圖書資訊學研究
JOURNAL OF LIBRARY AND INFORMATION SCIENCE RESEARCH

JLISR Open Peer Review Report

Reviewer Comments and Rebuttal to the Comments

Open Point 開放觀點：開放式同儕評閱機制

本刊新採行「開放觀點」（Open Point）機制，以留存作者與評閱者之間論證的寶貴文字紀錄，並鼓勵雙方在同意公開大部分的評論意見與審查回覆。希冀透過「開放式同儕評閱」（Open Peer Review）模式，提供學者窺見已被刊登文章背後，許多同樣值得被理解與被引用的觀點。這項機制有助於我們的作者、評閱者、讀者享有更真實的學術傳播精華。

圖書資訊學研究
JOURNAL OF LIBRARY AND INFORMATION SCIENCE RESEARCH

審 查 文 章 ：以集成學習方法改善 LDA 主題建模穩定性之研究

審 稿 者 ：

匿名評閱者 A （*僅公開評閱意見）

匿名評閱者 B （*僅公開評閱意見）

主編綜評

作　　者　：林頌堅

刊 登 卷 期 ：19 卷 1 期 (2024 年 12 月)

D　O　I　：10.30177/JLISR.202412_19(1).0001

說明：◎評閱者；★主編；●作者

圖書資訊學研究
JOURNAL OF LIBRARY AND INFORMATION SCIENCE RESEARCH

審 查 階 段

初 審

審稿者：匿名審稿者 A

評閱意見：

研究提出兩階段的方案，從而能夠更為穩定地擷取文本主題。除此之外，作者進一步討論該兩階段方案的應用性，以及改善兩階段方案實施成本的可能作法。作者以淺顯易懂的方式說明 Topic Modeling 的現有作法，幫讀者釐清且分類整理各種作法的簡易理論基礎，以及體系框架，對於之後欲從事類似研究的同儕，是很好的參考依據。並以很有邏輯性分析力方式，說服讀者兩階段作法的合理性。整體而言，本研究品質頗佳，研究結果有具有說服力，但有一些細節並未交代，以下有一些建議與意見請作者參酌。

- **作者回覆：感謝評審的意見與建議。**

1. 存在缺漏字的問題，如第 2 頁的「...最早期的主題建方法...」。請再次檢查本文。
   - **作者回覆：對全文再次檢查，盡量避免缺漏字與錯字。**

2. 作者很負責任地進行文件比例與所需基本模型數目搭配的分析（如表 4 與表 5 所示），說明可以用較低的成本達到接近完整地兩階段方案的績效。但建議還是說明所需時間，這樣可以更為具體地知道各種搭配的時間成本。
   - **作者回覆：已針對兩種語料庫的各種比例文本所需的 LDA 建模時間進行測量，結果如第 29 頁之表 5 與表 6。敬請參考。**

3. 表 2 說明建模語料庫的統計特性，建議可以加上兩列，亦即收錄為詞語統計條件的上下限（i.e., 20）

- 作者回覆：已於第 24 頁之表 3（原來之表 2，因前面增加一個表格，調整其編號）增列詞語出現發文數的選取範圍以及文本選擇詞語數的下限。

圖書資訊學研究
JOURNAL OF LIBRARY AND INFORMATION SCIENCE RESEARCH

審稿者：匿名審稿者 B

評閱意見：

本研究提出一個兩階段式的改良版主題建模方法，以改善傳統主題建模方法會有不穩定之主題輸出結果。本研究的作法第一階段透過一般 LDA 主題建模方法產生基本模型，再透過 UMAP 進行降維，其後採用 HDBSCAN 進行 clustering，經檢核後再進入第二階段的引導式主題建模，依據有效主題之機率形成群組代表主題特徵，再加入背景主題並補足應有之主題數目。實驗透過兩個資料集，一個為研究自行蒐集的中文 BBS 資料，另一為網路上之英文研討會論文語料庫。實驗結果證明本研究所提出之方法可以提升主題建模的穩定度。

本研究在方法上有一定程度的創新，透過一系列的資料前處理方法來達到主題模型建模穩定度的提升。根據方法的設計，此兩階段的建模方法很自然地可以達到穩定度的提升，主要原因在於作者對於原先主題分群的結果進行資料後處理，也就是說，作者將傳統主題模型產生的結果進行再次的處理而達到降低結果變異的結果，自然會產生穩定性較高的主題模型 (關鍵字分群)。降低結果的變異有許多方法可以運用，要讓 LDA 產生穩定的結果也有許多不同做法，但何以證明本研究的方法會導致更好的結果?需要更進一步的分析來佐證。

