網路爬蟲與資料分析
簡介

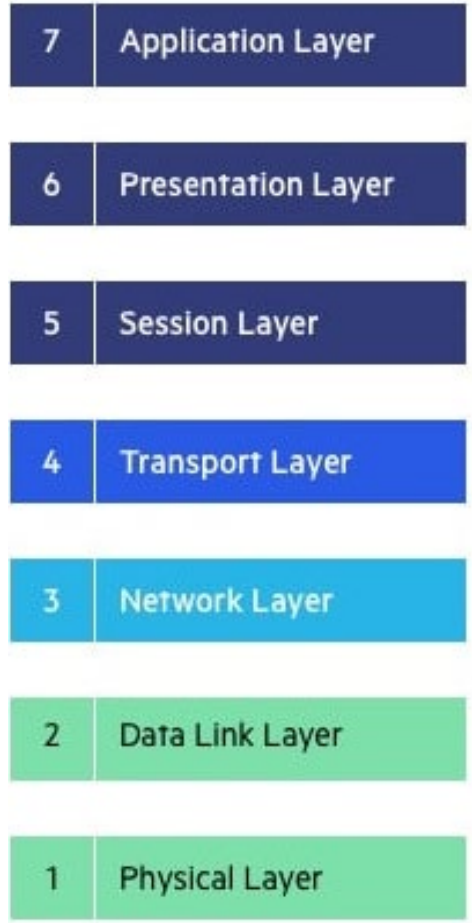
Instructor: 馬豪尚

什麼是網際網路(Internet)？

- 網際網路實際上並不是真正的網路，它是一個虛擬的概念，是由各種不同網路之間所串而連成的一個單一巨大國際網路，並在其上面提供網路服務。
- 為了能夠將各種不同網路連接起來，這些網路就必須以一組通用的協定相連。

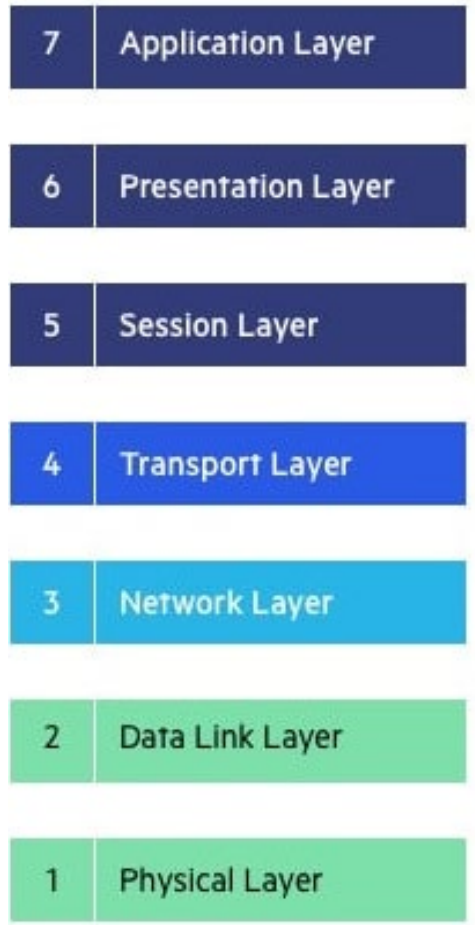
OSI 網路模型7層架構

- 應用層
  - 這層定義使用者的應用程式交換資料的方式
  - Web 瀏覽器和電子郵件客戶端
  - HTTPS、POP 、FTP
- 表示層
  - 這層負責準備資料以供應用層使用並定義資料格式的表現方式
  - 資料轉譯、加密和壓縮。
- 工作階段層
  - 這層負責處理開啟和關閉兩個裝置之間的通訊
  - 確保工作階段保持足夠長的開啟時間以傳輸所有進行交換的資料



