應用本體論設計與建置摘要系統
An Ontology-Based Article Summarization System

吳家威　　劉昭麟
國立政治大學資訊科學研究所
{g9010,chaolin}@cs.nccu.edu.tw

Abstract[†]

本文使用兩種方法實作英文文章的自動摘要系統，第一種以 Term-frequency 為主的自動摘要的系統，第二種使用本體論(Ontology)建置的自動摘要的系統。Ontology 的目的在於對知識的分享與利用；我們是否能透過 Ontology 進一步了解文章所描述的內容進而增加摘要的品質是本篇論文的主旨。我們實作一個 Ontology 的資料庫並嘗試運用 Ontology 對文章的主題(Subject)進行分析，在獲得文章中的數個主題後，區分每個主題的重要性，並依此資訊將文章的段落做主題相關度的評估計算；依照計算的結果選出一定比例的段落作為摘要內容。測試的內容為線上資料庫所提供的英文報紙以及該資料庫提供的由專家所選出的摘要。本文最後列出兩種方法的實驗結果。並針對兩種方法做比較與分析。

關鍵字：document summarization、ontology、document extracts、web information extraction

1. Introduction

近年來，文件資料隨著網際網路的發展而大幅增加，因此我們需要一種能夠快速且有效的方法幫助我們獲取所需資訊。一、是能快速且正確的搜尋到資料，二、是縮減文件的內容使閱讀的時間降低，增加閱讀的效率。於是摘要的建置變得十分的重要，除了能增加閱讀的效率之外也能減少搜尋的時間。但是摘要的建置必須花費相當高的成本，且摘要的產生包含了許多的主觀在內，必須要有專業人士才能撰寫出符合文章主題的摘要。若能成功的建置一套可靠的摘要系統便可快速、經濟的完成摘要，提供使用者精簡的內容閱讀。但分析、了解文章的內容選出重要的部分是非常困難的。即便是由人類來挑選摘要，不同的人由於主觀的差異挑選的內容也不盡相同。過去有許多針對此議題的研究，提出的方法主要可分為：語言相關的方法(Linguistic)、統計相關的方法以及兩者的結合 [5]。

由於自動摘要的問題十分困難於是大部分的研究大多將方向專注在從文件中圈選出摘要。將我們一般認識的摘要：從了解文章內容，詮釋、重新編寫摘要(Abstract)、最後產生輸出，簡化成為一個較簡單的問題：將文章的句子根據每個句子的特點加以排序，並據此抽取出摘要。於是，問題便成為如何分出每個句子的重要性。本論文也將研究問題簡化為相同的模式。

我們參考過去的文獻定義了一種以 Term-frequency 為主的摘要方法。同時我們也提出了一套 Ontology-based 的分析方法。並運用兩種方法分別實作自動摘要系統。資料的來源為線上資料庫 ProQuest 所提供的英文報紙的新聞資料，我們採用的新聞資料包括 New York Times 以及 Wall Street Journal。ProQuest 所提供的新聞資料會提供由專家所挑選的摘要，可以作為評估的標準。評估的方法是將系統選出的摘要與專家所挑選的段落進行比對，計算 Precision 和 Recall 的值，最後分別計算兩種方法的 F measure 的值作為比較的參考。

本論文的介紹次序如下：探討與本論文的相關研究，介紹 ontology 的定義與用途。對我們所使用的兩種方法做詳細的說明。陳述實驗進行的程序步驟。最後是實驗數據的展示並針對兩種方法加以探討。

[†] 發表於 2002 民生電子研討會論文集第 41 到 46 頁，台灣新竹，Dec. 2002。

2. Related Work

自動摘要的研究主要目的在於幫助使用者快速有效的了解文章的內容並且在損失最少資訊的條件下完成。而摘要的研究依產生的內容主要分為兩種：第一種方法稱為 abstract，將文件的內容分析後重新編寫，產生一份與原本文章的句子不完全相同的摘要。第二種方法稱為 extract，將文章的句子抽取一定比例作為摘要。由於產生一份重新編寫的摘要非常的困難，所以目前主要的方法大多以文件句子的選取當作摘要為主。本文所使用的方法為第二種方法，將文章的段落選取一定的比例作為摘要。

過去的研究已提出許多有關自動摘要的方法與討論如：Kupiec 等人於 1995 年提出的方法：利用字數統計，段落位置等特徵作為選取摘要的依據[11]。摘要的長度則約採用文章內容的 20%。Meown 等人則提出自動產生 abstracts，他們實作一套自然語言處理的系統(Natural Language Processing)，目的在分析新聞的內容並產生摘要，他們先建構許多的樣版(Template)把新聞內容與樣板對應，最後按照樣版的型式輸出[3]。Allan 等人則提出以新聞事件(Event)為主要核心以此追蹤新聞文章並產生摘要，而摘要包含了事件的前後資訊也加入了時間的概念 [1]。

