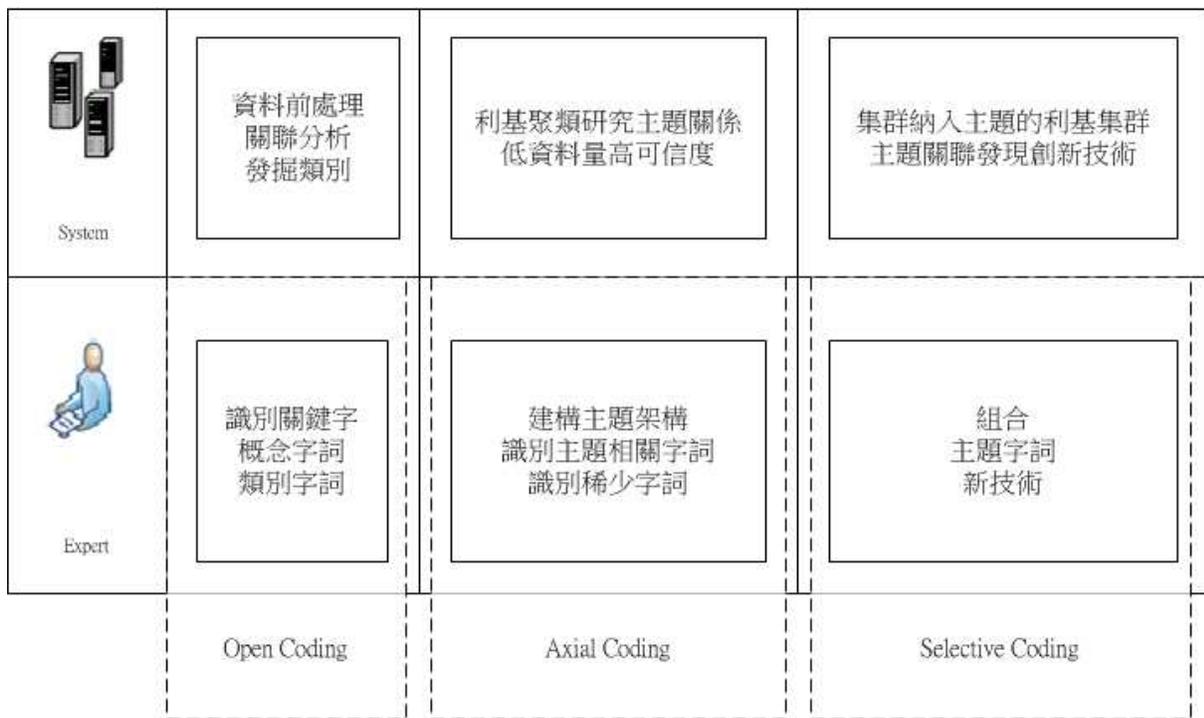


圖 2 質性研究架構圖

3.1 資料前處理

行斷詞處理與詞類分析；同時，進行停用字詞的研判與剔除；再者，並建立連詞檔案與同義字詞檔案，以進行同義字詞與連詞的歸併處理。以完成專利文件的資料清理，以整理出具意義的字詞(Meaningful terms)，以備後續探勘與分析之用。其歸類如下：

3-1-1H) 研究者定義他要研究的領域及相關的關鍵字詞。

3-1-1C) 研究者利用關鍵字詞從網路中篩選出相對應的文件資料。

3-1-2H) 基於他們的領域知識，研究者仔細閱讀文章同時針對文件進行斷詞、去除沒有意義的字詞，並且為字詞加上有意義的概念標籤。

3.2 關聯分析(開放式編碼)

開放性編碼就是藉著仔細檢驗而為現象取名字或加以分類的分析工作，也就是將資料轉為資訊，經過這個基礎步驟，我們才能將幾十到上百筆資料歸類為多個類別，經過此一步驟，看出類別的關聯，我們才能針對研究者，提出想法與假設，並進一步導出新發現。