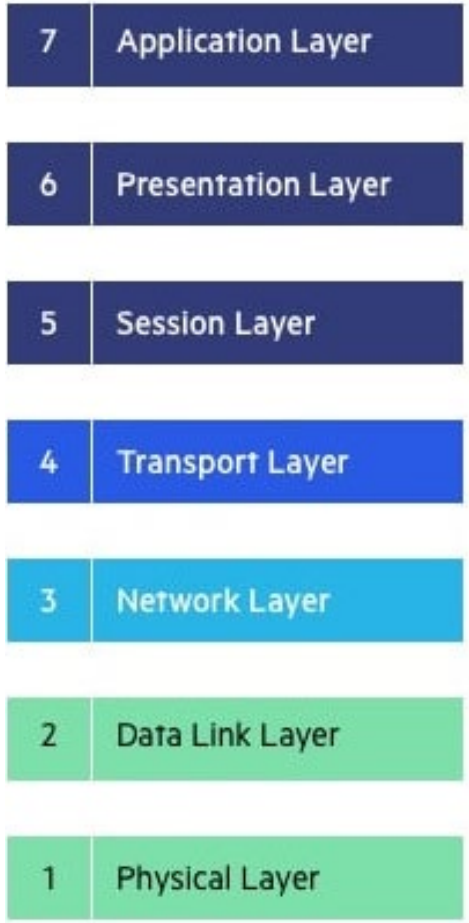
OSI 網路模型7層架構

- 傳輸層
  - 這層負責處理兩個裝置之間的端對端通訊
  - 從工作階段層取用資料，並在傳送至網路層之前分解為稱為"區段"的區塊
  - 接收裝置上的傳輸層負責將區段重組為工作階段層可以取用的資料
- 網路層
  - 這層負責促成兩個不同網路之間的資料傳輸
  - 在傳送者的裝置中將傳輸層中的"區段"分解為較小的單元(封包)
  - 接收端重新組裝這些封包



