教育資料與圖書館學

Journal of Educational Media & Library Sciences

http://joemls.dils.tku.edu.tw/

Vol. 59 , no. 2 (2022) : 201-232

主題相似性估計與其在主題建模穩定性測量之應用

Estimation of Topic Similarity and Its Application to Measuring Stability of Topic Modeling

林 頌 堅 Sung-Chien Lin

Assistant Professor

English Abstract & Summary see link at the end of this article

教育資料與圖書館學　59 : 2 (2022) : 201-231
DOI:10.6120/JoEMLS.202207_59(2).0014.RS.AM

研究論文

主題相似性估計與其在主題建模穩定性測量之應用

林頌堅

摘要

主題建模的穩定性測量針對相同文本集合以及在相同起始條件下，同一建模方法產生的模型能夠具有相似主題的程度。由於估計主題之間相似性的方法是主題建模穩定性測量的基礎，並且「主題對齊」是這項測量的關鍵步驟。本研究首先根據經由主題對齊之後獲得配對主題相同的比例，比較不同相似性估計方法，並觀察各種方法的相似性分數分布。最後，也分析主題數目對於穩定性測量的影響。本研究使用的主題建模方法是常用的潛在狄利克里分配（LDA）主題建模，並從 PTT BBS Book板上約30,000篇發文產生分析的模型。研究結果觀察到這些相似性估計方法配對主題相同的比例很高，但在配對主題上的相似性分數則有不同的分布，同時也發現隨著主題數目增加，主題建模的穩定性下降。

關鍵詞：主題建模，潛在狄利克里分配（LDA），穩定性測量，主題相似性估計，主題對齊

緒　論

主題建模（topic modeling）假設要分析的文本中包含一個或多個主題，而主題是由一組語意相關的詞語依據特定的比例構成，其目的便是利用數學或統計方法，找出文本集合中蘊含的主題結構。利用主題建模方法可以快速而有效率地協助分析文本內容，並且具有可以處理大量文本資料的可擴展性（scalability），已經愈來愈廣泛應用於各種文本分析的問題上，例如確認與檢索某些特定主題的文件；探討大眾傳播媒體（Jacobi et al., 2016）、政治演說（Quinn

世新大學資訊傳播學系助理教授
Email: scl@mail.shu.edu.tw

此篇文章之同儕評閱意見報告（Open Point）及導讀簡報（InSight Point）請至本刊網站查閱
2022/05/02投稿；2022/07/29修訂；2022/07/30接受

et al., 2010）或社群媒體（Elgesem et al., 2015）上討論的公共議題；追蹤新聞事件的發展（Kim & Oh, 2011）；分析網路評論上使用者對於產品各項設計與功能面向的評價與口碑（Tirunillai & Tellis, 2014）；發現電影等娛樂產品的心理主題特徵（psychological thematic features）與消費之間的關係（Toubia et al., 2019）；甚至應用在軟體工程（software engineering; Agrawal et al., 2018; Panichella et al., 2013; Sun et al., 2016）、研究評鑑（research evaluation; Nichols, 2014）上。以技術來說，機率潛在語意分析（probabilistic latent semantic analysis，簡稱pLSA；Hofmann, 1999）、潛在狄利克里分配（latent Dirichlet allocation，簡稱LDA；Blei et al., 2003）都是常見的主題建模技術，近年來也有將非負矩陣分解（nonnegative matrix factorization，簡稱NMF）技術應用於主題建模（Wang et al., 2012）。

以目前來說，LDA是最為研究者熟知、而且廣泛應用於各領域的主題建模技術（Lancichinetti et al., 2015）。LDA產生的主題模型是一種機率式生成模型（a generative probabilistic model）。假定所有的文件所成的集合中共有K個主題，將每一筆文件視為是由這K個主題依據特定的機率分布混合組成。每一個主題則由所有詞語出現在主題上的機率來表示（Blei et al., 2003），與主題相關的關鍵詞語具有較大的機率；反之，不相關的詞語的機率則相當小。所以主題模型包括兩組機率分布：所有文件上的主題機率分布與所有主題上的詞語機率分布，前者形成的矩陣在主題建模技術中稱為θ，而後者的矩陣則稱為ϕ。主題建模時，給定應用的文件集合和主題的數目K以及產生Dirichlet分布所需的先驗參數α和β，LDA演算法根據α和β隨機產生起始的θ和ϕ。然後再根據當前的θ和ϕ，將輸入文件的詞語分配到每個主題上，重新推導出更精確的θ和ϕ。反覆上述的訓練過程，使得文件的產生有最大的可能性。然後使用者便可以利用這兩組機率分布做為特徵，解讀文件內可能包含的主題以及主題可能的意義。為了更加有效地應用在探索和描述文本內容的主題結構，研究者除了探討與發展主題建模的應用領域之外，很多研究針對LDA的主題模型架構，提出各種不同的衍生模型，如關聯主題模型（correlated topic models，簡稱CTM；Blei & Lafferty, 2007）、動態主題模型（dynamic topic models，簡稱DTM；Blei & Lafferty, 2006）、階層式狄利克雷歷程混合模型（hierarchical Dirichlet processes，簡稱HDP；Teh et al., 2006）等等。另一方面，則是從LDA的模型品質著手，嘗試找出更有效、更穩定描述文本集合的模型。這些品質指標最為研究者所認識的是用來表示主題模型的文本預測能力的對數概似值（log-likelihood）和複雜度（perplexity; Griffiths & Steyvers, 2004），以及表示主題模型之可解釋性（interpretability）的主題協調性（coherence; Röder et al., 2015）。穩定性（stability）也是近來主題建模研究的議題之一（可參見Agrawal et al., 2018, 3.4. LDA, Instability and Tuning; Maier et al., 2018, Appendix）。

穩定性是同一的演算法在相同的輸入資料下，每一次執行能夠得到相同結果的測量指標。穩定性高的主題建模方法便是針對相同文件集合，在相同的主題數目（K）和先驗參數（α和β）等條件下，每次產生的各個模型上能夠有相似的主題。換句話說，產生的每一對模型之間有很高的一致性（agreement）。但是一般LDA的建模結果是不確定的（nondeterministic），在相同的條件下，某幾個模型上出現的主題可能並沒有出現在另外幾個模型中。如此一來，在應用主題建模技術分析文本集合的主題結構時，將無法確定此次建模所得到的主題是穩定或偶然出現的（Koltcov et al., 2016）。僅憑某一次建模所得到的主題模型做為文本內容分析的結果，可能會得到錯誤結論（Agrawal et al., 2018），造成分析結果的信度（reliability）有待商榷，影響主題模型的有用性（Maier et al., 2018）。本研究的目的便是針對LDA主題建模穩定性的測量進行分析。

主題建模穩定性的測量方法有很多，本研究依據De Waal與Barnard（2008）、Greene等（2014）、Belford等（2018）使用的主題建模穩定性測量架構進行研究。此測量架構的過程說明如下：首先在相同的輸入資料（文件集合、主題數目與先驗參數）下，重複進行多次主題建模，產生多個模型。然後，計算任何兩個模型間的一致性分數（agreement score），如果模型之間大多有較高的一致性，也就是主題模型上的主題幾乎都可以在另一個模型上找到相似的主題時，表示主題建模的穩定性較高。因此，將所有一致性分數的平均值做為主題建模穩定性的測量值。例如在主題建模時共產生M個模型，假設第i和j個模型間的一致性分數為$agreement_{ij}$。穩定性的測量值可表示為式(1)的形式，

$$stability \stackrel{\text{def}}{=} \frac{\sum_{i=1}^{M}\sum_{j=i+1}^{M} agreement_{ij}}{M(M-1)/2} \tag{1}$$

計算兩個模型之間的一致性分數則先找出這兩個模型中相似的主題配對，然後以配對的相似性分數平均值做為一致性分數。找出模型之間彼此最佳主題配對組合的步驟稱為主題對齊（topic alignment; Belford et al., 2018; De Waal & Barnard, 2008; Greene et al., 2014）。式(2)以數學形式表達上述想法，

$$agreement_{ij} \stackrel{\text{def}}{=} \frac{\sum_{k=1}^{K} sim(t_{ik}, t_{j\pi(k)})}{K} \tag{2}$$

在式(2)中，t_{ik}表示第i個主題模型的第k個主題，$t_{j\pi(k)}$則是t_{ik}經過主題對齊後在第j個主題模型上配對到的主題，$sim(t_{ik}, t_{j\pi(k)})$表示這個配對的相似性分數。如果經過主題對齊後，兩個模型在最佳配對組合內的主題之間大多具有較高的相似性分數，這兩個模型之間便有較高的一致性。

在上述主題建模穩定性的測量架構中，由於估計兩個主題t_{ik}和t_{jl}之間的相似性分數$sim(t_{ik}, t_{jl})$是測量的基礎，因此本研究將從主題相似性估計方法的分

析與比較開始。可用來估計主題之間相似性分數的方法很多，例如，Jaccard分數（簡稱JAC）、KL散度（Kullback-Leibler divergence，簡稱KLD）、JS散度（Jensen-Shannon divergence，簡稱JSD）和餘弦測量（cosine measure，簡稱COS）等等，不同的估計方法使用不同的主題特徵資訊，例如部分的關鍵詞語集合或詞語的出現機率，估計的方式也不相同。上述穩定性測量架構是建立在主題對齊獲得的主題配對組合上，如果配對相同的情形很高，則兩種不同相似性估計方法在穩定性測量的應用上將有相近的效果。因此，本研究則認為比較不同相似性估計方法時，應觀察不同相似性估計方法在最佳主題配對組合上配對相同的比例，瞭解不同方法應用於計算穩定性上是否有差異。此外，配對組合中可能包含相似性分數較高的配對，也可能包含分數較低的配對，本研究將觀察與比較各種相似性估計方法在配對主題上的相似性分數分布。

本研究並將討論主題數目對於穩定性的影響。主題數目（K）是主題建模相當重要的參數，目前有關主題建模穩定性的研究大多只有測量一種主題數目下的穩定性，只有Greene等（2014）、Ballester與Penner（2022）曾針對不同主題數目如何影響穩定性進行探討。但Ballester與Penner（2022）所使用的穩定性測量方法主要針對應用於文件叢集（document clustering）的主題建模方法上，所使用的概念不同於本研究使用的主題建模穩定性測量架構。Greene等（2014）認為較少的主題數目，將使得每個主題涵蓋的概念較大，出現機率分散在多個詞語上，可能出現的關鍵詞語種類較多；反之，主題數目增加時，每個主題的範圍縮小，主題上關鍵詞語彼此的相關性增加，但主題數目過度增加時，將使得主題的範圍過度狹隘，使得出現機率集中在少數詞語上（Greene et al., 2014）。由於主題建模的過程是反覆根據模型參數隨機地重新分配進行調整，因此可以推測主題數目將會對於主題建模的穩定性造成影響。但Greene等（2014）的研究使用的文本資料都已經有明確的主題，例如新聞語料庫上的版面資訊，而且主題數目都相當小。因此本研究將以主題不明確且數量較多的文本資料討論這個問題。