2-2-1C) 使用公式計算所有字詞的關聯值。

N is all words

$i = 1$ to $N - 1$

$j = i + 1$ to N

$$assoc(w_i, w_j) = \sum_{s \in allD} \min(|w_i|_s, |w_j|_s)$$

在公式中， s 是代表文件中的所有句子； $allD$ 則代表所有文件。

3-2-2C) 將計算結果視覺化，系統可以幫助繪製共同出現字詞的關聯圖。

3-2-1H) 研究者解讀關聯圖並從中確認字詞的概念及其所歸屬的類別，同時給予每個類別一個名稱，用來幫助研究者辨識各個類別主題的價值。

3.3 聚焦關聯分析(主軸編碼)

主軸編碼，是將開放式編碼中被分割的類別和次類別重新整合以類聚起來，於是相同概念資料的連結又被組合再一起，它與開放性編碼的差異在於主軸編碼對分類的類別性質和向性具有較清楚的方向，是較「明確目標導向」的（林秀玲，2008）。

3-2-1H) 我們將所有以冠上詞性的字詞分類到各類別主題上行成一範疇。

3-2-1C) 範疇從第一個類別主題字詞到最後一個類別主題字詞。

3-2-2C) 類別主題，用來篩選出多量且是有意義的字詞

3-2-3C) 分析結果視覺化產生字詞共同出現的關聯圖，建構單一類別主題關聯圖，then goto 3-1-1C。

3.4 整合趨勢分析(選擇性編碼)

為了統合分析資料，必須由資料中闡述出一條故事線，也就是將核心部分的分析概念用敘述故事的方式概念化，利用主軸編碼中的其他類別關聯圖連結在一起，藉此來進行分析。

3-4-1H) 研究者定義相關的 Topics' 字詞和相關的 value words。

3-4-1C) 利用這些字詞進行資料篩選與關聯分析。

肆、實驗結果與討論

本研究以經濟部智慧財產局(TIPO)的專利資料庫為對象，做為太陽能電池專利文件的來源。進入智財局的專利資料庫、並輔以連穎科技的 Patent Guider 軟體，輸入薄膜太陽能電池，進行 2011 年太陽能電池專利資料的蒐集，共有 25 篇。

1.資料前處理：(標詞性(名詞、形容詞) 去除無意義的字

2.開放式編碼：整合開放式編碼與資料前處理

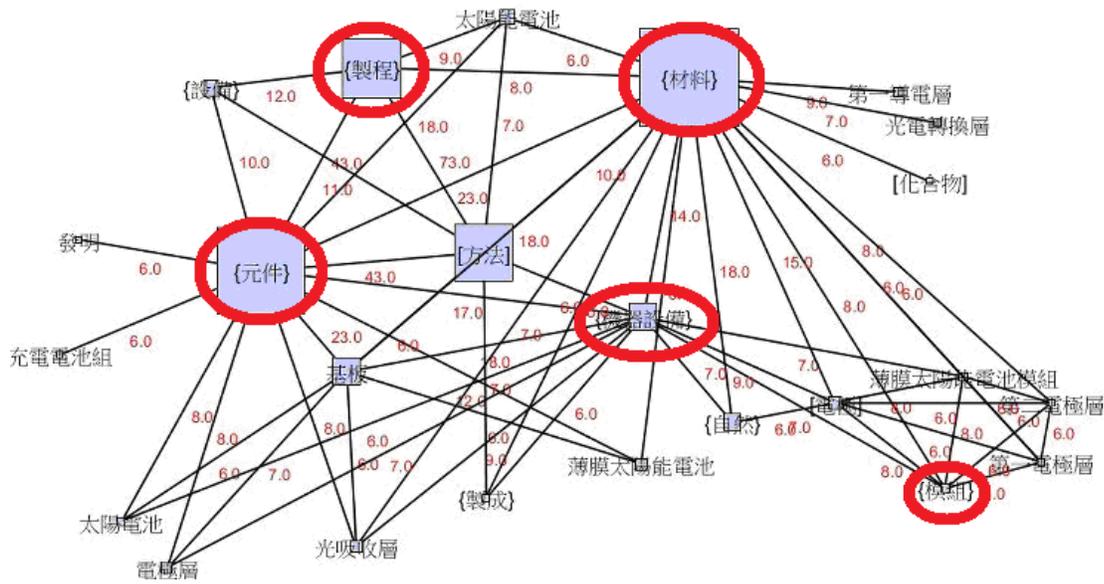


圖 3 薄膜太陽能電池關聯圖(頻率 2 關聯 6)