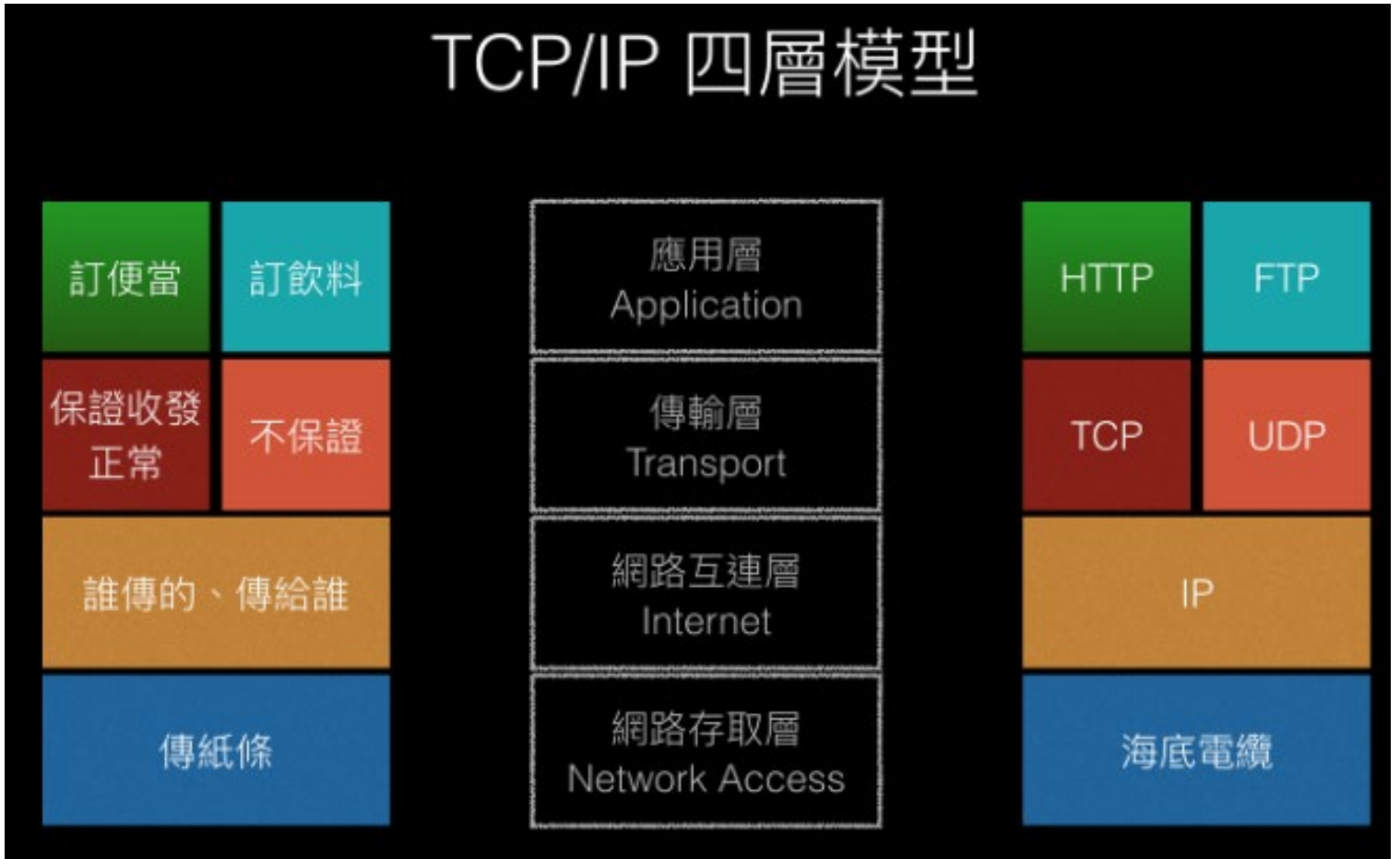
OSI 網路模型7層架構

- 資料連結層
  - 將網路層的封包分割成更小單位訊框 (Frame)
  - 負責網路內的流量控制以及錯誤控制
- 實體層
  - 這層負責網絡節點之間的物理有線或無線連接
  - 資料進一步轉換為 bit stream ，即為一連串的 0 與 1 字串，並轉換為傳輸介質所能傳輸的信號格式



TCP/IP 四層模型架構

› 現如今， 網際網路泛指以TCP/IP為主之通訊協定所架設而成之網路



全球資訊網 (World Wide Web)

- 全球資訊網是檔案、圖片、多媒體和其他資源的全球集合，可以理解為網際網路的一項服務，透過網際網路存取。
- 使用統一資源標誌符標識(URL)
  - 提供了一個全球命名標識系統，象徵性地標識服務、網頁伺服器、資料庫以及提供的檔案和資源
- 超文字傳輸協定(HTTP)
  - 全球資訊網的主要存取協定，全球資訊網的服務使用HTTP在軟體系統之間進行通訊和資料傳輸
- 超文字組成的系統，定義在超文字標記語言(HTML)內
  - 整體透過許多超連結互相連接，便於在資源之間導航

Web的組成要件

- 資源(Resource)
  - 嵌入在網頁上的文字、檔案、多媒體、互動式內容等
- 資源標識符（超連結Hyper link）
  - 為字符串，表示可能包含的通用地址
  - 例如https://ai.nutc.edu.tw/
- 傳輸協議(Transfer Protocol)
  - 規範瀏覽器之間的溝通方式
  - 例如http/https

使用統一資源標誌符標識(URL)

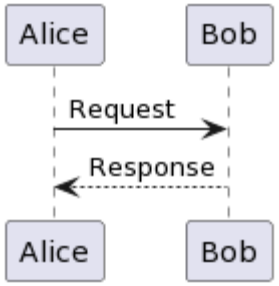
- 網路上的所有資源都是藉由一個URL來定位並存取
- 最早的URL為一長串的IP數字組成
- 後來演變成使用較容易識別的網域名稱以及網域名稱伺服器(DNS)來轉換IP位置並提供網域名稱

使用統一資源標誌符標識(URL)

- URL 由三部分組成
  - 安全協定 (https, ftp)
  - 網域名稱 (www.domain.com)
  - 文件路徑 (/directory/file.html)
- Example:
  - https://en.wikipedia.org/wiki/URL

安全協定　　網域名稱　　文件路徑

超文字傳輸協定(HTTP)



- 規範了客戶端請求與伺服器回應的標準，實際上是藉由 TCP 作為資料的傳輸方式。
- 例如使用者送出了一個請求，資料透過 TCP 協定傳遞給伺服器，並等待伺服器回應；然而這個一來一往的傳輸過程，資料都是 明文傳送。
- HTTPS - 加密過後的HTTP

超文字傳輸協定(HTTP)

- 使用者請求
  - GET: 向指定的資源發出「顯示」請求
  - HEAD: 與GET方法一樣，都是向伺服器發出指定資源的請求。只不過伺服器將不傳回資源的本文部份。
  - POST: 向指定資源提交資料，請求伺服器進行處理（例如提交表單或者上傳檔案）。
  - PUT: 向指定資源位置上傳其最新內容，若內容不存在則新增。