綜上所述，本研究將進行以下的觀察與分析：

㈠不同相似性估計方法在最佳主題配對組合上配對相同的比例，

㈡各種相似性估計方法在配對主題上的相似性分數分布，

㈢主題數目對於主題建模穩定性的影響。

本論文的章節結構如下：本節說明研究的動機與目的，簡要說明主題建模穩定性的測量方法與本研究將探討的問題；接下來，將對有關主題建模穩定性測量的研究以及其中最重要的主題相似性估計方法進行文獻回顧；再接下來說明研究中使用的文本資料、主題建模、主題相似性估計方法與穩定性的測量方法；最後的兩節分別是研究結果與結論。

二、相關研究

本節首先說明過去有關主題建模穩定性以及測量方法的研究，然後討論對穩定性測量相當重要的主題相似性估計方法。

㈠主題建模的穩定性以及測量方法

利用LDA主題建模程式進行文本內容分析的研究大多假定建立的主題是真實且一致的，結果具有相當的可重複產生性（reproducibility）。因此，這些研究除了調整模型的主題數目以外，對於產生主題在文件上的機率分布和詞語在主題上的機率分布的先驗參數α和β，往往採用程式預設的參數值，而且通常只採用一次建模所得到的結果，很少重複執行多次建模（Belford et al., 2018; Maier et al., 2018）。然而實際上，即便使用相同參數以及相同文本，每次建模產生的主題模型往往會有一些差異。這種不穩定的情形導致應用LDA主題建模在自動內容分析的有用性在近年越來越受到質疑（Belford et al., 2018; Chuang et al., 2015）。

根據以上的說明，測量主題建模的穩定性需要經由計算多次建立的主題模型之間的一致性分數，如果多個結果模型彼此一致的話，主題建模的結果便可認為是比較穩定的。比較這些模型的一致性有兩種做法：一種是Maier等（2018）與Belford等（2018）所建議的做法：在相同的參數下，對相同文本執行M次主題建模，獲得M個模型，然後計算全部$M(M–1)/2$對模型之間的一致性分數，再進行平均或以其中位數做為主題建模穩定性的測量值；另一種方法則是由Greene等（2014）提出，利用全部文件訓練、較完整的模型做為參考模型，以參考模型為主，計算它與其他（M-1）個只取部分文件訓練、較弱模型之間的一致性分數，再進行平均或取中位數。

既然主題建模的初始化與建模過程都是隨機的，每次建模所得到的主題次序與內容不大可能完全相同。在模型A上編號為k的主題可能與模型B上同樣編號k的主題相差很大，但與編號k'的另一個主題較相似。這種情形將造成計算兩個模型之間一致性的問題。因此，De Waal與Barnard（2008）、Greene等（2014）、Belford等（2018）建議在計算兩個主題模型的一致性分數時，可先將兩個模型之間主題的相似性分數輸入匈牙利演算法（Hungarian algorithm; Kuhn, 1955）進行主題對齊，一對一匹配兩個模型上相似的主題，獲得兩個模型的最佳主題配對組合。然後再以最佳配對組合內的主題配對估計這兩個模型的一致性分數。以下簡要說明上述研究應用匈牙利演算法進行主題對齊並計算一致性分數的方式，附錄中將提供匈牙利演算法的程序與一個簡單的主題對齊範例。

De Waal與Barnard（2008）提出根據兩個模型在文件上主題出現機率分布（θ）計算主題模型一致性的方法。他們建議先估計兩個模型之間主題的相似

性，然後將兩個模型共有K^2對的主題相似性分數輸入匈牙利演算法進行主題對齊，找出兩個模型的最佳配對組合。如果主題建模的方法穩定，同一筆文件在不同次的建模結果中彼此應具有相似的主題。因此，他們將主題視為文件的特徵，兩個模型在同一文件上的主題出現機率分布則是文件的兩組特徵值。當兩個模型進行主題對齊之後，可以利用所有文件上的主題機率分布，比較兩組特徵值的相關性，以相關性的高低表示模型之間一致性的大小。

Belford等（2018）和Greene等（2014）都以每一個主題上前T個出現機率較高的詞語集合代表各個主題。在估計兩個模型之間所有主題的相似性分數之後，將所有相似性分數輸入匈牙利演算法，找出兩個模型最佳的主題配對組合。兩個研究都將模型之間的一致性定義為最佳配對組合內每一對配對JAC之平均值。

Yang等（2016）首先利用匈牙利演算法進行主題對齊，然後將產生的主題配對組合應用在主題模型之間一致性的測量。他們提出三種主題模型一致性分數的測量方法：第一種方法先將要進行一致性測量的兩個主題模型分別應用於文件主題指定（document topic assignment），也就是文件中出現機率最高的主題。如果同一文件在兩個模型中指定的主題分別是主題對齊產生的配對主題，文件的主題指定便是一致的，而文件集合內主題指定一致的文件比例愈高，這兩個模型之間的一致性分數便愈高。第二種方法與Greene等（2014）、Belford等（2018）同樣使用每一個主題上前T個出現機率較高的詞語集合代表各個主題，但Yang等（2016）不使用配對主題JAC之平均值，而是將兩個模型的一致性定義為它們之間所有配對主題上詞語相同的比例，當配對主題的詞語相同比例愈高，兩個模型便愈一致。第三種方法則將主題模型應用在文件上每一個詞語的主題指定（token topic assignment），也就是決定文件上每一個詞語為主題模型上的哪一個主題，將文件集合內主題指定為同一對配對主題的詞語之比例視為主題模型之間的一致性。

使用匈牙利演算法對兩個模型中的主題進行主題對齊，其運算複雜度為$O(K^3)$，所以也有其他的研究採用較簡單的方法來計算主題模型的一致性。Maier等（2018）將兩個主題模型的一致性定義為完成配對的主題數量佔模型主題數目的比例。並定義某一個主題t_{ik}與另一個主題t_{jl}完成配對的條件為t_{ik}是其所屬模型中與t_{jl}相似性分數最高的主題，而且其分數超過0.7。

其他研究則提出不需先對模型進行主題對齊的穩定性測量方法。Belford等（2018）關於主題建模穩定性的研究中，除了前述利用匈牙利演算法對齊模型主題計算模型一致性分數的方法之外，另外提出其他兩種測量穩定性的方法：第一種方法對主題模型的每個主題，選出T個出現機率最高的關鍵詞語，然後以所有主題的關鍵詞語集合做為模型的代表特徵。主題模型對另一個模型的差異比率則定義為兩者關鍵詞語集合的集合差（set difference）大小佔所有可能主題

關鍵詞數目（$K \times T$）的比率。如果兩個關鍵詞語集合完全相同，這對模型的差異比率為0；如果完全不同，差異比率為1。計算出每一對模型的差異比率後，再以差異比率的平均值做為主題建模穩定性的測量值。如果所有模型的差異比率平均值接近0，表示建模的結果相當穩定。

Belford等（2018）的另一種方法利用經常用於測量叢集一致性（clustering agreement）的正規化交互資訊（normalized mutual information; Strehl & Ghosh, 2002）估算兩次建模結果之間的一致性。他們簡化主題建模方法所具有的機率叢集（probabilistic clustering）特性，只使用每個文件的主要主題（dominant topic），也就是該文件上出現機率最大的主題，視為一種將文件進行叢集分析（cluster analysis）所得到的劃分（partition）[1]。在第一次建模結果中，某一組具有相同主要主題的文件，如果在第二次結果也具有相同的主要主題，也就是兩次叢集的劃分結果相同。如果大多數相關的文件在兩次建模都具有相同主要主題的情形，此時便可獲得較高的正規化交互資訊，而可認為兩個模型相當一致。因此，將主題建模的穩定性定義為每一對模型之間正規化交互資訊的平均值。

Agrawal等（2018）利用主題中出現機率最高的前T個詞語代表主題，利用主題上的詞語在多次建模結果中重複出現的次數測量穩定性。當在M次的建模結果中，對於某一個主題，假定能發現與其有t個詞語相同的主題共有m次時，這個主題在t個詞語時的重複比例被定義為m/M。Agrawal等（2018）將整個模型在t個詞語時的穩定性分數定義為每個主題重複比例的中位數。在比較各種主題建模方法的穩定性時，將每種方法各產生M次的訓練結果，計算t從1到T個詞語時的穩定性分數。大抵來說，隨著t增加，穩定性分數會降低，但比較穩定的主題建模方法的分數下降程度較小，也就是主題中包含較多重複出現的詞語。

Ballester與Penner（2022）探討與比較LDA、NMF和Doc2Vec三種主題建模方法[2]的統計強健性（statistical robustness）、描述力（descriptive power）和反映真實（reflect reality）等三種品質，他們並認為主題建模方法提供了比其他文件叢集方法更好的文件相似性計算，因此在比較主題建模方法的品質時，應該著重其在文件相似性的計算上。Ballester與Penner（2022）說明統計強健性的意涵為「在相同資料上，以相同參數執行相同建模應該產生相同或至少極為相似的結果」，事實上便是本研究所探討主題建模時的穩定性。綜上所述，在測量某種主題建模方法的統計強健性時，先產生多個模型，然後計算每一對文件相似性分數在所有模型上的標準差（standard deviation），較大的標準差表示該對文件在

[1] Belford等（2018）所指之文件的主要主題也就是Yang等（2016）的文件主題指定結果。

[2] Doc2Vec利用類神經網路（neural network）的方式推導代表每個文件的特徵向量，稱為文件的嵌入（document embedding）。語意相似的文件，其嵌入之間的餘弦相似性較高。但因為無法解讀嵌入上的每一個元素代表的意義，嚴格來說，Doc2Vec並不能算是主題建模方法。

應用不同模型所得到的相似性分數有較大的差異。如果將所有成對文件的相似性分數標準差進行平均後，產生較大的數值，便表示這種主題建模方法並不是相當強健，也就是不穩定。Ballester與Penner（2022）建議文件相似性分數的計算，在LDA和NMF上可以利用文件的主題機率分布進行COS，Doc2Vec則可將COS應用在代表文件的嵌入（embedding）上。Ballester與Penner（2022）的研究指出，在三種主題建模方法中，Doc2Vec在各種主題數目下，相較於其他兩種方法，在強健性上都有不錯的結果；LDA則是在採用較多主題進行建模時較為強健，但在較少主題時並不理想；NMF則在主題數目增加時，有不佳的強健性。

㈡主題之間相似性估計方法

主題模型上包含的兩種資訊，文件上的主題出現機率分布θ和詞語在主題上的機率分布ϕ，都可以運用來估計主題之間的相似性分數。在使用θ進行估計時，主題可視為文件一種特徵表現。例如De Waal與Barnard（2008）建議兩個主題的相似性分數可定義為兩個主題在所有文件上出現機率的乘積總和。以兩個主題t_a與t_b為例，假定它們在D筆文件$d_1 \sim d_D$上的出現機率分別是$p_{1a}\sim p_{Da}$和$p_{1b}\sim p_{Db}$，De Waal與Barnard（2008）將它們的相似性分數定義為$\sum_{i=1}^{D} p_{ia}p_{ib}$。當兩個主題在各文件上的出現機率相似時，所估算得到的相似性分數較高。