近來的研究多半加入使用者的查詢，也就是摘要的內容與使用者所下的查詢必須配合，產生符合使用者需求的內容，使用者能藉此減少查詢後尋找資料的時間。

3. Ontology

近年來 ontology 又在資訊科學的相關領域中引起廣泛的討論[3]。然而，Ontology 的定義、建構與表達方式並沒有一定的準則或標準。我們採取 Gruber 的定義：Ontology 是一種對某一個概念的詳細描述，包括對於概念、關聯、實體的描述。並清楚的定義其所欲表達的概念，主要的目的可用於知識的分享與再利用 [6]。

基本上，Ontology 將世界分解為數個物件並加以描述，而描述的方法和呈現的方式都需視應用在什麼樣的系統而決定[4]。而本文所建構的 ontology 目的在幫助自動摘要系統分析文章的語意，將文章拆成數個主題並利用此資訊幫助系統取得摘要。因此我們將各個概念拆成數個 ontology subtree ,以樹狀的方式呈現。不同的 ontology subtree 所呈現的概念(Concept)即代表著各種不同的文章主題。

4. Data Source

文章的品質也是影響對自動摘要系統非常重要的因素，若文章的品質不一，例如：作者未使用固定的詞彙或者一致的描述風格，將使自動摘要問題變的更加困難。

實驗所需的新聞資料取自 ProQuest 線上資料庫所提供。採用 New York Times 以及 Wall Street Journal 的新聞內容。我們使用該資料庫原本具有的查詢語言搜尋以新力公司為關鍵字的文章，使用關鍵字 SONY 對新聞資料作查詢。總共所收集新聞 51 篇，共有 882 個段落，ProQuest 所選出的摘要共有 133 個段落(在此我們視為正確的摘要)。由於 ProQues 所提供的資料已經將文章切割段落。而段落可能不只一個句子。因此，本論文圈選的摘要皆以段落為基本的單位。

5. Methods

基本的構想是將文章的每一個段落標上分數，以分數高低作為排序的依據。其中分數越高的段落表示該段落可能在該篇文章的重要性越高，越有可能列入文章摘要中。之後，依照分數的高低來決定排序，並依序選出固定比例的段落作為系統所選出的摘要。

為了比較 ontology 的效用，我們的實作了包含 ontology-based 的兩種自動方法。

5.1 方法一：不用 ontology

第一種方法主要依照文章的非結構(Non-structure)的部分或者是過去的研究中較常使用到的規則(Rule)。而方法是來自參考過去的文獻所綜合的[4,5,11]。

- Term-frequency：首先統計文章中每個單字出現的次數，依出現次數的高低選

出最常出現的 N 個詞彙，作為計分的標的。而後再按照所選出來的詞彙對每個段落做投票計分。N 則是經過實驗挑選出來，我們設定 N=5。

- 段落所含字數的特徵：一般而言，若某一個段落所含的字數低於某個標準，則該段落被列入摘要的可能性很低。因此，我們設了一個門檻(Threshold)：一個段落必須包含 10 個詞彙以上。若低於所設的標準則不列入考慮。在計分的式子最前面所乘上的布林值目的就在過濾所含字數太短的段落上。
- 加分字：某些字出現在摘要的機會較高，因此我們認為若該字出現於某一段落中，該段落出現在摘要的機會也應該較高。我們先使用了 15 篇的摘要統計詞彙出現的次數作為加分的詞彙。
- 專有名詞的特徵：在段落中專有名詞出現的次數與段落選入摘要的機會有關。通常專有名詞出現越多該段落也越重要。我們使用一個段落中大寫字母出現的次數作為計算該特徵的依據。

