書目計量學

Lecture 06 -- 齊普夫經驗法則

陳光華
國立台灣大學圖書資訊學系
國立台灣師範大學社會教育學系
khchen@ntu.edu.tw

大綱

- 齊普夫與最省力法則
- 齊普夫經驗法則的形成
- 齊普夫經驗法則的基本原理
- 齊普夫經驗法則的發展
- 齊普夫經驗法則的應用

齊普夫

- 美國哈佛大學教授、著名的語言學家和心理學家
- 用大量的統計數據來驗證前人有關詞頻分布規律的研究成果
- 1948年出版《人類行為與最省力法則–人類生態學引論》。

齊普夫之最省力法則

- 例子
  - 從A地到B地，可以走各種不同的道路；從經濟上、安全上、時間上並結合本人的主觀條件（如身體情況）及客觀情況（所處的環境）等種種因素考慮，設法選擇一條最符合自己要求的道路，使得自己付出的“力”最小。
- 利用語言表達思想時，會受到兩個方向相反的力之作用
  - 單一化的力：希望儘量簡短
  - 多樣化的力：希望被對方理解
    - 單一化的力與多樣化的力取得平衡，使自然語言詞彙的分布呈雙曲線。

齊普夫經驗法則的形成

- 齊氏經驗法則形成的基礎
- 齊氏經驗法則的確立

齊氏經驗法則形成的基礎

- 頻率詞典（詞表）
  - 每一個詞在一定長度之文件中出現的頻率
  - 兩個最基本的數量指標
    - 詞的出現頻率、詞的序號
- 艾思杜的發現（1916）
  - 較長文章中，詞頻分布之定量化形式
  - $n_r \cdot r = C$（常數）
    - 詞的序號：*1, 2, …r, …D*
      （*1*：絕對頻率最大的詞，*D*：絕對頻率最小的詞）
    - 詞的絕對頻率：$n_1, n_2, \ldots, n_r, \ldots n_D$

E. Condon的公式

- 詞頻分布圖
  - 詞之序號的對數做為橫座標log r
  - 詞之絕對頻率的對數做為縱座標log n_r
- 定量公式：$f_r \times r = c$

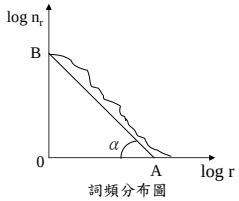
  令 $\tan \alpha = \gamma$

  $\log(r^{\gamma} \cdot n_r) = \log C$

  $n_r = \frac{C}{r^{\gamma}} \xrightarrow{\gamma = \tan \alpha = \tan 45^{\circ} = 1} n_r = Cr^{-1}$

  $\frac{n_r}{T} = \frac{C}{T} r^{-1}$（$\frac{n_r}{T} = f_r$，$\frac{C}{T} = c$）

  $\rightarrow f_r \times r = c$



詞頻分布圖

- *c*的值究竟是不是常數，還必須再加以驗證

齊氏經驗法則的確立

- 檢驗 *E.Condon* 關係式的可靠性和研究C的性質
  - 確定*c*是一個參數，使得 $\sum_{r=1}^{n} P_r = 1$
- 驗證了單參數詞頻分布公式之正確性
  - $f_r \times r = c$（或 $P_r \times r = c$）

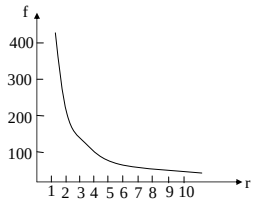
齊普夫經驗法則的基本原理

- 齊氏經驗法則的基本內容
- 齊氏經驗法則的圖像描述
- 齊氏經驗法則的侷限性

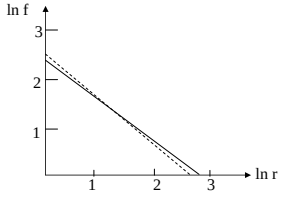
齊氏經驗法則的基本內容

- 齊普夫經驗法則
  - $f_r \times r = c$
    - f_r：頻次，r：等級序號
- 齊普夫經驗法則之“最省力法則”的解釋
  - 任何語言中，凡是使用頻率最高的詞，功能總是不會太大；因為其本身在這個場合中價值小，因而傳遞他們所需要的“力”就不大。