利用 3.1 的資料前處理進行分析得到圖 3，由此圖得知五大類。因圖 3 未能顯示出詳細的資料，而無法判斷其結果，利用主軸編碼分別得到詳細資料。

3.主軸編碼：總共有五個類別所組成，分別是元件、材料、製程、模組以及機器設備

以機器設備為關鍵字，再一次收集資料加以分析，得到研究圖(發現)

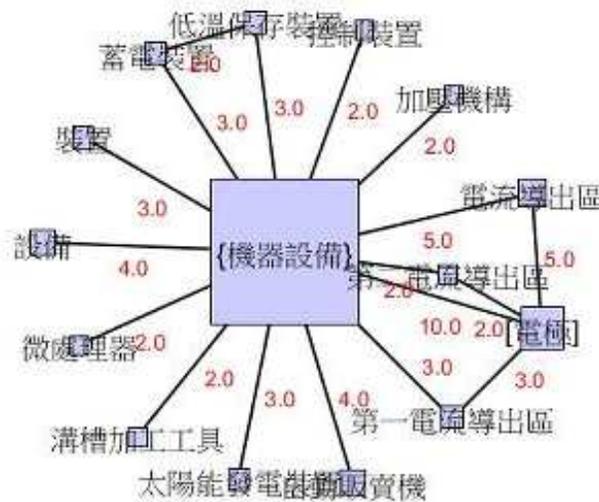


圖 4 機器設備(頻率 1 關聯 2)

由此圖所示，在此年度太陽能電池的機器設備，並沒有明顯的進步與擴展，由過去的一些機器設備，搭配較新的元件應用在太陽能電池上。

以模組為關鍵字，再一次收集資料加以分析，得到研究圖(發現)

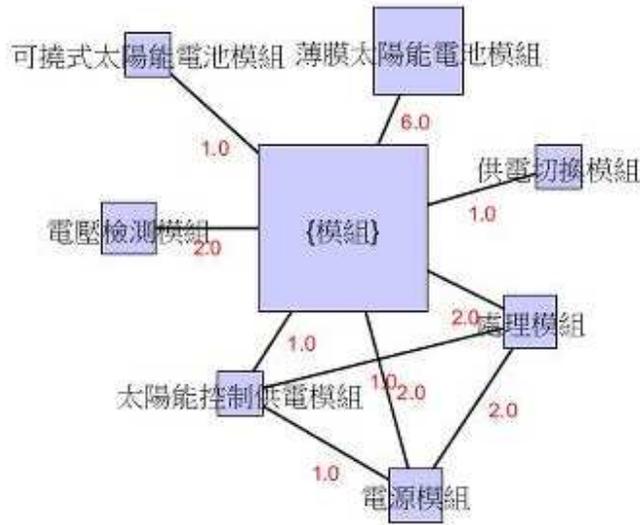


圖 5 模組(頻率 1 關聯 2)

可撓式太陽能電池包括基板、可撓式太陽能電池、接線、覆蓋層，其主要目的在於，太陽能電池以接線電性連接，再以覆蓋層將太陽能電池包覆在載體上，所以不需後續的塗裝作業，所以太陽能電池能夠方便且快速被包覆在載體上，且能夠依載體曲面覆蓋完成，以降低成本。

薄膜太陽能電池模組包括光電材料層、光電轉換層及發光材料反射層。具有發光材料反射層能讓薄膜太陽能電池模組獲得雙重光源以提升薄膜太陽能電池之發電效率。

以製程為關鍵字，再一次收集資料加以分析，得到研究圖(發現)