計分的公式如下：

$$G_i = w_1 f_1 + w_2 f_2 + \cdots + w_n f_n$$

G_i 為第 i 段落的分數，也是最後選擇摘要時排序的依據。 f_i 為第 i 個特徵所得的值。而 w_i 為第 i 個特徵值之權數。

5.2 方法二：Ontology-based

利用 ontology 分析整篇文章的主旨，在獲取語意層次的資訊後，幫助系統選擇挑選段落再匯集成摘要。首先，必須定義與建構 Ontology 供系統使用。

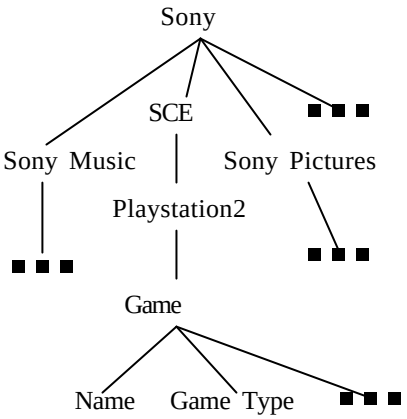
Ontology 的建構尚且沒有一個標準可依循。在參考相關文件後[10,12,13]，我們使用的方法如下：

一．定義 Ontology 的使用目的：本文所建立的 ontology 目的在幫助自動摘要系統分析文件內容。

二．定義 Ontology 的領域：本實驗所定義的 Ontology 主要的內容為 SONY 新力公司的產品資料，以及財務相關的資料，作為建構的主要領域[4]。

三．建構 Ontology：依據定義好的範圍收集相關的詞彙，並同時建立其同義詞的資料。將收集的詞彙依照資料模型(Data Model)，實體(Entity)，屬性(Attribute)，關聯(Relation)設計 ontology。例：

圖一、 我們的 Ontology 的一部份



通常一篇文章都包含一至數個主旨，而主旨是影響摘要的選取的重要考慮因素。我們希望藉由 Ontology 取得文章的主旨進而獲取摘要所需的資訊。

由於我們所選的資料主要來自報紙中有關新力公司的新聞，因此我們曾經考慮使用報紙的標題作為分析文章主旨的依據。但由於新聞的標題呈現議題的方式非常多元，能使用非常多種的描述發法來呈現主題，例如以下的新聞標題：*"Sony Looks Golden by Comparison"*。我們可以猜到新聞的內容大概是描述新力公司的營運表現相較其他公司表現的很好。但是，對於本文內容的分析卻幫助有限。因此，欲分析新聞內容表現的主題並非很直覺的單參照標題即可。在系統中我們將標題當作參考的特徵(Feature)之一，而非單只參照標題。以下是我們提出的方法：

首先，將文章內容的每個詞彙依序對應到 Ontology 中所定義好的詞彙中。若在 Ontology 中未定義該詞則予以忽略。若有相對應的詞則紀錄對

應到的次數。Ontology 使用的是樹狀的資料結構，每個節點所代表的概念都包含了其子節點所代表的概念，例如：電影“ Spider-man”單字出現，我們依據 ontology 認為“ Spider-man”的概念是其父節點“ 電影”概念的一部份。因此，在統計的過程中，子節點的次數會向上累加給其父節點，因為代表著相同的概念。而最上層的概念代表著文章候選的主題。最後統計的結果作為系統判斷文章主題的標準。前面有提過我們假設文章的主題可能不只一種，因此我們在選取主題時也不僅僅只取單一的主題。

在我們取得文章主題的資訊後，接下來便利用這些資訊來幫助系統選取摘要，也就是說，如何利用這些資訊區分出每個段落的重要性。根本的假設為：越接近主題的段落表示越重要且越能代表文章所要表達的意含。我們採用的方法如下：

一、將主題(在此即是 Ontology subtree)所定義的詞彙與每個段落的詞彙作對應的工作，並記錄出現的次數，作為主題的分數。

二、 一個段落可能也含有不同的主題，因此，必須將每個主題所計算的結果作加總並乘上權數。權數決定的基本假設為：一篇文章可能有數個主題，但主題之間也有主要與次要的差異，而主要的主題相對次要的主題而言應該更具文章的代表意義，因此被放入摘要的機會也越大。在計算出每個主題在段落所得之分數後必須乘上權數，乘上權數而的目的即在區分主題與主題之間的重要性高低。原則是越重要的主題越能代表該文章所乘上的權數也必須越高。

我們使用之前在計算每個 ontology subtree 所得到的分數作為權數。分數越高表示該 ontology subtree 所定義的詞彙在文章中出現的次數越多。最後將每個加權後的主題加總即為該段落所得之分數。以下我們舉一個簡單例子：

下段文章段落選自 *Wall Street Journal* *July* 26, 2002 標題為 *“Movie Helps Sony Post Profit”*。

Sony Chief ***Financial*** *Officer Teruhisa Tokunaka said box-office receipts of the film "****Spider-Man****" have reached $675 million, making it the fifth-largest-****grossing movie*** *ever (unadjusted for inflation) and boosting sales at Sony's* ***movie*** *business to 173.6 billion yen, a 28% increase from a year earlier. Mr. Tokunaka said operating* ***profit*** *at Sony's* ***electronics business****, which accounts for 70% of the company's annual sales, rose to 49 billion yen from 1.5 billion yen a year earlier.*

