如何在網路上減低自己被人肉搜索的風險?

動機 與 目的

網路社交平台已成為我們記錄生活、與人交流的重要媒介，在台灣使用度最高的就屬PTT與FB；其中 PTT是台灣本土第一大以電子布告欄系統形態架設的資訊與社交平台，平均每日上線使用者超過40萬以上。而FB則是全球最大以網頁形態呈現的社交網站，全球超過上億使用者。

在這些社交平台上，人們每天留下了各種的資訊「足跡」，也帶來被「人肉搜索」的風險，網路上一些所謂的“神人” 甚至只需要一個人的學校、IP等資訊，即可以把一個人的PTT ID與FB帳號做連結，進而找出個人的感情狀況、親友、照片、做過的事情等等。

因此，我們想要探究：

- 哪些資訊是敏感而容易洩漏個人資訊?
- 社群網路使用者分享資訊的習慣、心理與被人肉搜索的風險?
- 如何在網路上減低自己被人肉搜索的風險?

研究方法

STEP 1 利用人工方式蒐集資料

STEP 2 資料前處理 & 相似度計算

STEP 3 利用分析工具分析 & 實驗設置

STEP 4 觀察結果並討論

資料蒐集：

相似度計算：

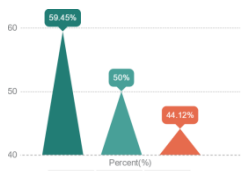
#1 LCS：姓名
Longest Common Substring
「最長共同子字串」。
出現於每一個字串的子字串，長度最長者。

#2 Binary：性別、血型、公司、職業、婚姻、網址
即是以1代表相同，0代表不同。

#3 Jaccard：出沒地、學校、興趣
只考慮有出現的項目，並計算對應率。number of matching presents / number of attributes with values present

#4 自我定義部分：生日
年月日拆成三項，依照對應的部份給定相似度。
e.g. 日+月 <-> 日+月：0.8
日+月+年 <-> 日+月+年：1
其他：0

資料分群：



依照資料的提供率分為高中低三群。

分析工具 與 演算法：

AD Tree
全名為：alternating decision tree，
是一個用來分類(classification)的機器學習方法。

可以 generalizes decision trees，
並且藉由整合弱的classifier，給予強的classifier較高的權重，以達到更好的training結果。
跟CART以及C4.5是不同的類別。

實驗設置：

◆ **Label 0 與 label 1 的意義：**
Label 1：不同平台(PTT及FB)上的同一個人

Label 0：不同平台(PTT及FB)上的不同人

利用程式random出不同的資料進行訓練
e.g. PTT * 10 + FB * 10 => label1 * 10 + label0 * 100
=> random出10個

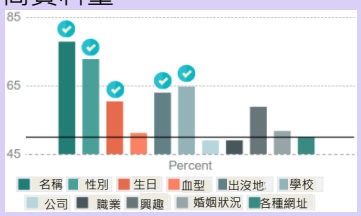
◆ **N-fold cross-validation：**
在資料總數少的時候非常適合。將資料切成n等分，交互進行training跟testing並將結果取平均。
本次專題取 N = 10。

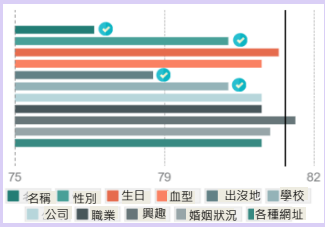
實驗結果與討論

＋ 一次只考慮一個屬性的相似度，並比較各個屬性的準確率，找出何者為高。

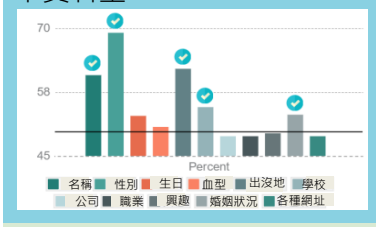
－ 以全資料屬性的實驗結果為基準，每次移除一個屬性的相似度，比較各個屬性被拿掉後，準確率的下降多寡。

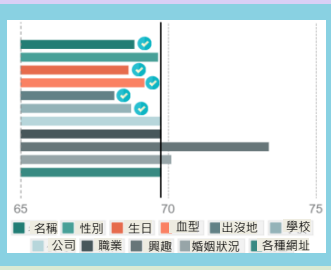
高資料量



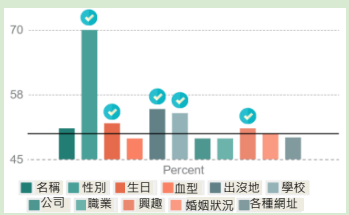


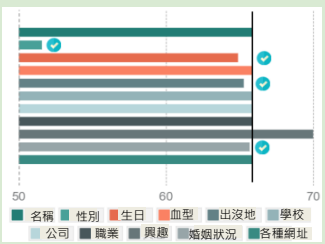
中資料量





低資料量





結論

重要屬性：
性別 生日 出沒地

1.PTT和FB兩邊填不同的資料或交錯填寫。

2. 避免同時填性別、生日和出沒地。

PRJ-NTPUCSIE-104-05

資工三 410285021 王瑞翊 410285029 王靜瑩