超文字傳輸協定(HTTP)

- 伺服器回應
  - 1XX: 訊息類 (收到請求，請求者繼續執行操作)
  - 2XX: 成功類 (操作被成功接受並處理)，例如：200 成功回應
  - 3XX: 重定向類 (需進一步操作才能完成)，例如：301 成功轉向
  - 4XX: 客戶端錯誤類 (請求語法錯誤或無法完成請求)，例如：404 找不到資源
  - 5XX: 伺服器錯誤類 (後端的問題)，例如：500 伺服器錯誤

HTML範例程式碼

```
<HTML>
    <HEAD>
    <TITLE>The title of the webpage</TITLE>
    </HEAD>
    <BODY>  <P>Body of the webpage
    </BODY>
</HTML>
```

HTML基礎架構

- 基本上 <head> 裡面的內容都不是給"人"看得，而是給機器運作、搜尋用得。
- 主要放置得標籤用來告訴搜尋引擎，這個網頁有什麼樣的內容、控制網頁與外部程式碼的連結、定義網頁使用的樣式等等。
- HTML5常用的標籤有 <title>、<meta>、<link>、<script>、<style>、<base> 等等
- 網頁真正會跑給使用者看的東西全部都在 <body>

網路爬蟲

- 網路爬蟲是一個透過程式「自動抓取」網站資料的過程，在這資訊爆炸的時代中，資料的收集是相當重要的工作項目之一，但如果透過人工的方式來收集網站資料，效率低之外也會花費掉非常多的時間
- 資料的收集與整理這份工作，可以透過網路爬蟲來協助，我們只要先制定好規則，網路爬蟲就可以自動依照這規則收集和擷取資料並整理出我們所需的格式
  - Excel、CSV等

網路爬蟲的應用

- 找飯店，Trivago!
- Skyscanner 機票搜尋
- 股票應用程式
- 美食推薦應用

網路爬蟲的原理

- 請求網頁內容
  - 網路爬蟲進行的第一步驟都是向目標網站請求特定網址（URL）的內容
- 抓取所需資料
  - 伺服器返回應網頁的 HTML文件後，在此步驟，網路爬蟲主要是將 HTML 文件做「解析」並「取出」所需的資料
- 儲存資料
  - 將取出的資料儲存在 CSV 檔案、Excel 表或是資料庫當中

網路爬蟲合法嗎?

- 透過網路爬蟲每天自動到別人的網站中抓取內容，這時你可能會開始思考一個問題，這樣可以嗎？
- 取決於如何抓取以及怎麼使用抓取到的資料
  - 遵守 robots.txt 的規範
  - 不造成網站伺服器的負擔
- 確認網站是否有提供 API，如有提供API可以直接使用API所定義的程式語法取得資料

Robots.txt

- 通常Robots.txt都在根目錄下，例如
  www.yahoo.com/Robots.txt
  www.google.com/Robots.txt

yahoo

```
User-agent: *
Disallow: /p/
Disallow: /r/
Disallow: /bin/
Disallow: /caas/
Disallow: /blank.html
Disallow: /includes/
Disallow: /_td_api
Disallow: /tdv2_fp
Disallow: /nel_ms
Disallow: /fp_ms
Disallow: /sports_fp_ms
Disallow: /search_ms
Disallow: /_tdpp_api
Disallow: /_remote
Disallow: /_multiremote
Disallow: /_tdhl_api
Disallow: /digest
Disallow: /fpjs
Disallow: /myjs
```

google

```
User-agent: *
Disallow: /search
Allow: /search/about
Allow: /search/static
Allow: /search/howsearchworks
Disallow: /sdch
Disallow: /groups
Disallow: /index.html?
Disallow: /?
Allow: /?hl=
Disallow: /?hl=*&
Allow: /?hl=*&gws_rd=ssl$
Disallow: /?hl=*&*&gws_rd=ssl
Allow: /?gws_rd=ssl$
Allow: /?pt1=true$
Disallow: /imgres
Disallow: /u/
Disallow: /preferences
Disallow: /setprefs
Disallow: /default
Disallow: /m?
Disallow: /m/
Allow:    /m/finance
Disallow: /wml?
Disallow: /wml/?
Disallow: /wml/search?
Disallow: /xhtml?
Disallow: /xhtml/?
Disallow: /xhtml/search?
```