- **作者回覆：感謝評審的意見與建議。針對各項建議答覆如下：**

  ...作者將傳統主題模型產生的結果進行再次的處理而達到降低結果變異的結果，自然會產生穩定性較高的主題模型 (關鍵字分群)。降低結果的變異有許多方法可以運用，要讓 LDA 產生穩定的結果也有許多不同做法，但何以證明本研究的方法會導致更好的結果?需要更進一步的分析來佐證。一併於以下第 1 點回覆。

以下是其他具體建議：

1. 本研究僅跟傳統 LDA 進行實驗比較，但過去已有許多學者提出提升穩定性的做法，作者應該嘗試比較其他更具有代表性的方法，作為研究的 baseline(s)。

- **作者回覆：目前雖然已經提出相當多提升穩定性的做法。但由於各種做法使用的策略不同，應用於不同語料庫時需要調整的參數與程序相當多，主要的幾種方法並沒有程式釋出（特別是與本研究較相關的 Lancichinetti et al., 2015、Maier et al., 2018 和 Koltcov et al., 2016），也缺乏各種參數的設定資料做為調整各種方法最佳成效的依據。因此，目前尚未有比較不同方法成效的研究。近期要完成不同做法的比較研究有困難。未來將可以針對此研究缺口，進行深入的比較研究，因此已將評審的建議列為結論的未來研究方向，請參見第 32 頁。**

2. 文章結構應要有顯著的調整。第一章的段落結構可讀性差，建議應該大幅度調整與切分段落，提升可讀性。此外，研究動機段落應該要特別費心論述，設法將研究缺口做詳盡的描述，如本方法跟過去提升模型穩定性方法的差異，以及補足了甚麼缺陷。

- **作者回覆：(1)已重新撰寫緒論，刪除較不相關的敘述，提升可讀性。(2)緒論中並增加本方法跟過去改變模型參數的推斷方法和選取較穩定的主題兩種主要提升模型穩定性策略之比較，討論這些策略在理論或實務上的問題，相關說明請參見第 5 頁。**

3. 研究方法應該嘗試用更正式的寫法陳述，例如採用虛擬碼陳述演算法流程。

- **作者回覆：已將第一階段與第二階段的虛擬碼演算法分別補充於第 17 頁和第 19 頁。**

4. 實驗設計宜多加考慮其實務意義。舉例來說，本研究強調的是主題模型的結果穩定性，何以需要探討改善方法時間?舉不同文本比例代表之實務意義為何?

- **作者回覆：在實務應用上，需要同時考慮方法的效能、穩定性與效率等各種因素。由於本研究運用集成學習的概念，需要比傳統 LDA 主題建模花費更多的計算時間。為了克服此一問題，本研究嘗試在第一階段只隨機取出語料庫中的部分文件進行建模。為避免節省了時間，卻達不到提高穩定性的效果，因此也針對這個技巧進行實驗與討論。以上相關說明已加入第一節（第 5 頁）。**

5. 請增加段落描述 LDA 之實務應用，並嘗試引用相關之國內外文獻。本論文目前國內期刊文獻引用甚少，請適度引用。

- **作者回覆：已於緒論中增加段落描述 LDA 之實務應用並引用國內外研究，請參見第 3 到 4 頁。**

圖書資訊學研究
JOURNAL OF LIBRARY AND INFORMATION SCIENCE RESEARCH

審稿者：主編綜評

評閱意見：

1. 文獻與相關呈現表達方式參考評閱者建議修正與調整。
   - **作者回覆：已經於緒論中修改與調整論文的表達方式，並增加國內外 LDA 主題建模相關應用的文獻。**

2. 其中請特別留意評閱者二之建議。
   - **作者回覆：已根據評閱者二之建議修改與補充論文。**

3. 評閱者二提到建議 1 關於 baseline 實驗，請說明辯證原因或限制。
   - **作者回覆：目前雖然已經提出相當多提升穩定性的做法。但由於各種做法使用的策略不同，應用於不同語料庫時需要調整的參數與程序相當多，主要的幾種方法並沒有程式釋出（特別是與本研究較相關的 Lancichinetti et al., 2015、Maier et al., 2018 和 Koltcov et al., 2016)，也缺乏各種參數的設定資料做為調整各種方法最佳成效的依據。因此，目前尚未有比較不同方法成效的研究。近期要完成不同做法的比較研究有困難。已將評閱者的建議列為結論的未來研究方向。**

圖書資訊學研究
JOURNAL OF LIBRARY AND INFORMATION SCIENCE RESEARCH

審 查 階 段

複 審

審稿者：匿名審稿者 B

評閱意見：

謝謝作者群的回覆，針對上一輪的意見作者群已做了相對應的修改，論文初稿品質有顯著提升。具體來說，除了第一題與過去學者所提出之方法進行比較，作者將之列為未來研究方向、沒有較正面的回應外，其餘問題均已明確處理。

- **作者回覆：感謝評閱者的辛勞。**