齊氏經驗法則的圖像描述



齊普夫分布曲線



齊普夫分布對數曲線

- 橫座標：等級序號r
- 縱座標：相應的頻率f
- 等級 r 與頻率 f 皆取對數
- 虛線：$\ln r+\ln f=\ln c$
  實線：$b\ln r+\ln f=\ln c$
  （斜率為b）

齊氏經驗法則的侷限性

- 對出現頻率特別高的詞和特別低的詞，並不能完地反映分布規律
  - 低頻率的詞，序號相同的很多。
  - 高頻率的詞，序號相同的詞隨著頻率的增高而越來越少。

齊普夫經驗法則的發展

- 朱斯（M. Joos）修正式
- 孟德爾伯特（ B. Mandelbrot ）修正式
- 布斯（B. Booth）之齊普夫第二經驗法則

朱斯修正式

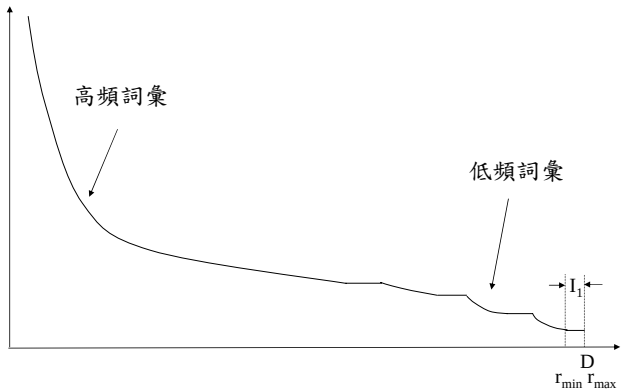
- 單參數詞頻分配律→雙參數詞頻分布律
  - $p_r = cr^{-\gamma}$ 中，c和r的負指數（以 γ 表示）都是參數。
  - γ 不是一個常數而是一個參數
    - 當詞典收詞多時， γ 會增大，即圖像中的 α 角會增大；當詞典收詞小時， γ 會減少，即圖像中的 α 角會變小。
- 雙參數詞頻分布律之公式
  - $p_r = cr^{-\gamma}$
  - $\gamma > 0, c > 0$，對於 $r = 1, \ldots\ldots, D$ ，參數 γ , c要使
  - 當 $\gamma = 1$時，公式變為$p_r = cr^{-1}$，即是齊普夫的單參數詞頻分布律

$$\sum_{r=1}^{n} p_r = 1$$

孟德爾伯特修正式

- 三參數頻率分布律 $\sum_{r=1}^{n} p_r = 1$
  - $p_r = c(r+a)^{-b}$
    - $0 \leqq a < 1, b > 0, c > 0$，對於 $r = 1, \ldots\ldots, D$，參數a、b、c，要使
    - 參數c：與出現概率最高的詞之概率大小有關。
    - 參數b：與高概率詞的數量之多少有關，對於$r<50$的高頻率詞，b是r的非減函數，隨著r的增大，參數b並不減少。
    - 參數a：與詞的數量n有關。
  - 當$a=0$，公式形式為$p_r = cr^{-b}$（朱斯修正式）
  - 當$a=0, b=1$時，公式形式為$p_r = cr^{-1}$（齊普夫公式）

齊普夫詞彙分佈圖



齊普夫第二經驗法則

- 齊普夫經驗法則可以分成兩大經驗法則
  - 高頻詞分布的經驗法則（第一經驗法則）
  - 低頻詞分布的經驗法則（第二經驗法則）
    - 齊普夫之推導
    - 布斯之修正