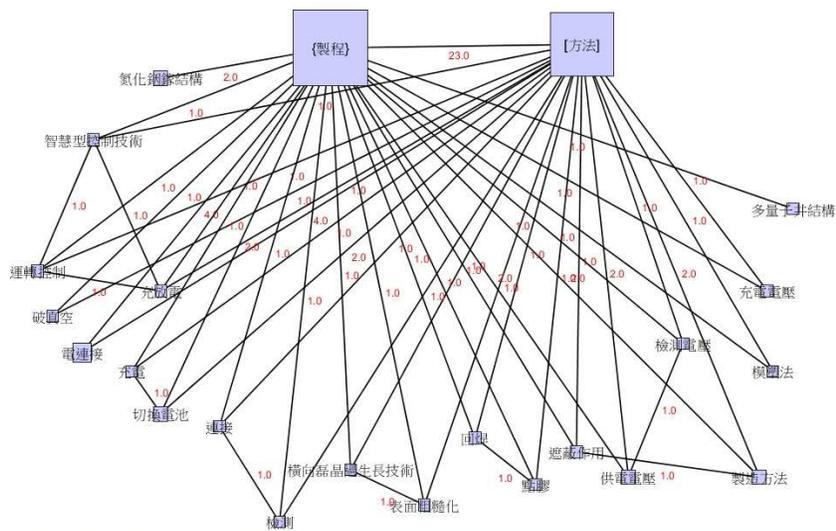


圖 6 製程(頻率 1 關聯 2)

智慧型控制技術以控制元件或控制系統以改變結構之特性，達到提升建物結構安全性、舒適性或功能性為目的。

以元件為關鍵字，再一次收集資料加以分析，得到研究圖(發現)

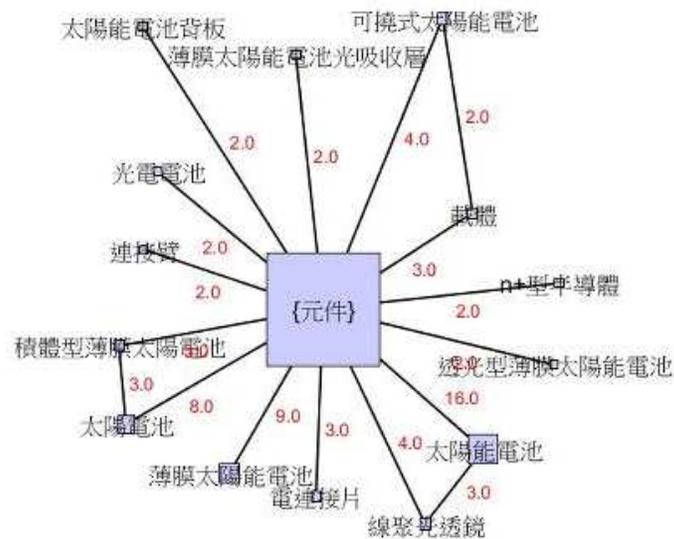


圖 7 元件(頻率 1 關聯 2)

根據圖中所示，看出新元件有可撓式太陽能電池、透光型薄膜太陽能電池、積體型薄膜太陽能電池等等，可撓式太陽能電池是可以捲曲的電池，可用於凹凸不平地區，透光型薄膜太陽能電池為一種可以與玻璃結合，呈現如同一般玻璃一樣可透光，不像一般不透光太陽能產品，這讓建築設計師有更大的對建築的設計空間。

以材料為關鍵字，再一次收集資料加以分析，得到研究圖(發現)