另一方面，在運用ϕ估計主題之間的相似性時，可依據主題的代表特徵分為三類。下面以表1上的簡單例子說明上述三類代表特徵以及使用這些特徵的相似性估計方法。在這個例子中，詞彙中詞語的總數共有七個，分別是w_1~w_7，t_a、t_b與t_c是三個要進行相似性估計的主題，表格上的數值則代表對應的詞語在主題上的機率。

表1 以詞語在主題上機率（ϕ）估計主題相似性的簡例

主題	w_1	w_2	w_3	w_4	w_5	w_6	w_7
t_a	0.09	0.15	0.10	0.02	0.25	0.18	0.21
t_b	0.22	0.18	0.03	0.19	0.17	0.05	0.16
t_c	0.09	0.12	0.08	0.10	0.33	0.07	0.21

1. 以ϕ上的機率值代表主題的特徵：將表上的主題t_a表示為[0.09, 0.15, 0.10, 0.02, 0.25, 0.18, 0.21]形式之特徵向量。利用詞語出現在主題的機率值代表主題的特徵，因此相似主題彼此有相似的機率分布。常見的相似性估計方法有KLD、JSD、Pearson相關係數、COS和折扣累積效益（discounted cumulative gain，簡稱DCG）等。以表1為例，如果採用JSD做為相似性估計方法，t_a與t_b的JSD分數約為0.09，t_a與t_c的分數則為0.03。因為JSD分數愈小，兩個機率分布便愈相似，所以t_a與t_c比t_a與t_b更相似。

2. 以ϕ上前T個機率值較高的關鍵詞語所形成的集合代表主題的特徵：假定取機率值較高的前四個關鍵詞語代表主題，表1上的主題t_a便可表示為$\{w_5, w_7, w_6, w_2\}$，而t_b與t_c則分別可表示為$\{w_4, w_2, w_5, w_7\}$和$\{w_5, w_7, w_2, w_4\}$。相似的主題上彼此應具有相似的關鍵詞語集合，這類的相似性估計方法包括JAC和Dice分數等。如果採用JAC做為相似性估計方法，雖然t_b與t_c的關鍵詞語在集合內次序不同，但兩者包含相同的關鍵詞語。因此，t_a與t_b和t_a與t_c的JAC都是0.6。
3. 以ϕ上所有詞語或前T個關鍵詞語的順序代表主題的特徵：這類的方法有Spearman等級相關係數、Kendall τ係數（Kendall's τ coefficient，簡稱KEN）和等級偏向重疊分數（rank biased overlap，簡稱RBO）。假定採用RBO做為相似性估計方法利用關鍵詞語的順序比較兩個主題的相似性，表1上的各個主題同樣取機率值較高的前四個關鍵詞語代表。此時，因為t_a與t_c上的關鍵詞語順序比較相似，其分數約為0.85，比t_a與t_b的分數0.27大。也就是t_a與t_c比t_a與t_b更相似。

Mantyla等（2018）、Kim與Oh（2011）以及Niekler與Jähnichen（2012）比較不同的主題相似性估計方法的相關研究。Mantyla等（2018）採用Spearman等級相關係數、JAC和RBO等多種方式估計主題之間的相似性，然後計算主題建模的穩定性測量。結果發現這些方法所得到的穩定性分數之間有很高的正相關性。Kim與Oh（2011）、Niekler與Jähnichen（2012）的研究雖然不是針對使用相同文本集合進行多次訓練產生的主題模型，但他們將主題相似性運用在不同時間區段所產生的主題模型，找出各個模型中相似的主題，追蹤主題在時間上的演化情形，也是主題相似性估計的應用。Kim與Oh（2011）比較JAC、KLD、JSD、COS、KEN和DCG等六種方式來估計前後時期兩個主題之間的相似性，以找出最相似的主題，做為新聞中持續出現的議題（issues）。在Niekler與Jähnichen（2012）的研究中，他們以每天的新聞建立主題模型，挑選出每天都出現的主題，然後應用了JSD、COS和Dice分數等三種方法估計兩個日期中所有主題之間的相似性。

㈢本節小結

從上述的文獻探討可觀察到目前在主題建模穩定性的測量方法中，較主流的測量架構是首先訓練出多個模型，然後針對每兩個模型計算其一致性分數，以所有一致性分數的平均值做為穩定性的測量值（Belford et al., 2018; De Waal & Barnard, 2008; Greene et al., 2014）。計算兩個模型之間的一致性分數時，由於每個模型上的主題次序與內容不大可能完全相同，因此關鍵步驟是進行主題對齊，找出兩個模型之間的最佳主題配對組合，然後再比對兩個模型應用於文件主題指定結果或估計配對主題的相似性。由於主題建模方法的應用不僅是瞭解目前文件集合內各文件具有的主題分布θ，更可進一步利用各主題上的詞語機

率分布ϕ去推論新進文件上的主題分布，且透過主題詞語機率分布，較容易解讀主題建模的結果，因此本研究將採用Belford等（2018）和Greene等（2014）使用的主題建模穩定性測量框架，並利用配對主題相似性的平均值計算模型之間的一致性分數。

另一方面，Belford等（2018）和Greene等（2014）以每一個主題上前T個出現機率較高的詞語集合來估計主題之間的相似性。相較於使用所有詞語的機率值或機率的大小順序，這種方式所使用的數據量相當少。僅使用主題詞語機率分布上少部分資訊來估計主題相似性，是否會對相似性的估計結果，甚至主題建模的穩定性測量結果產生影響，值得進一步探討。因此本研究將應用多種主題相似性方法於主題建模的穩定性測量，計算主題對齊產生配對結果相同的比例，比較各種相似性估計方法，並觀察這些方法在模型之間最佳主題配對組合上的相似性分數分布。

最後，目前只有Greene等（2014）以及Ballester與Penner（2022）等少數研究曾針對不同主題數目如何影響穩定性進行探討，但Ballester與Penner（2022）所使用的穩定性測量方法主要針對應用於文件叢集的主題建模方法上，而Greene等（2014）使用的文本資料都已經有明確的主題，且主題數目都相當小。因此本研究將使用主題不明確且數量較多的文本資料，分析主題數目對穩定性測量的影響。

三、研究方法

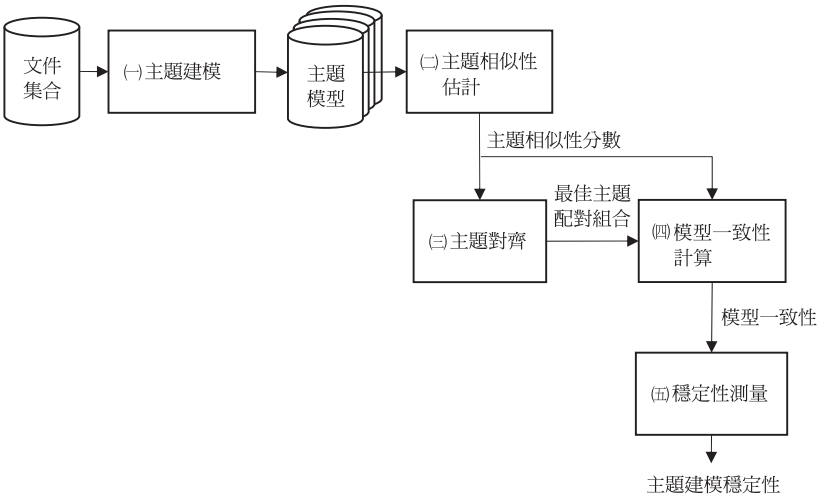
本研究在測量主題建模的穩定性時所採用De Waal與Barnard（2008）、Greene等（2014）、Belford等（2018）使用的主題建模穩定性測量架構，其過程如圖1所示：㈠對文本資料的集合，利用LDA主題建模程式，進行多次主題建模，訓練出M個模型；㈡估計每一對模型上各個主題之間的相似性；㈢利用這些相似性分數，經過主題對齊後，找出平均相似性分數最佳的主題配對組合；㈣以最佳配對組合上的相似性分數計算這一對模型的一致性分數；㈤將所有$M(M-1)/2$對模型的一致性分數進行平均，做為穩定性的測量值。以下說明本研究使用的文本資料、主題建模以及各種相似性估計方法，最後是主題模型一致性的計算與主題建模的穩定性測量。

㈠文本資料

本研究從批踢踢實業坊電子布告欄系統（PTT BBS）上蒐集估計主題建模穩定性的文本資料。本研究選擇PTT BBS書板（https://www.ptt.cc/bbs/book/index.html）上網友發布的文章，利用自行撰寫的程式蒐集發文內容，建立語料庫，蒐集的時間範圍自2009年1月起到2021年4月，共獲得32,895筆發文。

由於書板上發文內容主要以中文書寫，所以在進行主題建模前，需要先經

圖1　主題建模穩定性測量過程示意圖



過斷詞處理（word segmentation）。本研究採用中央研究院詞知識庫小組開發的ckiptagger（Li et al., 2020）做為斷詞系統，將輸入的發文內容切分為詞語的序列。同時，也將斷詞結果輸入ckiptagger的詞類標示（part-of-speech tagging）模組，標示出每個詞語對應的詞類。

要進行主題建模，可先建立建模用的詞典，彙整語料庫內所有文本資料出現的詞語，過濾較不重要的停用詞。本研究在過濾停用詞時，依據詞語的詞類和在整個語料庫上出現的總次數和發文數進行過濾，保留普通名詞、專有名詞、地方詞、名物化動詞與非謂形容詞等詞類的詞語，但刪除出現次數少於50次或出現發文數在總發文數1/10以上的詞語，共得到3,043種不同詞語。

最後建立建模用的文本集合。以詞典統計每筆發文上出現的詞語種類，選擇內容中至少包含五種詞語的發文，做為文本集合。最後集合內共計20,287筆發文，所有發文上出現的詞語總數為1,579,116個詞。

㈡主題建模

將詞典與分析的文本集合以及設定的參數輸入主題建模，產生主題模型。本研究採用python上較多人使用的主題建模套件gensim（https://radimrehurek.com/gensim/），版本為3.6.0。但為了得到較佳的主題模型結果，在gensim上取用University of Massachusetts開發的主題建模軟體Mallet（MAchine Learning for LanguagE Toolkit, http://mallet.cs.umass.edu/）進行建模，使用的Mallet版本為2.0.8。

在本研究中，固定先驗參數α和β，針對不同主題數目（K = 5, 10, 15, …, 100），各建立20個模型[3]。對於模型數目的選擇上，需要足夠多的模型才能確認

[3] 主題建模所使用的先驗參數α設為50，β採用Gensim套件的預設值，訓練次數（iteration）也採用Gensim套件預設值1,000。

每次執行產生的模型主題是否不穩定，但又因為需要估計每一對模型的主題相似性，其複雜度為$O(M^2)$，模型數目過多，也會影響研究的效率。本研究參考Mantyla等（2018）的研究設計，將模型數目設為20。這些主題模型將用來測量模型間主題的相似性，分析與比較不同的相似性估計方法以及探討主題數目對主題建模穩定性的影響。