假設我們分析文章後得到的主題包括 movie、electronic business、financial 三個主題，分數分別為 20、10、15。則計算該段分數時遇到與 movie 有關的詞彙如：movie，Spider-Man 需要加上 20，而遇到 electronic business 相關的則加上 10。20 與 10 即 movie 與 electronic business 的權數。所有主題的分數作加總即此段的分數。

關於取幾個主題作為計分的依據我們的做法是權數小於 10 的主題直接視為雜訊予以刪除。而主題的選取數量最大值為 3，我們相信這個假設應是合理的。因為，一篇文章能介紹的主題應是有限的，而摘要只能包含最重要的幾個主題；因此 ,即便文章包含許多主題也應該刪除較不重要的主題。

三、 依每段所得到的分數高低加以排序，並取一定比例的段落作為系統最後選出的摘要。

計算公式如下：

$$P_i = w_1 o_1 + w_2 o_2 + \cdots + w_n o_n$$

P_i 為第 i 段落所得的分數，o_i 為第 i 個主題(即第 i 個 ontology tree)所代表的的分數，w_i 為第 i 個權數。

6. Experiments

首先，將文件輸入至系統中，若系統選出的摘要段落與我們事先取得正確的摘要比對一致則算正確。

評估的方式：依照分數的高低予以排序，依序

選出一至十個段落作為摘要，例如：系統選出的摘要段落為 3 段，則依照得分的高低依序選出 3 個段落，總共 51 篇文章每篇文章 3 段，共選出 153 個段落。最後，依照各個選取段落數目計算其精準度(precision)與召回率(recall)。

未了反應當文章押縮超過一定程度後無法環傳足夠的段落來評估，我們採用調整過的召回率及 F-value[5]。

$$\text{Precision}=\frac{J}{K}$$

$$\text{Recall}=\frac{J}{\min(M,K)}$$

$$\text{F}=\frac{2}{\frac{1}{precision}+\frac{1}{recall}}$$

J 為選取的段落且同時為正確的摘要的數目，K 為共選取幾段作為摘要，M 為正確的摘要共有幾個段落。J、M、K 皆為所有文章實驗結果的加總。例如：每篇選 3 段作為摘要，共有 51 篇，則 K=153 個段落。

7. Results

以下列出實驗結果的數據。圖二為使用 Ontology-based 方法所建置的摘要系統的實驗結果。圖三為使用隨機選取段落作為摘要選取方式的實驗結果，主要目的在於評估摘要正確率的最低標準 (base-line)。將隨機選取摘要的方法與 Ontology-based 的方法比較可以發現：系統採用 Ontology-based 的方法對摘要的正確率有明顯的提昇，由此可以證明 Ontology-based 的方法對文章摘要有顯著的成果。(以下三個表格的第一列數列 10~1 表示選擇多少個段落作為摘要)

圖二. Method 2 (ontology-based)

	10	9	8	7	6	5	4	3	2	1
Precision	0.24	0.27	0.29	0.32	0.36	0.42	0.46	0.52	0.54	0.70
Recall	0.94	0.93	0.91	0.87	0.84	0.81	0.70	0.60	0.54	0.70

10~1 為選取的段落數

圖三、Randomly Selected Paragraph

	10	9	8	7	6	5	4	3	2	1
Precision	0.13	0.14	0.14	0.13	0.14	0.16	0.18	0.17	0.16	0.19
Recall	0.22	0.21	0.21	0.19	0.20	0.21	0.21	0.18	0.16	0.19

圖四為方法一(term-frequency 為主)與方法二(ontology-based)F-measure 的比較。

圖四、4. F-measure 比較

	10	9	8	7	6	5	4	3	2	1
Method 2 Ontology	0.39	0.41	0.45	0.47	0.51	0.56	0.55	0.55	0.54	0.70
Method 1 Non-ontology	0.38	0.41	0.44	0.46	0.49	0.53	0.55	0.55	0.51	0.52
Random	0.22	0.21	0.21	0.19	0.20	0.21	0.21	0.18	0.16	0.19