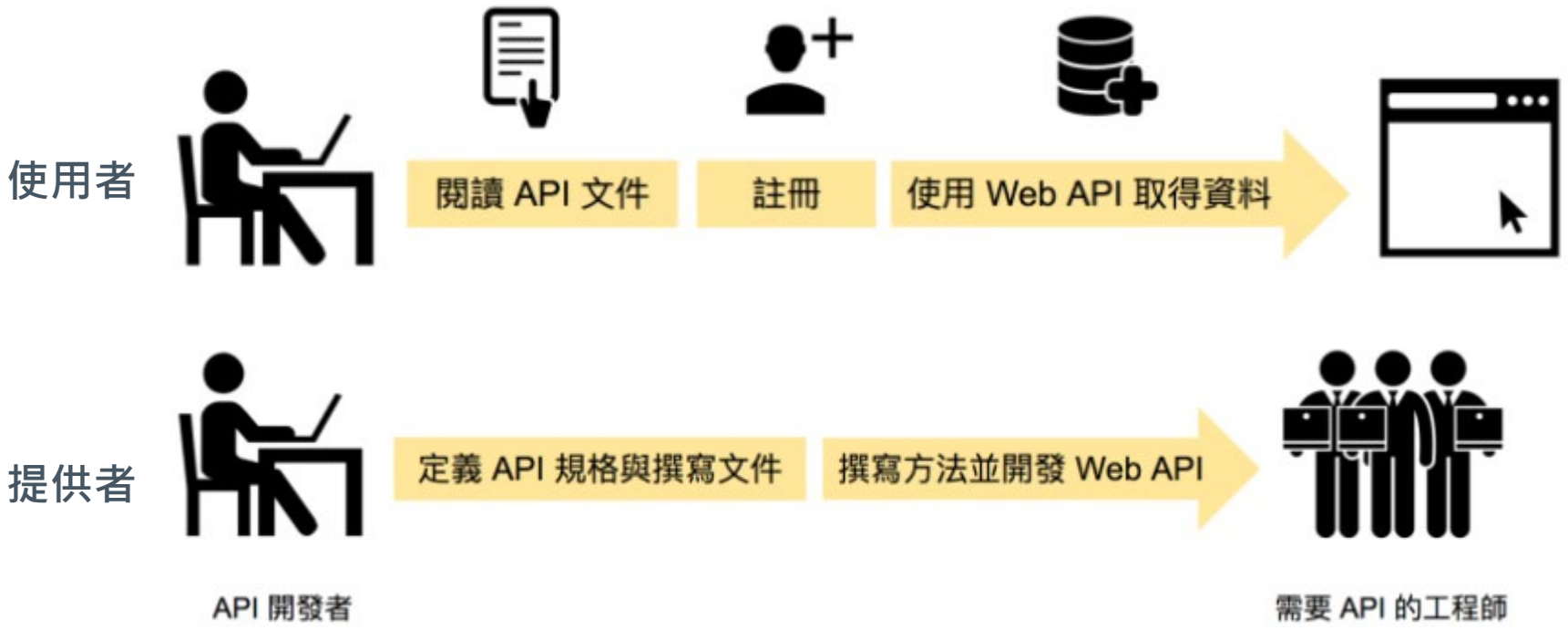
- 爬蟲本身不被法律禁止，可以採集對大眾、所有人公開的「公開資訊」，但用途須合理，如教學使用。
- 爬取非商業網站，像是國家政府資訊或公開資訊觀測站資料...等，這種對外公開且提供公開查詢服務的網站，一般不構成侵權，基本上可以抓取。
- 爬取商業網站，有些商業網站雖然沒有設定爬蟲哪些可以、哪些不可以爬取，但這種資料不代表可以隨意抓取，建議先取得對方授權同意才可執行。
- 使用爬蟲影響正常業務，比如搶購、搶門票、搶車票...等，會影響原網站使用者體驗的就不行。
- 雖然已經取得合法授權使用，但沒有遵循告知的使用目的進行使用，比如約定上只能分析使用，但卻用來販售資訊，這種也不可以。
- 用來爬取未公開、沒有經過許可、帶有敏感資訊(個資)的資料，無論如何都是非法的行為。