㈢ 相似性估計方法

根據前面對於相似性估計方法的探討，主題的特徵可以使用：1.詞語的出現機率分布，2.機率較高的關鍵詞語集合，3.詞語的出現機率順序等方式代表。本研究從各類的代表特徵中選出六種方法應用於主題建模的穩定性測量，使每種主題特徵類型至少有一種方法。本研究並調整各種方法輸出結果，使得所有的估計範圍在0與1之間，並且主題之間愈相似者，其估計分數愈大。以下說明這六種方法以及本研究如何調整與應用。

1. JS散度（JSD）

JSD可以估計兩個機率分布之間的差異，是以KLD為基礎的延伸，目的是為了改善KLD不對稱、計算分數的範圍不定（有可能為無限大）等問題。JSD的範圍在0與1之間，如果兩個機率分布愈相似，則它們之間的JSD愈小（Kim & Oh, 2011）。因此，當我們以詞語的機率分布分別代表主題，假定第i個模型的第k個主題t_{ik}和其在第j個模型上第l個主題t_{jl}的詞語機率分布分別是ϕ_{ik}和ϕ_{jl}，利用JSD估計這兩個主題的相似性分數$sim_{JSD}(t_{ik}, t_{jl})$時，可以定義為1減去它們之間的JSD分數$JSdiv(\phi_{ik}||\phi_{jl})$，也就是$sim_{JSD}(t_{ik}, t_{jl}) \stackrel{\text{def}}{=} 1 - JSdiv(\phi_{ik}||\phi_{jl})$。

2. 正規折扣累積效益（NDCG）

DCG是一種排序品質的測量方法，經常用來評估搜尋引擎演算法的有效性（Järvelin & Kekäläinen, 2002）。由於搜尋引擎的檢索應該盡量將相關性高的答案排在結果前列。所以當評估搜尋引擎時，其成效的計算方式是將所有預測結果與正確答案相比的相關性分數除以正確答案所在位置的對數值，藉此減少後列正確答案的重要性，最後將這些經過折扣的分數加總起來。在本研究的應用上，考慮為了使相關性分數估計的結果範圍在0與1間，將採用正規折扣累積效益（normalized discounted cumulative gain，簡稱NDCG）。並且因為NDCG的計算並不是對稱的，也就是$\mathrm{NDCG}(\phi_{ik}, \phi_{jl}) \neq \mathrm{NDCG}(\phi_{jl}, \phi_{ik})$，所以將主題$t_{ik}$與主題$t_{jl}$的相似性分數$sim_{NDCG}(t_{ik}, t_{jl})$定義為$(\mathrm{NDCG}(\phi_{ik}, \phi_{jl}) + \mathrm{NDCG}(\phi_{jl}, \phi_{ik}))/2$，使其具備對稱性。

3. 餘弦測量（COS）

COS以兩個向量之間夾角大小（Maier et al., 2018），評估這兩個向量方向的相似度，如果這兩個向量的方向完全相似，COS的測量結果為1，如果完全相反，測量結果為–1。因此，在測量主題t_{ik}與另一個模型的主題t_{jl}的相似性分數

$sim_{cos}(t_{ik}, t_{jl})$時，可以將它們的詞語機率分布$\phi_{ik}$和$\phi_{jl}$視為是特徵向量，利用COS進行估計，也就是$sim_{cos}(t_{ik}, t_{jl}) \stackrel{\text{def}}{=} Cos(\phi_{ik}, \phi_{jl})$。雖然COS的結果範圍在–1到1之間，但因為$\phi$上詞語的機率值都是大於或等於0，所以相似性分數$sim_{cos}(t_{ik}, t_{jl})$的值範圍在0與1之間。

4. Jaccard 分數（JAC）

JAC經常使用於估計兩個集合的相似性，其計算方式為兩個集合的交集內的元素個數除以聯集內的元素個數。如果兩個集合內的元素相當相似時，它們的交集和聯集中的元素都和它們相似，所以其JAC接近1；反之，兩個集合內的元素相當不同時，它們交集內的元素個數比起聯集內的元素個數少很多，此時的JAC接近0。以出現機率較高的前面數個關鍵詞語所形成的集合代表主題時，便可採用JAC估計任何一對主題的相似性分數（Belford et al., 2018）。例如主題t_{ik}與主題t_{jl}的前T個機率最高的詞語所成的集合分別是R_{ik}與R_{jl}，估計相似性分數$sim_{JAC}(t_{ik}, t_{jl})$，可以定義為$Jaccard(R_{ik}, R_{jl})$。本研究參考Belford等（2018）的研究設定，選用前10個機率最高的關鍵詞語所成的集合代表主題，也就是$T = 10$。

5. 等級偏向重疊分數（RBO）

正如先前在相關研究的討論，利用JAC估計兩個主題的相似性分數只考慮兩個主題的關鍵詞語的重疊性，並沒有考慮詞語對主題的重要性，而這樣的重要性反應在詞語在主題上的機率以及其順序。RBO可以考慮關鍵詞語在主題上的重要性，使得機率較大的詞語在計算相似性分數時能夠有比較大的影響力（Webber et al., 2010）。RBO的結果範圍在0與1之間，如果RBO分數為0，表示這兩個主題上的關鍵詞語完全不同；如果RBO分數較大，表示這兩個主題的關鍵詞語與其重要性順序都很接近（Mantyla et al., 2018）。因此，本研究將相似性分數$sim_{RBO}(t_{ik}, t_{jl})$定義為$RBO(R_{ik}, R_{jl})$。與JAC相同，本研究選用前10個機率最高的詞語所成的集合代表主題，但需要注意的是輸入$RBO(R_{ik}, R_{jl})$的R_{ik}與R_{jl}上的每一個詞語則必須按照它們在主題上的機率排序。

6. Kendall τ係數（KEN）

KEN是用來估計兩個數列之順序關聯性的相關係數，是一種無母數（non-parametric）統計方法。KEN計算某一個數列上的資料項目與其他項目的相對順序關係在另一個數列上是否能夠維持的個數，當其相對順序關係都能夠保持時，它們的KEN值為1，如果都無法保持時，它們的KEN值為–1。本研究在估計主題t_{ik}與主題t_{jl}的相似性分數$sim_{KEN}(t_{ik}, t_{jl})$時，基於詞語機率分布$\phi_{ik}$和$\phi_{jl}$上的機率值估計兩個主題上詞語之順序關聯性的相關係數，也就是$Kendall(\phi_{ik}, \phi_{jl})$。但是KEN的結果範圍在–1與1之間。本研究將小於0的值都調整為0，使相似性分數$sim_{KEN}(t_{ik}, t_{jl})$的值範圍在0與1之間。

這六種方法的前三項（JSD、NDCG和COS）都是根據所有詞語的機率做為主題特徵，JAC和RBO則是選取部分機率較高的關鍵詞語集合代表主題，其中RBO還考慮詞語的機率順序，最後KEN則是以所有詞語的機率順序做為代表主題的特徵。

㈣ 主題模型一致性計算與主題建模穩定性測量

在估計兩個模型之間的所有相似性分數之後，將這些分數輸入匈牙利演算法，進行主題對齊，產生最佳主題配對組合。在獲得最佳主題配對組合之後，本研究將計算配對結果相同的比例，比較各種相似性估計方法，並觀察這些方法在模型之間最佳主題配對組合上的相似性分數分布。

最後，將每一對最佳主題配對的相似性分數進行平均，做為這兩個主題模型之間一致性分數，並以每一對主題模型之間一致性分數的平均值做為穩定性的測量值。本研究將分析主題數目對穩定性測量的影響。

四、主題建模穩定性的測量結果分析

㈠ 不同相似性估計方法之間主題配對相同的比例

穩定性測量的目的是評估在相同起始條件下，主題建模方法每次產生模型具有相似主題的程度，換言之，在對兩個模型進行主題對齊後，模型上大多數的主題能否配對到最相似的主題是穩定性測量的重要因素。因此，如果不同方法在主題對齊後獲得近似的最佳配對組合，表示這些方法應用在穩定性測量上有相近的效果。本研究針對LDA主題建模在相同起始條件下產生的20個主題模型，統計六種方法中配對結果相同的方法數量。由於研究時程與篇幅所限，目前只將主題數目設定為25，未來可進一步觀察不同主題數目下的主題配對相似性分數分布。結果如表2，左欄是主題對齊配對結果相同的方法數量，右欄則是該類型配對佔所有配對（4,750對[4]）的百分比。由於在本研究中，並沒有發現六種方法都不同的主題配對結果，因此便沒有呈現在表2上。

表2　25個主題的主題建模主題對齊結果相同的方法數量佔比

主題對齊配對結果相同情形	佔比（%）
六種方法都相同	76.99
僅其中五種方法相同	8.51
僅其中四種方法相同	8.59
僅其中三種方法相同	4.80
僅其中兩種方法相同	1.11

[4] 本研究針對LDA主題建模在相同起始條件下產生20個主題模型，總共190（20 × 19 / 2）對模型，每對模型產生25個主題配對，因此共有4,750對。

表2的配對結果可觀察到不同方法之間有很高比例獲得相同的情形。六種方法都相同的配對結果達到76.99%，四種或四種以上方法相同配對其佔比總和更達到94.09%（76.99% + 8.51% + 8.59%）。換言之，如果主題建模的穩定性測量目的是「在相同的主題數目（K）和先驗參數（α和β）下，針對相同文件集合的每次建模產生的主題應該是相似的」前提下，應用本研究探討的這幾種相似性估計方法可以達到大致相同的效果。

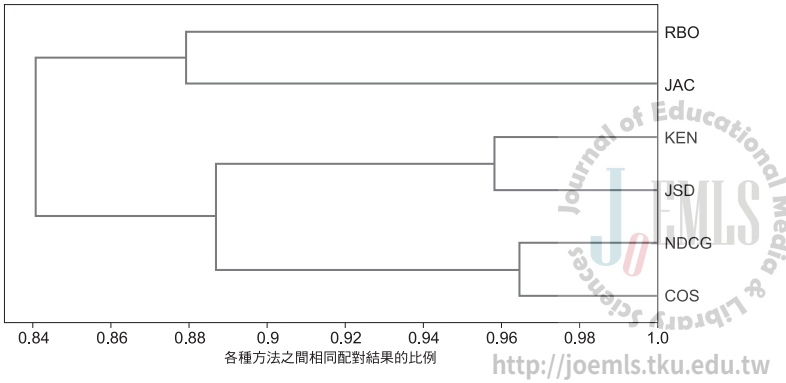
接下來，為了進一步瞭解不同的相似性估計方法之間哪些有更為接近的配對結果？本研究比較不同的方法，兩兩間具有相同配對結果的比例，比較結果呈現於表3。由於每對方法之間具有相同配對結果的比較結果是對稱的，因此表3只呈現比較結果的下半部，查看兩種方法具有相同配對結果的比例，可從表上兩種方法分別對應的行與列上取得所需的數值，例如查看COS與NDCG的比較結果，可從表上COS這一行與NDCG這一列上的數值取得。

表3　兩種相似性估計方法具有相同配對結果比例

	JSD	NDCG	COS	JAC	RBO
NDCG	90.88%				
COS	90.15%	96.48%			
JAC	85.81%	84.59%	84.06%		
RBO	84.67%	87.71%	86.80%	87.92%	
KEN	95.83%	89.68%	88.69%	85.07%	84.38%