由數據資料可以發現：方法二較方法一的表現有些微的領先，但在大部分的情況下差距的程度有限。我們認為兩種方法都能取得一部份重要的摘要特徵值，能夠反應一部分段落的重要性。Term-frequency 為主的方法能夠獲取到文章中最常出現的詞彙，抽取到某些重要的詞彙但卻不見得具有語意上的意義，這類的文字以 Ontology-based 的方法即偵測不到，因為在設計 Ontology 時不會考慮到這類的詞彙。然而，兩種方法所用運用的資訊也有部分是重疊的，因為 term-frequency 為主的方法所統計出來的詞彙通常也能部分反映文章主題的資訊。

然而，在只選一個段落時我們可以發現 Ontology-based 的方法能較精準的找到正確的主題，並依此找到最重要的段落。當只選一個段落時，此段落選出憑據特徵主要來自主題的包含與否而 ontology-based 即能依此特性正確的找出該段落。因此，當一篇文章需要高度的壓縮其內容大小時 ontology-based 的方法能夠提供較精準的內容給

使用者。然而這項特性也能夠提供將兩種方法合併時的重要參考。

我們認為 Ontology-based 的方法無法明顯的超越 term-frequency 為主的方法主要原因為：文章的主題若超過 Ontology 所定義的範圍時，Ontology 產生的摘要正確率會變的相當不理想*。或者該詞彙與 Ontology 的詞彙意義相近但卻沒有定義在同義詞中也會降低 Ontology-based 摘要方法的正確率。

---

*附註:本文所實驗的文章只依據新聞的長度過濾過短的新聞，對於文章的主題是否包含於 Ontology 中則未經過過濾。

8. Discussion

實驗的結果反映了 ontology-based 的方法對於自動摘要的效果。然而 Ontology 若能依據系統需求足夠詳細的描述一個領域，對於文章的分析將有非常高的價值。但即便是在一個限定的領域，設計、維護一個 Ontology 也是相當不容易的，需要花費大量的成本。因此要運用在實際的系統上相信仍需一番努力。

未來我們希望能將 Ontology 的架構與同義詞部分予以增加及強化，增加其對應詞彙的能力；對於 Ontology 所涵蓋的範圍也有改進的必要。關於段落的選取數目我們希望能設計一套辦法針對每一篇文章的摘要段落數量做評估，提供系統選取的依據。最後，若能提出一套方法適當的將兩種方法作整合，相信能對系統的表現有相當正面的助益。

References

[1] James Allan, Rahul Gupta, and Vikas Khandelwal, Temporal Summaries of News Topics, In *ACM SIGIR 2001,10-18*, 2001.

[2] Jade Creating, Vibhu Mittal, Jaime Carbonell, Jamie Callan, Evaluating Multi-Document Sentence Extract Summaries, In *Proceedings of the 9th International Conference on Information and Knowledge Management*, 2000

[3] B. Chandrasekaran, J. R. Josephson, and V. R. Benjamins, "What Are Ontologies, and Why Do We Need Them?," *IEEE Intelligent Systems*, **14**(1):20-26, 1999.

[4] Wesley T.Chuang, Jihoon Yang, Extracting Sentence Segment for Text Summarization：A Machine Learning Approach, In *ACM SIGIR 2000*, 152-159, 2000.

[5] Jade Goldstein, Mark Kantrowitz, Vibhu Mittal, Jaime Carbonell, Summarizing Text Documents: Sentence Selection and Evaluation Metrics, In *ACM SIGIR 1999*, 121-128, 1999.

[6] Tom Gruber, Ontology Definition , http://www-ksl.Stanford.edu/kst/what-is-an-ontology.html

[7] Bing Liu, Ninqing Hu, and Wynne Hsu, Multi-Level Organization and Summarization of the Discovered Rules, In *Proceeding of the SIGKDD* 2000, 208-217, 2000.

[8] Kathleen Meown, Dragomir R.Radev,Generating Summaries of Multiple News Articles, In *ACM SIGIR* 1995, 74-82, 1995.

[9] David D. McDonald, The View from the Trenches: Issues in the Ontology of Restricted Domains, In *Proceedings of the international conference on Formal Ontology in Information Systems 2001, 22-33*, 2001.

[10] Deborah L. McGuinness, http://www.ontology.org/main/papers/iccs-dlm.html.

[11] Julian Kupiec, Jan Pederseon, Francine Chen, A Trainable Document Summarizer. In *ACM SIGIR 95*,68-73 , 1995.

[12] John F. Sowa. *Knowledge Representation: Logical, Philosophical, and Computational Foundations*, Brooks/Cole Pub Co, 1999.

[13] Hicham Snoussi, Laurent Magnin, and Jian-Yun Nie, Heterogeneous Web Data Extraction using Ontology, *AOIS, 2001*.

[14] Data source:http://www.sony.com