剎那古籍處理技術盤點

古籍協會2022.11

技術總覽

文字處理技術：✅ 漢字拼形--解決所有漢字的自由表達

文件處理技術：✅ 文層與文釘--讓使用者與底文脫鉤的關鍵

文庫處理技術：

- ✅ 離線跨平台全文搜尋
- ☐ 互文式文庫--異質性資料庫的自由組合與呈現。

文字

缺字(缺碼)是所有古籍處理無法迴避的,

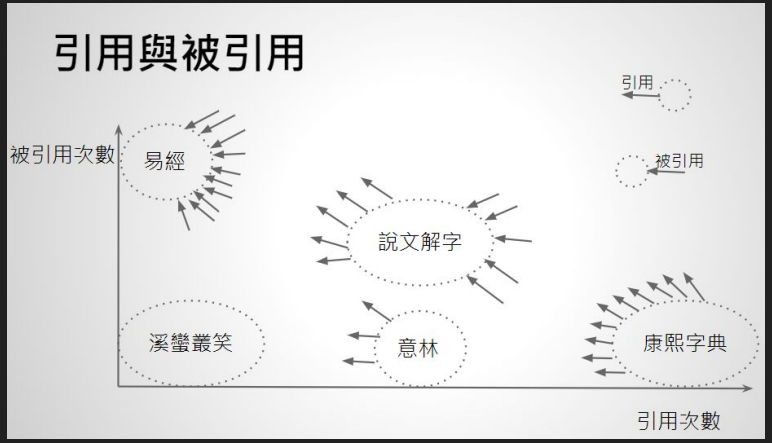
它不因硬體和網路技術的發展而自動解決。

目前已基本解決, 歷史過程詳見漢字拼形。

大規模應用需要軟體廠商之配合植入操作系統。

demo

文件(引用視角)



順連：某一段文字引用另一段文字。

逆連：從任意文字，找到引用它的文字。

core text (root) 核心文件 ：常被引用但不引用其他文件。(常用經文)

hub text (trunk) 樞紐文件 ：引用次數及被引用次數之乘積較大者。

thesis (fruit) 論文 ：有引用其他文本之文件。

reference work (jam) 工具書：引用涵蓋面廣泛者。

ps. 傳統 HTML 的連接<a>是從「文字」到「文件」

文件(文獻學視角)

Ferdinand de Saussure：文本語意由「語序軸」與「聯想軸」組成

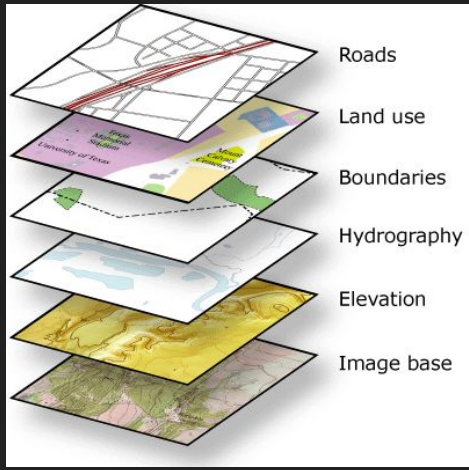
底文：原始創作的文字。Syntagmatic axis (無著作權)

加值：媒材訊息、校注、內容結構、語意標籤。Associative axis (可以有著作權)

文層：使底文與加值脫勾。地圖喻、小畫家與 Photoshop。

文釘：定位一小段底文（首字....尾字）。

補充說明：文件版本的繁衍與維護問題



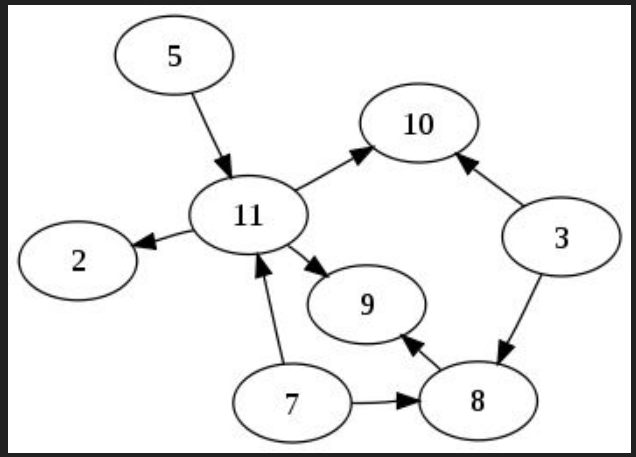
文庫

由各種類型的文件以「有向無環圖」組織成互文式文庫。

文本階級：字詞、片語（辭典最小單元）、句(平行語料最小單元)、段落（定址命名單元）、篇章節、作品（書）、叢書。

有別於傳統的全文式或表格式資料庫。

古籍文庫是一個Graph結構的「有機體」。



文件入庫之順序

將文件切分語義完整的小段落，採用 PageRank 計算權重。

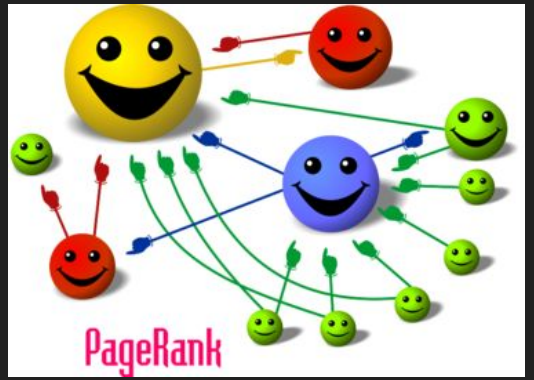
按權重決定加工順序，達到事半功倍之效。

連結分布很不平均、密集在極少數的核心文本與樞紐文本。

漢字也有類似特徵、前100個常用字，貢獻了70%以上的字頻。

康熙字典80%的引用集中到25本古籍。

(十三經、前四史、老莊)



古籍區塊鏈

Block Chain：凡創造出來的永不消滅。是去中心化數字貨幣的基礎。

古籍最重要的就是可信度，不能在流通過程中被任何修改，但又要保有演化和散佈的能力。在資源不足的情況下，這兩者有天然矛盾。

古人花大力氣刻在石頭上（第五次巴利三藏結集、房山石經、碑林），滿足「可信度」和稍有失真(拓版)的「散佈」。

就像是數字空間上無限的石頭。Block Chain 同時滿足了「可信度」「散佈」以及「演化」。

Web 3.0

web 1.0 廣播式（發佈者消費者涇渭分明，讀）	News
web 2.0 互動式（消費者參與信息的創造，讀+寫）	Social Media
web 3.0 去中心（信息主權回歸創造主體，讀+寫+主權）	Metaverse??

未來你發明了一「金句」，知道被誰引用，就可以建立報酬回流機制。

如果有人精校一本經典，大家樂於引用，可以從web 3.0的架構自動得到收益，以繼續改善。