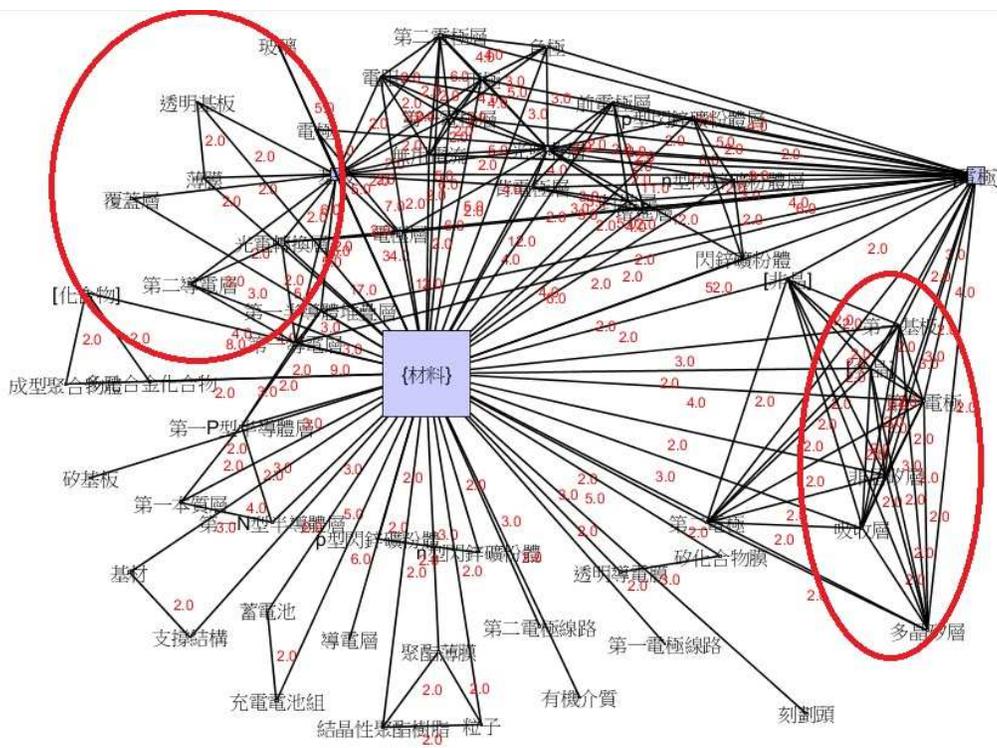


圖 8 材料(頻率 1 關聯 2)

由於此圖的關聯值較低，所以可以看到更多的材料在圖裡，從關聯圖中看到，關聯較密的是太陽能電池的基本材料，有電極層、多晶矽層、基板等等，而其餘關聯較低的則是其他類型太陽能電池的材料，像薄膜為薄膜太陽能電池的材料，而透明基板為透明行薄膜太陽能電池的材料等等。

5. 選擇性編碼:

