

Submodular Selection

物件挑選問題一直是被研究的課題，演算法書籍裡耳熟能詳的背包客問題，就是一經典NP-hard問題，當能夠挑選物件數量非常大時，則無法在能夠忍受的時間裡選出最好的物件。所幸在一定的條件下，有近似的解決方法能找到不錯的解答。

在這裡以text summarization做為方法的介紹，常見自動摘要系統有兩種摘要方法，一個是從原文擷取句子做為摘要，而另一種為自動產生摘要，擷取句子做為摘要方法是為一物件挑選問題，後面以如何挑選出最佳摘要句來介紹近似解。

若一篇文章包含 V 個句子，則摘要時會挑選出 $\text{top}k$ 個句子做為摘要，通常 $\text{top}k \ll |V|$ ，既然要挑選摘要句，則要有衡量挑選出的句子的方法，在這裡以 $f(S)$ 做為衡量挑選出句子和 V 關係的function，此關係可以是相似度、多樣性程度等等...， S 為挑選出的句子集合，則目標是挑出 S 集合可以最大化 $f(S)$ ，以數學形式表達為

$$\max_{S \subseteq V} \{f(S) : |S| \leq k\}$$

此解無法在有效時間內解答出，因此有了greedy forward-selection演算法，當 f function符合一定的數學條件後，利用incremental method即可找出近似解，近似解準度 ≥ 0.632 倍最佳解。這是多令人振奮人心的數學框架！只要設計好 f function符合條件，就可以將許多挑選問題利用此演算法解出答案，而 f 的靈活性可以加進許多想要的衡量方式，挑選出想要的句子出來。

要符合的條件有兩個，第一個為monotone，當 $f(S) \leq f(T)$ ， $S \subseteq T$ ，第二個為submodular， $f(S \cup T) + f(S \cap T) \leq f(S) + f(T)$ ， $S, T \subseteq V$ 。第一個條件為單調性，當挑選的摘要句子比較多時， $f(\cdot)$ 值理應較大，反之則小，而第二個條件以著名的邊際效應式子來解釋會比較符合摘要問題。邊際效應以數學式表達為

$f(S \cup \{s\}) - f(S) \leq f(R \cup \{s\}) - f(R)$ ， $R \subseteq S \subseteq V$ ，此式子也符合submodular條件。邊際效應所描述為當資源少時，每多一個小資源能夠增加的效益很大，可是當資源夠多時，再增加一個小資源效益並不大。套在摘要情境為，當挑選的句子數量少時，每多挑一個能夠增加的效益很大，可是當挑選的句子已經夠多時，再多挑一個則增加的效益並不大，因此兩個條件完全符合摘要所需的效果。

下面將講述一些符合此兩條件的 $f(\cdot)$ ，用以衡量 S 和 V 的關係。

1. 摘要為濃縮文章中的重點，因此必須要考量挑選出的句子能夠包含其他句子含意，因此衡量 S 和 V 的相似度是一重要考量要點，以式子表達為 $f_{\text{facility}}(S) = \sum_{i \in V} \max_{j \in S} w_{ij}$ ，此式子考量 V 和 S 中所有元素的相似度，透過 \max 來最大化 S 和 V 相似度，最後挑選出的 S 就會包含所有句子的含意。
2. 好的摘要句不但要濃縮文章重點還要避免挑選出的摘要有相同含意，而造成冗餘摘要，因此設計出能夠懲罰含有冗餘摘要的 $f(\cdot)$ 也很重要。以下簡介一個簡單的懲罰式子， $f_{\text{penalty}}(S) = \sum_{i \in V \setminus S} \sum_{j \in S} -\lambda \sum_{i, j \in S: i \neq j} w_{ij}$ ， $\lambda \geq 0$ ，式子前面一項確保 S 和 V 的相似度，後面一項則是懲罰挑選出冗餘摘要句子，此式子一樣符合monotone和submodular條件，因此保證其最後準度。

講完兩個簡單的 $f(\cdot)$ ，以下簡介incremental method如何將句子一個一個從文章中挑選出來當成摘要句，原始目標函數 $\max_{S \subseteq V} \{f(S) : |S| \leq k\}$ 因為NP-hard問題無法使用，這裡利用greedy方式每次從文章中挑選出能夠增加最大利益的句子出來，遂將目標函數改為 $S^* \in \arg \max_{S \in V \setminus S} f(S \cup \{S\}) - f(S)$ 。

incremental method就是每次從文章中挑選出 S^* ，加進 S 集合中，直到 $|S| = topk$ 。

要修改 $f(\cdot)$ 可以利用+-法簡單增減function的feature，就像上述懲罰式子一樣增加一個懲罰feature就可以保證符合兩條件，若要更嚴苛的feature，可以參考相關的submodular論文。

monotone, submodular function的擴展性強且易實作非常適合拿來應用

Reference

- [1] Graph-based Submodular Selection for Extractive Summarization'10
- [2] Multi-document Summarization via Budgeted Maximization of Submodular Functions'10
- [3] A Class of Submodular Functions for Document Summarization'11