在表3上，任何兩種方法具有相同配對結果的比例全都在84%以上，也就是任何兩種方法之間都有相當接近的配對結果。本研究並將表3的結果輸入完整連結叢集演算法（complete-linkage clustering algorithm），找出配對結果接近的方法。從圖2的叢集結果可以觀察到這些方法可分為兩組，第一組是COS、NDCG、JSD和KEN，第二組則是JAC和RBO，在同組內的各種方法有更接近的配對結果。前一組方法都是運用詞典中所有詞語在主題上的機率ϕ的數值或順序，後一組則都僅利用機率最大的前10個關鍵詞語上的資訊。這可能是造成這兩組方法之間有差異的主要緣故。

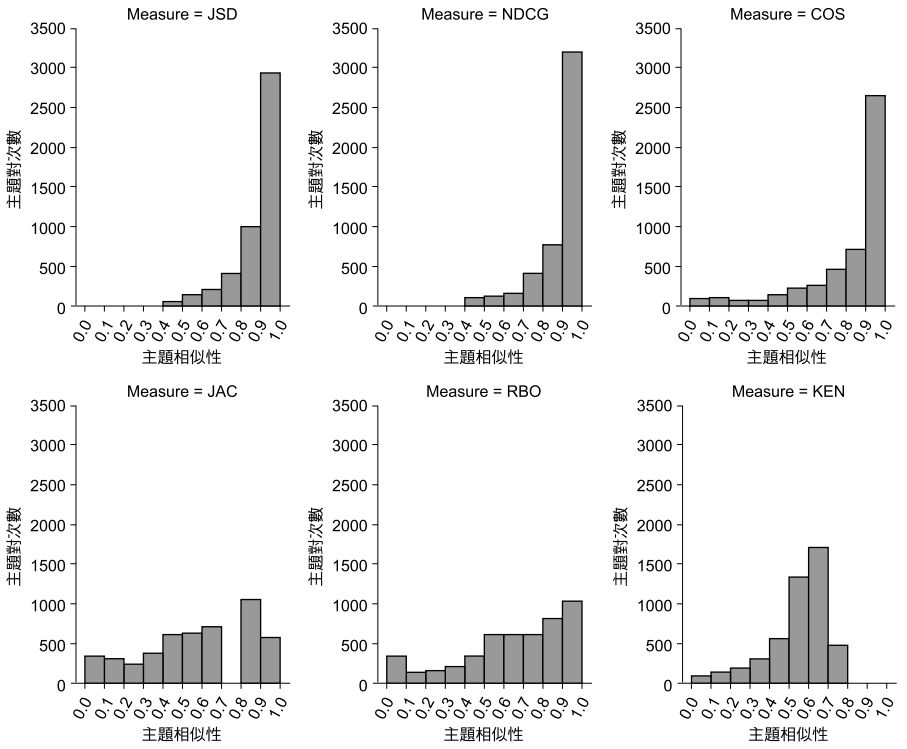
圖2　根據相同配對結果比例，將六種主題相似性估計方法分組結果



㈡最佳主題配對組合中各種相似性估計方法的分數分布

本研究接著觀察各種相似性估計方法在最佳主題配對組合上配對的分數分布，瞭解不同方法使用的主題特徵資訊與計算方式的差異。圖3表示各種估計方法的主題配對相似性分數分布。

圖3　各種相似性估計方法的主題配對相似性分數分布



以下依據代表主題的特徵方式，將六種相似性估計方法分為三組進行說明：

1. JSD、NDCG和COS

這三種方法都是利用所有詞語在主題上的出現機率作為特徵。JSD和NDCG的結果範圍相似，大約分布在0.4到1.0之間，且大部分配對主題之間都具有相當高的相似性分數，0.9到1.0之間分別有2,945對（佔全部4,750對主題的62.00%）與3,201對（67.39%），少於0.5的配對則各佔1.22%與2.29%。COS的分布範圍則在0.0到1.0之間，但大部分配對結果也有相當高的相似性分數，相似性在0.9到1.0之間有2,650對（55.79%），少於0.5的配對僅佔9.68%，且在0.0到0.1之間，佔全部的1.75%。

2. JAC和RBO

JAC和RBO都是採用關鍵詞語集合做為代表主題的特徵。相較於其他方法，這兩種方法的相似性分數分布較為分散，分布範圍在0.0到1.0之間，0.5

以上的配對分別佔全部的61.71%與76.11%。JAC所測量得到的主題相似性分數如果為1，配對主題彼此之間具有完全相同的關鍵詞語，RBO的主題相似性分數如果為1，配對主題除了具有完全相同的關鍵詞語外，關鍵詞語的順序也相同。在本研究中，JAC和RBO分別有566對（11.92%）與54對（1.14%）主題的相似性分數為1。在六種方法中，以這兩種方法產生最多相似性分數為0的情形。JAC和RBO分別有230對（4.84%）與235對（4.95%）的相似性分數為0。這個情形表示有些配對的主題有完全不同的關鍵詞語。

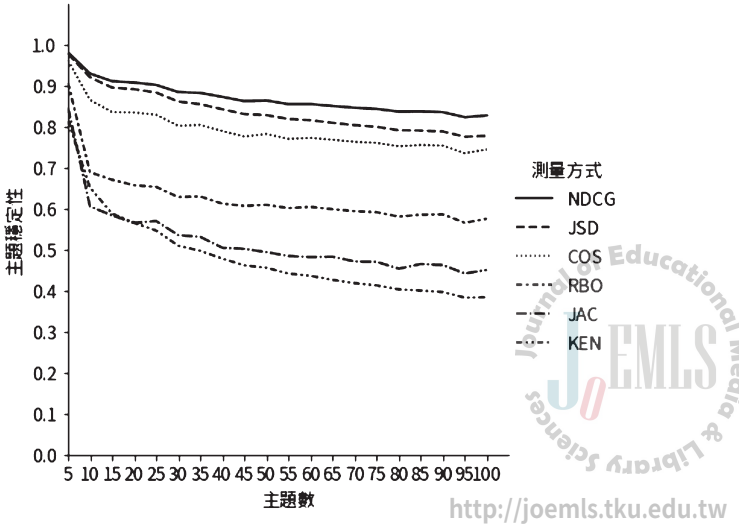
3. KEN

KEN的主題配對相似性分數範圍則在0.0到0.8之間，大部分分布在0.5到0.8之間，佔全部的73.56%，且其中最高只到0.76。在本研究中，相較於其他相似性估計方法，KEN根據某一個數列上的資料項目與其他項目的順序關係在另一個數列上能否維持的差異估計主題相關性，需要考慮詞典中每對詞語在配對的兩個主題上具有一致的出現機率順序。但每個主題上仍包含許多不相關且機率相當小的詞語，可能造成相似的主題卻無法有完全相同的順序，影響KEN的計算，因此所得到的相似性分數明顯地較其他方法為低。

㈢主題數目對於主題建模穩定性的影響

本研究以不同的主題數目（$K = 5 \sim 100$）各建立20個主題模型，然後測量主題建模在各種主題數目下的穩定性，圖4上的折線從上到下分別是在不同主題數目下運用NDCG、JSD、COS、RBO、JAC和KEN等方法測量得到的穩定性。在圖4上可觀察到不論何種估計方法，隨著主題數目增加，穩定性都有明顯下降的情形。並且在各種主題數目之中，以五個主題的主題模型最為穩定，

圖4　不同主題數目對於主題建模穩定性的影響



且其程度遠較其他數目的模型高出相當多。其原因可能是由於主題數目增加，使得模型中的主題概念範圍縮小，從而造成在主題中的詞語、順序和出現機率容易有變化的情形，可能讓有愈多的主題無法在另一個模型中比對到相當相似的主題，進而使得測量到的穩定性變差。

五、結　論

隨著LDA主題建模在文本分析的應用愈來愈廣泛，主題建模的穩定性測量也愈受到重視。在De Waal與Barnard（2008）、Greene等（2014）與Belford等（2018）使用的穩定性測量架構中，主題之間的相似性分數估計方法是測量主題建模穩定性的基礎，並且產生最佳主題配對組合的「主題對齊」是這個程序的關鍵步驟，然而過去的研究較少比較不同相似性估計方法對主題建模穩定性的影響，也缺乏針對主題對齊的結果進行探討。本研究採用PTT BBS書板約30,000筆發文做為分析的文本集合，並應用JSD、NDCG、COS、JAC、RBO和KEN等相似性估計方法，比較不同方法經由主題對齊之後產生配對結果相同的比例，並觀察各種相似性估計方法在配對主題上的相似性分數分布。最後並探討主題數目對於主題建模穩定性的影響。研究結果有以下的發現：

㈠本研究提出以具有相同配對結果的比例來比較不同的相似性估計方法在測量主題建模穩定性的效果，並發現本研究所探討的六種相似性估計方法配對結果相同的情形比例相當高。因此，在穩定性測量的應用上，例如本研究進行的主題數目對於穩定性的影響，各種方法大致上都有相同的效果。但本研究也發現方法上運用詞典中所有的詞語，或只利用少數的關鍵詞語仍會輕微影響配對結果是否相同。

㈡本研究觀察六種相似性分數估計方法，在經由主題對齊演算法產生主題配對組合上的相似性分數分布，目前主題建模的穩定性測量研究尚未有關於這方面的探討。在六種方法中，運用所有詞語在主題上的出現機率做為主題特徵的JSD、COS和NDCG等三種方法可以明顯地觀察到大部分配對有相當高的相似性分數。換言之，未來將可運用這三種方法搭配匈牙利演算法進行主題對齊，然後以較高的相似性分數選取出兩個模型中相似的主題。JAC和RBO兩種方法僅使用少數出現機率較大的關鍵詞語做為主題特徵，使得相似性分數的分布較分散，較難透過觀察分數決定主題是否配對到另一個模型上最相似的主題。但是利用JAC和RBO方法可發現配對中關鍵詞語完全相同或完全不同的主題。KEN的估計方式是根據每一對詞語在配對的兩個主題上是否維持一致順序，然而絕大多數詞語與配對的兩個主題並不相關。

㈢本研究發現，主題數目對於穩定性有很大的影響，使用不同的相似性估計方法都可觀察到，主題數目愈大時主題建模愈不穩定的現象。Greene等

（2014）認為，主題建模時較多的主題將造成較小的主題範圍，使得每次建模產生的主題多不相同，容易造成建模時的不穩定，本研究的結果與他們的推論相符合。然而本研究是針對LDA主題建模的穩定性進行探討，有別於Greene等（2014）針對NMF主題建模的研究。此外，Greene等（2014）的實驗假定文件僅有一個主題，且整個文本集合內的主題數目並不多；本研究則以較實際的主題建模應用為考量，假定分析的文件中可能包含多個主題，且考慮較大範圍的主題數目對主題建模穩定性的影響，較符合實際情況。

根據上述的研究結果，我們建議以下課題做為未來研究的方向：

㈠由於研究時程與篇幅的限制，本研究在進行主題建模的穩定性測量時，將模型個數固定為20，並在使用JAC和RBO兩種方法僅使用10個關鍵詞語做為主題特徵，未來可探討不同模型個數與關鍵詞語數目對穩定性測量的影響。