機器設備

此圖是以開放式編碼中的關聯圖與主軸編碼之機器設備關聯圖結合所產生。

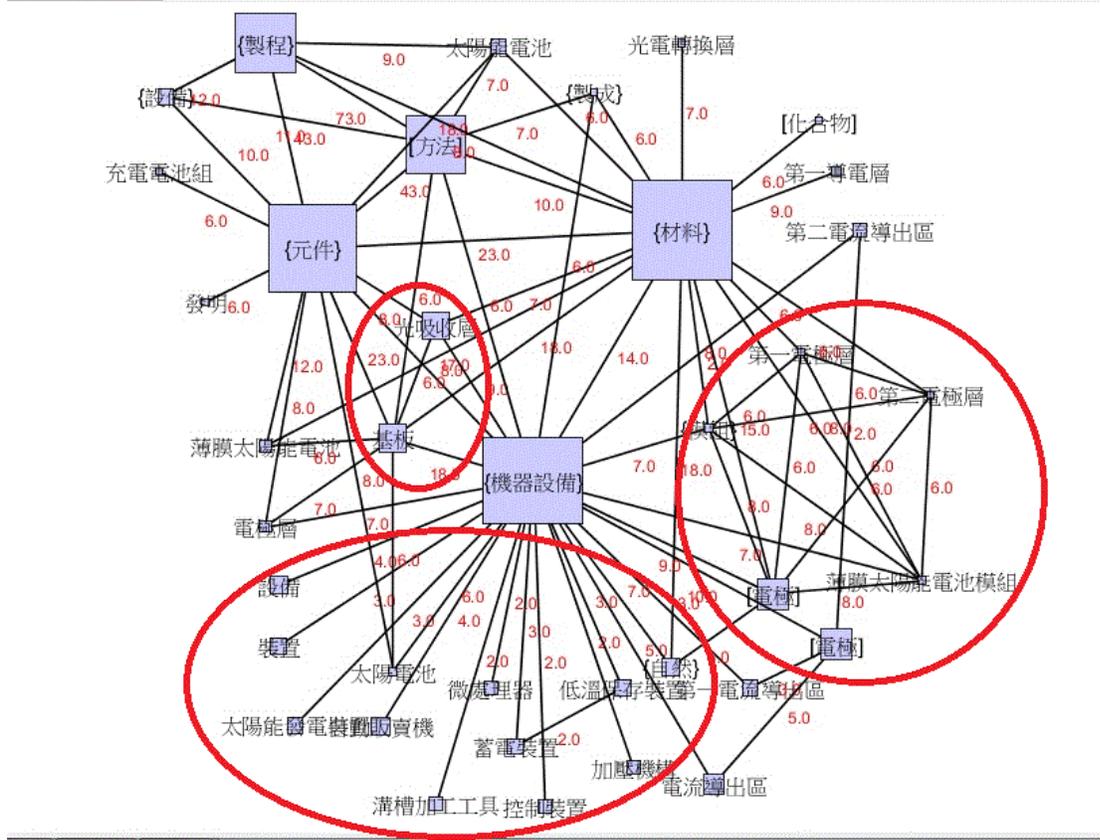


圖 9 機器設備 (頻率 1 關聯 2)

從圖中可以看出，使用的機器設備有控制裝置、加壓機構、低溫保存裝置、蓄電裝置、加壓機構、電極.....等等，可以看出機器設備的變動並不大，電極是頻率最高的機器設備，相對於電極機器設備而言，其他設備雖然出現頻率不高，但依舊是不可缺少的重要設備。

元件

此圖是以開放式編碼中的關聯圖與主軸編碼之元件關聯圖結合所產生。

製程

此圖是以開放式編碼中的關聯圖與主軸編碼之製程關聯圖結合所產生。

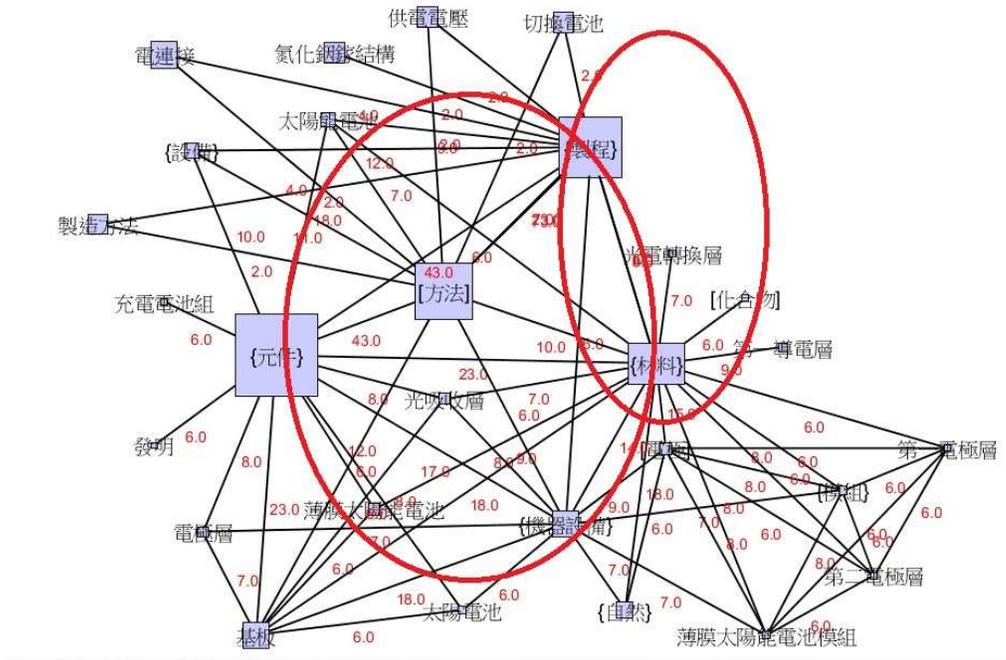


圖 12 製程(頻率 1 關聯 2)