網路著作權合理的使用原則

- 利用的目的性與性質
  - 教育人員於網站上下載101的照片運用在教材中，讓學生認識台灣風景，但因照片只限於教學之用，無營利目的，就無侵權的疑慮。
- 著作之性質
  - 如憲法、法律、命令...等法律條文性質特殊，且內容有關於公共之利益，所以引用法條不會侵害著作權。
- 利用的「質和量」佔著作之比例
  - 若在網站上複製文章指更換標題，並發表在自己的網站上，就侵犯了著作權；但即使只是節錄文章中的一段文字，若該段文字為文章之精隨，也有可能侵犯著作權。
- 利用的結果是否會造成著作人的利益損害
  - 在網路上「合法下載」的歌曲只供自己欣賞，既轉載與販售，頁沒有造成著作權人的利益損害，就不構成侵害。

應用程式介面
Application Programming Interface, API

› 應用程式介面 (API) 是用於打造應用程式軟體的一組副程式定義、協定與工具。一般而言，API 是指各種軟體組件之間一套明確定義的溝通方法

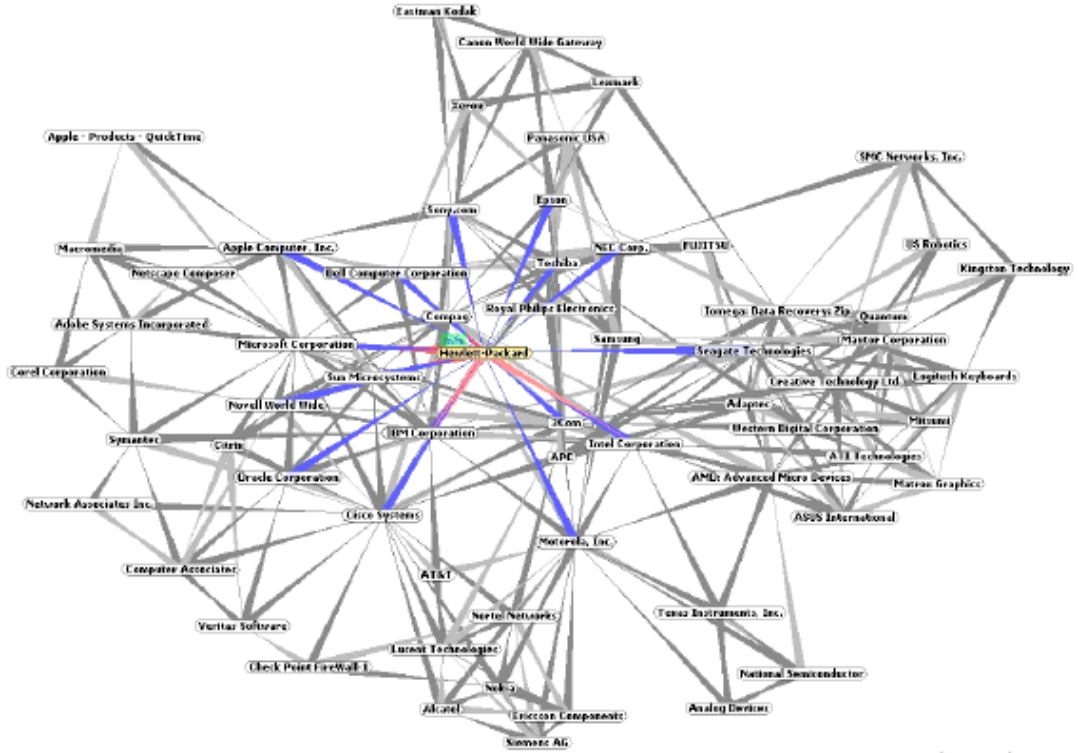


網路文本探勘 (Web Text Mining)

搜尋引擎

- 全文檢索
  - 將全部的文字訊息儲存起來
  - 使用者必須詳細的規劃自己的查詢
- 關鍵字查詢
  - 字詞切割
  - 關鍵字定義與比對
- 自然語言處理

網路圖探勘(Web Graph Mining)



網路圖探勘(Web Graph Mining)

- 網路圖分析
- 網路連結分析
- 網頁重要程度分析
- 異常使用者偵測