㈡目前在測量主題建模的穩定性時，大多根據估計的相似性分數計算模型之間的一致性，未來也許可以參考Maier等（2018）利用可能正確配對的主題數量佔比，發展適合直接解讀模型品質的測量方式，並找出各個模型中比較穩定的主題。例如整合不同方法的特性，先運用JSD、COS或NDCG等方法搭配匈牙利演算法進行主題對齊，確認兩個模型中較相似的主題，然後再利用JAC或RBO等方法選取模型中較穩定的主題或排除不穩定的主題。

㈢在運用主題建模進行文本分析時，主題數目是一個相當重要的輸入參數，主題數目決定了模型上主題彼此之間的差異與可解釋性（interpretability），主題數愈多，產生的主題具有愈加狹隘（narrow），而特定（specific）的意義，導致多個不同的主題可能具有相似的概念；反之，主題數目太少，將使得主題的意義廣泛，理應區分的概念被包含同一主題內（Maier et al., 2018）。過去已有相當多研究利用複雜度或主題協調性決定最佳的主題數目，甚至藉由人力檢視（Maier et al., 2018），Greene等（2014）提出利用穩定性發現最佳主題數目的概念。本研究則建議未來可嘗試運用與整合各種主題模型品質指標決定最佳的主題數目。

㈣最後，也是最重要的，在累積更多主題建模穩定性測量的經驗，對這項主題模型品質有較深入的瞭解後，可進一步嘗試發展提升主題建模穩定性的方法，使得相同輸入條件下每次產生模型上的主題盡可能相似，讓文本分析的結果具有高信度。目前已有一些有關這方面的研究，例如前述的Chuang等（2015）、Lancichinetti等（2015）、Koltcov等（2016）、Agrawal等（2018）、Maier等（2018）和Mantyla等（2018）。

參考文獻

Agrawal, A., Fu, W., & Menzies, T. (2018). What is wrong with topic modeling? And how to fix it using search-based software engineering. *Information and Software Technology*, *98*, 74-88. https://doi.org/10.1016/j.infsof.2018.02.005

Ballester, O., & Penner, O. (2022). Robustness, replicability and scalability in topic modelling. *Journal of Informetrics*, *16*(1), 101224. https://doi.org/10.1016/j.joi.2021.101224

Belford, M., Mac Namee, B., & Greene, D. (2018). Stability of topic modeling via matrix factorization. *Expert Systems with Applications: An International Journal*, *91*, 159-169. https://doi.org/10.1016/j.eswa.2017.08.047

Blei, D. M., & Lafferty, J. D. (2006). Dynamic topic models. In W. W. Cohen & A. Moore (Eds.), *Proceedings of the 23rd international conference on machine learning* (pp. 113-120). ACM. https://doi.org/10.1145/1143844.1143859

Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of science. *The Annals of Applied Statistics*, *1*(1), 17-35. https://doi.org/10.1214/07-AOAS114

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993-1022.

Chuang, J., Roberts, M. E., Stewart, B. M., Weiss, R., Tingley, D., Grimmer, J., & Heer, J. (2015). TopicCheck: Interactive alignment for assessing topic model stability. In R. Mihalcea, J. Chai, & A. Sarkar (Eds.), *Proceedings of the 2015 conference of the North American Chapter of the Association for Computational Linguistics: Human language technologies* (pp. 175-184). Association for Computational Linguistics. https://doi.org/10.3115/v1/N15-1018

De Waal, A., & Barnard, E. (2008). Evaluating topic models with stability. In F. Nicolls (Ed.), *Proceedings of the 19th annual symposium of the Pattern Recognition Association of South Africa* (pp. 79-84). Pattern Recognition Association of South Africa.

Elgesem, D., Steskal, L., & Diakopoulos, N. (2015). Structure and content of the discourse on climate change in the blogosphere: The big picture. *Environmental Communication*, *9*(2), 169-188. https://doi.org/10.1080/17524032.2014.983536

Greene, D., O'Callaghan, D., & Cunningham, P. (2014). How many topics? Stability analysis for topic models. In T. Calders, F. Esposito, E. Hüllermeier, & R. Meo (Eds.), *ECML PKDD 2014: Machine learning and knowledge discovery in databases* (pp. 498-513). Springer. https://doi.org/10.1007/978-3-662-44848-9_32

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, *101*(Suppl. 1), 5228-5235. https://doi.org/10.1073/pnas.0307752101

Hofmann, T. (1999). Probabilistic latent semantic analysis. In K. B. Laskey & H. Prade (Eds.), *UAI'99: Proceedings of the fifteenth conference on uncertainty in artificial intelligence* (pp. 289-296). Morgan Kaufmann.

Jacobi, C., Van Atteveldt, W., & Welbers, K. (2016). Quantitative analysis of large amounts of journalistic texts using topic modelling. *Digital Journalism*, *4*(1), 89-106. https://doi.org/10.1080/21670811.2015.1093271

Järvelin, K., & Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, *20*(4), 422-446. https://doi.org/10.1145/582415.582418

Kim, D., & Oh, A. (2011). Topic chains for understanding a news corpus. In A. Gelbukh (Ed.), *CICLing 2011: Computational linguistics and intelligent text processing* (pp. 163-176). Springer. https://doi.org/10.1007/978-3-642-19437-5_13

Koltcov, S., Nikolenko, S. I., Koltsova, O., Filippov, V., & Bodrunova. S. (2016). Stable topic modeling with local density regularization. In F. Bagnoli et al. (Eds.), *INSCI 2016: Internet science* (pp. 176-188) Springer. https://doi.org/10.1007/978-3-319-45982-0_16

Kuhn, H. W. (1955). The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, *2*(1-2), 83-97. https://doi.org/10.1002/nav.3800020109

Lancichinetti, A., Sirer, M. I., Wang, J. X., Acuna, D., Körding, K., & Amaral, L. A. N. (2015). High-reproducibility and high-accuracy method for automated topic classification. *Physical Review X*, *5*(1), 011007. https://doi.org/10.1103/PhysRevX.5.011007

Li, P.-H., Fu, T.-J., & Ma, W.-Y. (2020). Why attention? Analyze BiLSTM deficiency and its remedies in the case of NER. *AAAI Conference on Artificial Intelligence*, *34*(05), 8236-8244. https://doi.org/10.1609/aaai.v34i05.6338

Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., Pfetsch, B., Heyer, G., Reber, U., Häussler, T., Schmid-Petri, H., & Adam, S. (2018). Applying LDA topic modeling in communication research: Toward a valid and reliable methodology. *Communication Methods and Measures*, *12*(2-3), 93-118. https://doi.org/10.1080/19312458.2018.1430754

Mantyla, M. V., Claes, M. & Farooq, U. (2018). Measuring LDA topic stability from clusters of replicated runs. In M. Oivo (Chair), *ESEM'18: Proceedings of the 12th ACM/IEEE international symposium on empirical software engineering and measurement* (Article No. 4). ACM. https://doi.org/10.1145/3239235.3267435

Nichols, L. G. (2014). A topic model approach to measuring interdisciplinarity at the National Science Foundation. *Scientometrics*, *100*(3), 741-754. https://doi.org/10.1007/s11192-014-1319-2

Niekler, A., & Jähnichen, P. (2012). Matching results of latent Dirichlet allocation for text. In N. Rußwinkel, U. Drewitz, & H. Van Rijn (Eds.), *Proceedings of the 11th international conference on cognitive modeling* (pp. 317-322). Universitätsverlag der TU.

Panichella, A., Dit, B., Oliveto, R., Di Penta, M., Poshynanyk, D., & De Lucia, A. (2013). How to effectively use topic models for software engineering tasks? An approach based on genetic algorithms. In D. Notkin, B. H. C. Cheng, & K. Pohl (Eds.), *Proceedings of the 35th international conference on software engineering: ICSE 2013* (pp. 522-531). IEEE. https://doi.org/10.1109/ICSE.2013.6606598

Quinn, K. M., Monroe, B. L., Colaresi, M., Crespin, M. H., & Radev, D. R. (2010). How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science*, *54*(1), 209-228. https://doi.org/10.1111/j.1540-5907.2009.00427.x

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. In X. Cheng & H. Li (Chairs), *WSDM '15: Proceedings of the eighth ACM international conference on web search and data mining* (pp. 399-408). ACM. https://doi.org/10.1145/2684822.2685324

Strehl, A., & Ghosh, J. (2002). Cluster ensembles - A knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, *3*, 583-617.

Sun, X., Liu, X., Li, B., Duan, Y., Yang, H., & Hu, J. (2016). Exploring topic models in software engineering data analysis: A survey. In Y. Chen (Ed.), *Proceedings of the 17th IEEE/ACIS*

international conference on software engineering, artificial intelligence, networking and parallel/distributed computing (SNPD) (pp. 357-362). IEEE. https://doi.org/10.1109/SNPD.2016.7515925

Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, *101*(476), 1566-1581.

Tirunillai, S., & Tellis, G. J. (2014). Mining marketing meaning from online chatter: Strategic brand analysis of big data using latent Dirichlet allocation. *Journal of Marketing Research*, *51*(4), 463-479. https://doi.org/10.1509/jmr.12.0106

Toubia, O., Iyengar, G., Bunnell, R., & Lemaire, A. (2019). Extracting features of entertainment products: A guided latent Dirichlet allocation approach informed by the psychology of media consumption. *Journal of Marketing Research*, *56*(1), 18-36. https://doi.org/10.1177/0022243718820559

Wang, Q., Cao, Z., Xu, J., & Li, H. (2012). Group matrix factorization for scalable topic modeling. In W. Hersh (Chair), *SIGIR '12: Proceedings of the 35th international ACM SIGIR conference on research and development in information retrieval* (pp. 375-384). ACM. https://doi.org/10.1145/2348283.2348335

Webber, W., Moffat, A., & Zobel, J. (2010). A similarity measure for indefinite rankings. *ACM Transactions on Information Systems*, *28*(4), Article 20. https://doi.org/10.1145/1852102.1852106

Yang, Y., Pan, S., Song, Y., Lu, J., & Topkara, M. (2016). Improving topic model stability for effective document exploration. In G. Brewka (Ed.), *IJCAI'16: Proceedings of the twenty-fifth international joint conference on artificial intelligence* (pp. 4223-4227). AAAI.