由此圖可以得知，製程與、元件、材料、機器設備都有直接的關聯，除了模組之外。在這一張圖裡可以看到光電轉換層藉由材料與製程製作所產生的，光電轉換成又稱為光電轉換器它是利用材料的光電效應製作成的探測器。除了看到製程與材料的關係之外，光吸收層也是製程、元件、材料與機器設備的關聯，雖然製程只是一個技術方面的過程，但是製程是太陽能電池中不可或缺的一環。

模組

此圖是以開放式編碼中的關聯圖與主軸編碼之模組關聯圖結合所產生。

3. 陳達仁、黃慕萱，2009，專利資訊檢索、分析與策略，台北，華泰文化事業股份有限公司。
4. 顏怡文校訂，楊德仁編，2009，太陽能電池材料 Solar Cell Materials，台北，五南圖書出版股份有限公司。
5. 林明憲，2007，太陽電池技術入門，台北，全華出版社。
6. J. Ritchie & J. Lewis、藍毓仁，2008，質性研究方法，臺北市，巨流圖書。
7. 曾元顯，2004，『專利文字之知識探勘：技術與挑戰』，現代資訊組織與檢索研討會。
8. 蔡玉娟，2003，資料探勘之關聯法則發展與應用。
9. 盧俊宏，民 98，以紮根理論初探升遷動機 —以國營事業電力工員轉換為職員為例，國立中山大學企業管理學系研究所碩士論文。
10. Anselm Strauss. Juliet Corbin 著，質性研究入門：紮根理論研究方法，2001，濤時出版社。
11. 徐宗國譯，Strauss,A & Corbin,J 著，1990，Basics of Qualitative Research： Grounded Theory Procedures and Techniques.Newbury Park,CA:Sage，台北，巨流圖書公司，1997。
12. Sullivan, D. Document Warehousing and Text Mining. Wiley, 2001.
13. .Strauss , A & Corbin, J . (1990).Basics of Qualitative Research： Techniques and Procedures for Developing Grounded Theory，吳芝儀、廖梅花譯，質性研究入門:紮根理論研究方法，嘉義:濤石文化，2001。
14. Chao-Fu Hong: Qualitative Chance Discovery - Extracting competitive advantages，2009.

Using Text Mining and Grounded Theory for Technological Situation Analysis on Solar Cells

Chao-Fu Hong

Aletheia University Department of Information Management
au4076@au.edu.tw

Tzu-Fu Chiu

Aletheia University Department of Industrial Management and Enterprise Information
chiu@mail.au.edu.tw

Wu-Zong Qi

Aletheia University Department of Information Management
julh1991@hotmail.com

Deng-Qi Hong

Aletheia University Department of Information Management
harry60078@hotmail.com

Chen-Bai Cheng

Aletheia University Department of Information Management
peteras444@gmail.com

Li-Yi Zhang

Aletheia University Department of Information Management
peteras444@gmail.com

Liao-Jian Ming

Aletheia University Department of Information Management
z8051319@yahoo.com.tw

Abstract

A Qualitative Chance Discovery (QCD) model, formed by integrating Text Mining techniques and Grounded Theory (GT) processes, was applied in this study. The model's first step, similar to the open coding of GT, researcher can recognize topic's keywords from an association map and then to focus on that topic. Secondly, based on the selective coding he/she can organize the topic's information to discover innovation. Finally, after analyzing solar cell's patent data, researcher can find out the innovative techniques which are connected with other techniques that can help him/her to see the potential technology.

Keywords: grounded theory, text mining, qualitative chance discovery model, solar cell, patent analysis