齊普夫之推導

$p_r = n/T$

$r = c/p_r \Rightarrow n = \frac{cT}{r}$

- p_r為第r位詞出現的機率
- T為詞的總體集合中不同詞出現的總次數
- n為序位為r的詞彙的絕對頻率

- 僅出現一次的詞彙，可能有許多個，一般而言，

$$0.5 \le \frac{cT}{r} < 1.5 \Rightarrow 2/3 < \frac{r}{cT} \le 2$$

$$r_{\min} = \frac{2cT}{3} < r \le 2cT = r_{\max} = D$$

$$I_1 = r_{\max} - r_{\min} = 2cT - \frac{2cT}{3} = \frac{4cT}{3}$$

$$I_1 / D = (\frac{4cT}{3}) / 2cT = 2/3 = 0.667$$

齊普夫之推導（續）

- 出現n次的詞彙，也可能有許多個，一般而言，

$$n - 1/2 \le \frac{cT}{r} < n + 1/2 \Rightarrow \frac{1}{n+\frac{1}{2}} < \frac{r}{cT} \le \frac{1}{n-\frac{1}{2}}$$

$$r_{\min} = \frac{2cT}{2n+1} < r \le \frac{2cT}{2n-1} = r_{\max}$$

$$I_n = r_{\max} - r_{\min} = \frac{2cT}{2n-1} - \frac{2cT}{2n+1} = \frac{4cT}{4n^2-1}$$

$$I_n / I_1 = (\frac{4cT}{4n^2-1}) / (\frac{4cT}{3}) = \frac{3}{4n^2-1}$$

布斯之修正式

$p_r = n/T$

$p_r = cr^{-1}$，$c = rp_r$

$r_n = c/p_r = c\,p_r^{-1} = \frac{cT}{n}$

- p_r為第r位詞出現的機率
- T為詞的總體集合中不同詞出現的總次數
- 第r_n位詞出現的次數為n

- 出現n次的詞之數量

$$I_n = r_n - r_{n+1} = \frac{cT}{n} - \frac{cT}{n+1} = \frac{cT}{n(n+1)}$$

$$I_1 = \frac{cT}{1(1+1)} = \frac{cT}{2} \text{-----} a\text{式}$$

$D = cT/1$（由前述$r_n = \frac{cT}{n}$）

代入a式得

$I_n/D = [n(n+1)]^{-1}$

故 $I_n / I_1 = \frac{cT/[n(n+1)]}{cT/2} = \frac{2}{n(n+1)}$ ----齊普夫第二式

假設有D個不同詞彙，則最高序位數為D，其至少出現一次

齊普夫經驗法則的應用

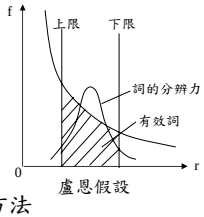
- 文獻索引和詞表編制
- 資訊檢索
- 在圖書資訊管理的應用

詞表編制

- 敘詞表和標引可提高計算機檢索的效率
- 根據齊普夫的頻率分布方法，通過標引試驗，找出被引文獻與敘詞使用頻率的分布特徵，確定合乎需要的參數值
- 選用原始文獻中的術語，統計其發生的頻率，研究分布特徵，最後決定合乎使用頻率的詞

文獻索引

- 自動索引
  - 利用電腦對每個詞的頻率進行統計分析，篩選出適於索引的詞彙
  - 盧恩（Luhn）之頻率自動索引方法
    - 排除高頻詞
    - 去後綴
    - 找出相應的詞幹
- 索引加權
  - 史派克瓊斯（Spack Jones）的加權方法
  - 如果有N篇文章，某個檢索詞涉及其中的n篇，那麼給這個詞log(N/n)+1的權值，可得到較佳的檢索效果。



盧恩假設

資訊檢索

- 用於估計資訊系統所需的儲存量
- 倒置檔的大小，取決於同屬性欄位中不同詞之數量以及每個詞的出現頻率
- 按照齊普夫經驗法則，計算詞頻出現的機率
  - 假定某一種字段共有D個不同詞彙，總數為T
  - p_r：隨意從有關欄位抽取描述r等級詞屬性值的機率

$$p_r = \frac{r\text{等級詞出現頻次}}{T}$$

  - 研究發現詞頻機率可以滿足下式

$p_r = A/r$（A為常數，A值近似於0.1）

在圖書資訊管理的應用

- 處理與語言文字有關的問題
- 將“最省力法則”的原理應用於圖書資訊事業之管理
  - 合理地選擇圖書館或資訊中心的最佳地理位置，使得使用者都能以最省力的途徑方便到達
  - 用以設計圖書館、資訊中心資料庫的排架

結語與討論