附錄：匈牙利演算法及其在主題對齊的應用

以下首先說明匈牙利演算法的輸入、目的與步驟，接著以一個範例說明本研究利用匈牙利演算法進行主題對齊的方法與過程。

匈牙利演算法的輸入為一個行與列的數目都為K的矩陣。矩陣上的行與列分別代表要進行配對的兩組項目，矩陣上的元素則表示配對項目之間的差距。以第i行第j列上的元素為例，上面的值表示第一組的第i個項目與第二組的第j個項目之間的差距。匈牙利演算法的步驟如下：

Step 1：針對矩陣上的每一行，減去這行元素中的最小值。

Step 2：針對目前矩陣上的每一列，減去這列元素中的最小值。

Step 3：嘗試用最少的垂直線與水平線通過矩陣上所有為 0的元素。如果線的數目少於行數K，進行Step 4。否則，直接進行Step 5。

Step 4：找出整個矩陣中不是0的元素中的最小值m，將所有不是0的元素減去m。並且找出兩條線交叉上為0的元素，取代為m。然後，返回Step 3。

Step 5：選擇一個行與列的配對組合，使得每一行或每一列都只有一個0被選上。

本研究在計算兩個模型的主題以及每一對主題之間的相似性分數之後，將應用匈牙利演算法進行主題對齊。以附圖1(a)上的矩陣為例，表示兩個主題數目為5的模型上每一對主題之間的相似性分數，並且這些相似性分數的值在0到1之間。由於匈牙利演算法是計算兩組項目之間上最小差距總和的配對，但本研究希望取得兩個模型上相似性分數總和最大的主題配對，所以首先以1減去相似性矩陣上每個元素的值，轉換為差距矩陣，如附圖1(b)所示。假定第1個模型的第1個主題與第2個模型的第1個主題之間的相似性分數為0.29，這個值放在附圖1(a)相似性矩陣的第1行第1列上，當轉換為附圖(b)的差距矩陣時為1 – 0.29 = 0.71。

接著進行演算法的Step 1，找出每一行最小的元素，然後減去這個元素的值。例如第1行的元素為[0.71, 0.93, 0.08, 0.24, 0.85]，其中以0.08為最小的元素值，因此將這行上所有的元素減去這個值，結果為[0.63, 0.85, 0, 0.16, 0.77]。其他各行也是經過如此運算，結果為附圖1(c)上的矩陣。然後進行Step 2，將每列的元素減去該列的最小元素值，結果如附圖1(d)上的矩陣。

在演算法的Step 3，利用最少的線通過目前矩陣上所有出現0的元素。先選取上面有最多0的行或列開始，以線通過這個行或列。處理完後，如果矩陣上還有0尚未被通過，再選取目前上面有最多0的行或列，以線通過。重複進行上面的處理過程，一直到矩陣上所有的0都有線通過為止。以附圖1(d)上的矩陣為例，先選擇具有3個0的第3列，以線覆蓋此列。然後，再依序利用線通過第3行、第5行以及第1列和第5列。結果如附圖1(e)所示。目前矩陣上共有五條線，與行和列的數目相等，因此接著進行Step 5。

Step 5先選取只有單獨一個0的行，例如附圖1(f)上的第1、2和4等行。將這些行上的0所在列上其餘的0進行標記，例如根據第1行上的0，也就是(1, 3)（表示第1行、第3列，以下的表示方式與此相同）位置的0，標記同樣在第3列上其餘的0，包括(3, 3)和(5, 3)等位置上的0，附圖1(f)呈現這個處理的示意圖。如果有沒有被選取的行，便再

次選取除了已經標記的0之外，只有單獨一個0的行，並且將這些行上的0所在列上其餘的0進行標記，以上面的例子為第3行與第5行。反覆進行上面的處理過程，一直到所有的行被選取為止。

最後，矩陣上沒有被標記0所在的位置便是最佳的配對組合。在這個例子中，最佳配對組合包括(1, 3)、(2, 5)、(3, 2)、(4, 1)和(5, 4)等，如附圖1(g)所示。在這種配對情形下主題之間相似性分數的總和比其他任何配對組合的分數總和大，其總和為0.92 + 0.85 + 0.66 + 0.97 + 0.72 = 4.12。

附圖1　應用匈牙利演算法進行主題對齊之說明

	1	2	3	4	5
1	0.29	0.07	0.92	0.76	0.15
2	0.44	0.17	0.81	0.71	0.85
3	0.73	0.66	0.79	0.01	0.04
4	0.97	0.11	0.16	0.22	0.31
5	0.08	0.26	0.83	0.72	0.55

(a) 相似性矩陣

	1	2	3	4	5
1	0.71	0.93	0.08	0.24	0.85
2	0.56	0.83	0.19	0.29	0.15
3	0.27	0.34	0.21	0.99	0.96
4	0.03	0.89	0.84	0.78	0.69
5	0.92	0.74	0.17	0.28	0.45

(b) 轉成為差距矩陣

	1	2	3	4	5
1	0.63	0.85	0	0.16	0.77
2	0.41	0.68	0.04	0.14	0
3	0.06	0.13	0	0.78	0.75
4	0	0.86	0.81	0.75	0.66
5	0.75	0.57	0	0.11	0.28

(c) 減去每行的最小值

	1	2	3	4	5
1	0.63	0.72	0	0.05	0.77
2	0.41	0.55	0.04	0.03	0
3	0.06	0	0	0.67	0.75
4	0	0.73	0.81	0.64	0.66
5	0.75	0.44	0	0	0.28

(d) 減去每列的最小值

	1	2	3	4	5
1	0.63	0.72	0	0.05	0.77
2	0.41	0.55	0.04	0.03	0
3	0.06	0	0	0.67	0.75
4	0	0.73	0.81	0.64	0.66
5	0.75	0.44	0	0	0.28

(e) 以最少的線通過所有的0

	1	2	3	4	5
1	0.63	0.72	0	0.05	0.77
2	0.41	0.55	0.04	0.03	0
3	0.06	0	0*	0.67	0.75
4	0	0.73	0.81	0.64	0.66
5	0.75	0.44	0*	0	0.28

(f) 選取單獨只有一個0的行（如圖上的第1行），並標記這些行上的0（如第1行第3列上的0）所在列上其餘的0（如第3列上的0*）

	1	2	3	4	5
1	0.63	0.72	0	0.05	0.77
2	0.41	0.55	0.04	0.03	0
3	0.06	0	0	0.67	0.75
4	0	0.73	0.81	0.64	0.66
5	0.75	0.44	0	0	0.28

(g) 找出配對的項目

Journal of Educational Media & Library Sciences 59 : 2 (2022) : 201-231
DOI:10.6120/JoEMLS.202207_59(2).0014.RS.AM

Research Article

Estimation of Topic Similarity and Its Application to Measuring Stability of Topic Modeling

Sung-Chien Lin

Abstract

Topic modeling stability is a measurement of the extent to which models produced by the same modeling approach for the same corpus and with the same initial conditions have similar topics. Since the method used for calculating similarity between topics is considered the basis for measuring topic modeling stability and topic alignment is a key step in the measurement, the present study first calculated the proportion of identical paired topics among the optimal combinations of paired topics generated using different topic similarity calculation methods, and then observed the distribution of similarity scores of paired topics for each method. Finally, this study performed an analysis of the effects of the number of topics on topic modeling stability. The topic modeling method used in this study is commonly used LDA topic modeling, and the corpus used to establish topic models including about 30,000 posts was collected from the PTT Bulletin Board System (BBS) Book message board. The results indicated that there is a high proportion of identical paired topics among the different methods of measuring similarity, although the similarity scores of paired topics for each method had different distributions due to the different kinds and amounts of information of word distribution in each topic they used. The results also revealed that with the increase of the number of topics, the stability noticeably decreased.

Keywords: *Topic modeling, latent Dirichlet allocation (LDA), Stability measurement, Topic similarity estimation, Topic alignment*

SUMMARY

Introduction

Topic modeling can reveal topic structures contained in a corpus and aid in the rapid and effective analyses of large amounts of text. Currently, latent Dirichlet allocation (LDA; Blei et al., 2003) is regarded as the most popular topic

Assistant Professor, Department of Information and Communications, ShihHsin University, Taipei, Taiwan
E-mail: scl@mail.shu.edu.tw

Please visit JoEMLS website to read the Peer Review Report (Open Point) and Article Summary (InSight Point) of the article.
http://joemls.ntu.edu.tw
2022/05/02 received; 2022/07/29 revised; 2022/07/30 accepted

modeling technique among researchers and is widely used for problems involving text analyses (Lancichinetti et al., 2015). However, in practice, even with the same parameters and corpus, the models produced with this technique somewhat differ from each other, calling into question the reliability of the analysis results (Maier et al., 2018). This problem casts doubt on the usefulness of LDA topic modeling (Belford et al., 2018; Chuang et al., 2015).

Topic modeling stability is a measurement of the extent to which models produced by the same modeling approach for the same corpus and with the same initial conditions have similar topics. Several methods can be used to measure topic modeling stability. For instance, in the present study, the framework used for measuring topic modeling stability (De Waal & Barnard, 2008; Greene et al., 2014) involved producing multiple topic models through repeated modeling with the same corpus and number of topics and then performing topic alignment between any two topic models by using the Hungarian algorithm to determine the optimal combination of topic pairs. In this combination, the mean similarity of the topic pairs was the agreement score of the two models, whereas the mean of the agreement scores was the measurement of the topic modeling stability.

According to this measurement framework, the method used for calculating similarity between topics is considered the basis for measuring topic modeling stability. Belford et al. (2018) and Greene et al. (2014) used Jaccard's score (JAC) to calculate topic similarity; however, their approach considered only a small portion of information in the word distribution of each topic. Therefore, in the present study, the following six methods for measuring topic similarity were used and compared: Jensen–Shannon divergence (JSD), normalized discounted cumulative gain (NDCG), cosine measure (COS), JAC, rank-biased overlap score (RBO), and Kendall's τ coefficient (KEN). Topic alignment is a key step in this measurement framework. If two different methods for measuring topic similarity yield highly similar optimal combinations of topic pairs, the two methods may have similar stability measurement outcomes. The distribution of the similarity score of paired topics can also indicate which methods are more likely to identify the topics that appear in most models after topic alignment.

This study performed the following analysis tasks:

Task 1: Conduct an analysis of the proportion of identical paired topics among the optimal combinations of paired topics generated using different topic similarity calculation methods.

Task 2: Perform an analysis of the distribution of similarity scores of paired topics for each method.

Overall, the study conducted by Greene et al. (2014) is regarded as one of the few studies analyzing the effects of the number of topics on topic modeling

stability. However, the corpus used in that study had few topics, which were already clearly defined. Therefore, a corpus with a greater number of topics was used in the present study.

Task 3: Perform an analysis of the effects of the number of topics on topic modeling stability.

Research Methods

Word segmentation, part-of-speech tagging, and stop word removal were performed on 32,895 posts collected from the PTT Bulletin Board System (BBS) Book message board. Posts containing at least five words were selected to form a corpus for analyzing topic modeling stability. The final corpus included 20,287 posts and 1,579,116 words. The topic modeling inputs consisted of this corpus and a dictionary. For each different number of topics (K = 5, 10, 15, …, 100), a total of 20 models were created with fixed prior parameters α and β.

Next, the six methods of measurement mentioned earlier were used with any two topic models to calculate the similarity between each topic pair. The results of each method were then adjusted to be between 0 and 1. The greater the similarity between any two topics was, the greater the score was. The similarity scores of all pairs of topics between every two models were then entered into the Hungarian algorithm to align the topics and obtain an optimal combination of topic pairs. Analysis Tasks 1 and 2 were then performed.

Finally, the agreement score between every two topic models was obtained by averaging the optimal topic pair similarity scores. Analysis Task 3 was then performed using the mean agreement score between each pair of topic models as the stability measurement.

Research Result

Task 1

This task involved assessing whether different methods of measuring similarity had the same effect when measuring stability based on the proportion of identical topic pairs in the optimal combinations of topic pairs. The results obtained indicated a high proportion of identical paired topics among the different methods of calculating topic similarity. The proportion of identical paired topics among the six methods reached 76.99%, and the total proportion even increased to 94.09% in four or more methods. For any two methods, the proportion of identical topic pairs was 84% or higher, suggesting that any two methods had similar stability outcomes. However, slight differences were observed between the methods that involved the use of all word distribution data, such as JSD, NDCG, COS, and KEN, and the methods that involved only a few keywords, such as JAC and RBO.

Task 2

If a method for calculating topic similarity can yield a high similarity score between postalignment paired topics, then this means that this method can differentiate between similar topics within different models and thereby identify stable topics in each model. In this study, rather high similarity scores were observed among most of the paired topics when JSD, NDCG, and COS were used, which are methods that involve the use of the occurrence probability of all words in each topic, showing that these methods can easily identify stable topics in models. JAC and RBO are methods that involve the use of a set of keywords to represent topics. In this study, these two methods yielded similarity scores that were scattered across a wide range. In addition, approximately 5% of the similarity scores were 0, because the corresponding paired topics had completely different keywords. The KEN method considers every word to have a consistent order of occurrences among paired topics. However, each topic contains several irrelevant and low-probability words, which may cause similar topics to exhibit dissimilar orders and hence lower the similarity scores.

Task 3

This task entailed measuring the stability of topic models with different numbers of topics. The results revealed that with the increase of the number of topics, the stability noticeably decreased. This may be because with the increase of the number of topics, the topic ranges in the model became narrower, and the distribution of words in the topic became more prone to change. This may have resulted in an increasing number of topics being unable to align with similar topics in another model, thereby lowering the stability.

Suggestions and Future Research

In this study, topic alignment was performed using the Hungarian algorithm, and the agreement score between models was calculated on the basis of the similarity scores between paired topics. Future researchers may refer to Maier et al. (2018) and use the proportion of possible pairs as an indicator of model stability to develop a method of measurement that is suited to direct interpretation.

During text analyses with topic modeling, the number of topics is considered a key parameter that determines the scope, accuracy, and interpretability of the model. Several studies have employed perplexity or topic coherence as an indicator of topic model quality to determine the optimal number of topics, and some have even involved manual reviews (Maier et al., 2018). Therefore, we suggest integrating stability with other quality indicators to determine the optimal number of topics.

Finally and most importantly, the methods used to improve topic modeling stability should be further developed. Increasing the level of stability can help increase the possible similarity in topics among all models produced under the same input conditions and thereby enhance the reliability of the text analysis results. Among the current studies investigating this topic are those of Chuang et al. (2015), Lancichinetti et al. (2015), Koltcov et al. (2016), Agrawal et al. (2018), Maier et al. (2018), and Mantyla et al. (2018).

ROMANIZED & TRANSLATED REFERENCES FOR ORIGINAL TEXT

Agrawal, A., Fu, W., & Menzies, T. (2018). What is wrong with topic modeling? And how to fix it using search-based software engineering. *Information and Software Technology*, *98*, 74-88. https://doi.org/10.1016/j.infsof.2018.02.005

Ballester, O., & Penner, O. (2022). Robustness, replicability and scalability in topic modelling. *Journal of Informetrics*, *16*(1), 101224. https://doi.org/10.1016/j.joi.2021.101224

Belford, M., Mac Namee, B., & Greene, D. (2018). Stability of topic modeling via matrix factorization. *Expert Systems with Applications: An International Journal*, *91*, 159-169. https://doi.org/10.1016/j.eswa.2017.08.047

Blei, D. M., & Lafferty, J. D. (2006). Dynamic topic models. In W. W. Cohen & A. Moore (Eds.), *Proceedings of the 23rd international conference on machine learning* (pp. 113-120). ACM. https://doi.org/10.1145/1143844.1143859

Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of science. *The Annals of Applied Statistics*, *1*(1), 17-35. https://doi.org/10.1214/07-AOAS114

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993-1022.

Chuang, J., Roberts, M. E., Stewart, B. M., Weiss, R., Tingley, D., Grimmer, J., & Heer, J. (2015). TopicCheck: Interactive alignment for assessing topic model stability. In R. Mihalcea, J. Chai, & A. Sarkar (Eds.), *Proceedings of the 2015 conference of the North American Chapter of the Association for Computational Linguistics: Human language technologies* (pp. 175-184). Association for Computational Linguistics. https://doi.org/10.3115/v1/N15-1018

De Waal, A., & Barnard, E. (2008). Evaluating topic models with stability. In F. Nicolls (Ed.), *Proceedings of the 19th annual symposium of the Pattern Recognition Association of South Africa* (pp. 79-84). Pattern Recognition Association of South Africa.

Elgesem, D., Steskal, L., & Diakopoulos, N. (2015). Structure and content of the discourse on climate change in the blogosphere: The big picture. *Environmental Communication*, *9*(2), 169-188. https://doi.org/10.1080/17524032.2014.983536

Greene, D., O'Callaghan, D., & Cunningham, P. (2014). How many topics? Stability analysis for topic models. In T. Calders, F. Esposito, E. Hüllermeier, & R. Meo (Eds.), *ECML PKDD 2014: Machine learning and knowledge discovery in databases* (pp. 498-513). Springer. https://doi.org/10.1007/978-3-662-44848-9_32

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, *101*(Suppl. 1), 5228-5235. https://doi.org/10.1073/pnas.0307752101

Hofmann, T. (1999). Probabilistic latent semantic analysis. In K. B. Laskey & H. Prade (Eds.), *UAI'99: Proceedings of the fifteenth conference on Uncertainty in artificial intelligence* (pp. 289-296). Morgan Kaufmann.

Jacobi, C., Van Atteveldt, W., & Welbers, K. (2016). Quantitative analysis of large amounts of journalistic texts using topic modelling. *Digital Journalism*, *4*(1), 89-106. https://doi.org/10.1080/21670811.2015.1093271

Järvelin, K., & Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, *20*(4), 422-446. https://doi.org/10.1145/582415.582418

Kim, D., & Oh, A. (2011). Topic chains for understanding a news corpus. In A. Gelbukh (Ed.), *CICLing 2011: Computational linguistics and intelligent text processing* (pp. 163-176). Springer. https://doi.org/10.1007/978-3-642-19437-5_13

Koltcov, S., Nikolenko, S. I., Koltsova, O., Filippov, V., & Bodrunova. S. (2016). Stable topic modeling with local density regularization. In F. Bagnoli et al. (Eds.), *INSCI 2016: Internet science* (pp. 176-188) Springer. https://doi.org/10.1007/978-3-319-45982-0_16

Kuhn, H. W. (1955). The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, *2*(1-2), 83-97. https://doi.org/10.1002/nav.3800020109

Lancichinetti, A., Sirer, M. I., Wang, J. X., Acuna, D., Körding, K., & Amaral, L. A. N. (2015). High-reproducibility and high-accuracy method for automated topic classification. *Physical Review X*, *5*(1), 011007. https://doi.org/10.1103/PhysRevX.5.011007

Li, P.-H., Fu, T.-J., & Ma, W.-Y. (2020). Why attention? Analyze BiLSTM deficiency and its remedies in the case of NER. *AAAI Conference on Artificial Intelligence*, *34*(05), 8236-8244. https://doi.org/10.1609/aaai.v34i05.6338

Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., Pfetsch, B., Heyer, G., Reber, U., Häussler, T., Schmid-Petri, H., & Adam, S. (2018). Applying LDA topic modeling in communication research: Toward a valid and reliable methodology. *Communication Methods and Measures*, *12*(2-3), 93-118. https://doi.org/10.1080/19312458.2018.1430754

Mantyla, M. V., Claes, M. & Farooq, U. (2018). Measuring LDA topic stability from clusters of replicated runs. In M. Oivo (Chair), *ESEM'18: Proceedings of the 12th ACM/IEEE international symposium on empirical software engineering and measurement* (Article No. 4). ACM. https://doi.org/10.1145/3239235.3267435

Nichols, L. G. (2014). A topic model approach to measuring interdisciplinarity at the National Science Foundation. *Scientometrics*, *100*(3), 741-754. https://doi.org/10.1007/s11192-014-1319-2

Niekler, A., & Jähnichen, P. (2012). Matching results of latent Dirichlet allocation for text. In N. Rußwinkel, U. Drewitz, & H. Van Rijn (Eds.), *Proceedings of the 11th international conference on cognitive modeling* (pp. 317-322). Universitätsverlag der TU.

Panichella, A., Dit, B., Oliveto, R., Di Penta, M., Poshynanyk, D., & De Lucia, A. (2013). How to effectively use topic models for software engineering tasks? An approach based on genetic algorithms. In D. Notkin, B. H. C. Cheng, & K. Pohl (Eds.), *Proceedings of the 35th international conference on software engineering: ICSE 2013* (pp. 522-531). IEEE. https://doi.org/10.1109/ICSE.2013.6606598

Quinn, K. M., Monroe, B. L., Colaresi, M., Crespin, M. H., & Radev, D. R. (2010). How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science*, *54*(1), 209-228. https://doi.org/10.1111/j.1540-5907.2009.00427.x

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. In X. Cheng & H. Li (Chairs), *WSDM '15: Proceedings of the eighth ACM international conference on web search and data mining* (pp. 399-408). ACM. https://doi.org/10.1145/2684822.2685324

Strehl, A., & Ghosh, J. (2002). Cluster ensembles - A knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, *3*, 583-617.

Sun, X., Liu, X., Li, B., Duan, Y., Yang, H., & Hu, J. (2016). Exploring topic models in software engineering data analysis: A survey. In Y. Chen (Ed.), *Proceedings of the 17th IEEE/ACIS international conference on software engineering, artificial intelligence, networking and parallel/distributed computing (SNPD)* (pp. 357-362). IEEE. https://doi.org/10.1109/SNPD.2016.7515925

Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, *101*(476), 1566-1581.

Tirunillai, S., & Tellis, G. J. (2014). Mining marketing meaning from online chatter: Strategic brand analysis of big data using latent Dirichlet allocation. *Journal of Marketing Research*, *51*(4), 463-479. https://doi.org/10.1509/jmr.12.0106

Toubia, O., Iyengar, G., Bunnell, R., & Lemaire, A. (2019). Extracting features of entertainment products: A guided latent Dirichlet allocation approach informed by the psychology of media consumption. *Journal of Marketing Research*, *56*(1), 18-36. https://doi.org/10.1177/0022243718820559

Wang, Q., Cao, Z., Xu, J., & Li, H. (2012). Group matrix factorization for scalable topic modeling. In W. Hersh (Chair), *SIGIR '12: Proceedings of the 35th international ACM SIGIR conference on research and development in information retrieval* (pp. 375-384). ACM. https://doi.org/10.1145/2348283.2348335

Webber, W., Moffat, A., & Zobel, J. (2010). A similarity measure for indefinite rankings. *ACM Transactions on Information Systems*, *28*(4), Article 20. https://doi.org/10.1145/1852102.1852106

Yang, Y., Pan, S., Song, Y., Lu, J., & Topkara, M. (2016). Improving topic model stability for effective document exploration. In G. Brewka (Ed.), *IJCAI'16: Proceedings of the twenty-fifth international joint conference on artificial intelligence* (pp. 4223-4227). AAAI.

http://joemls.tku.edu.tw