201917603

(19)中華民國智慧財產局

(12)發明說明書公開本 (11)公開編號：TW 201917603 A

(43)公開日： 中華民國 108 (2019) 年 05 月 01 日

(21)申請案號：107129401 (22)申請日： 中華民國 107 (2018) 年 08 月 23 日

(51)Int. Cl.： *G06F17/27 (2006.01)* *G06F17/30 (2006.01)*

(30)優先權：2017/10/23 中國大陸 201710992297.8

(71)申請人：香港商阿里巴巴集團服務有限公司 (香港地區) ALIBABA GROUP SERVICES LIMITED (HK)
香港

(72)發明人：曹紹升 (CN)；楊新星 (CN)；周俊 (CN)；李　小龍 (US)

(74)代理人：林志剛

申請實體審查：有　申請專利範圍項數：19 項　圖式數：5　　共 38 頁

(54)名稱

基於集群的詞向量處理方法、裝置以及設備

(57)摘要

本說明書實施例公開了基於集群的詞向量處理方法、裝置以及設備，方案包括：集群包括伺服器集群和工作機集群，工作機集群中的各工作機並行地分別讀取部分語料，並從讀取的語料中提取詞及其上下文詞，從伺服器集群中的伺服器獲取對應的詞向量並進行訓練，由伺服器根據一個或者多個工作機對相同詞的詞向量分別的訓練結果，對訓練前保存的相同詞的詞向量進行更新。

指定代表圖：

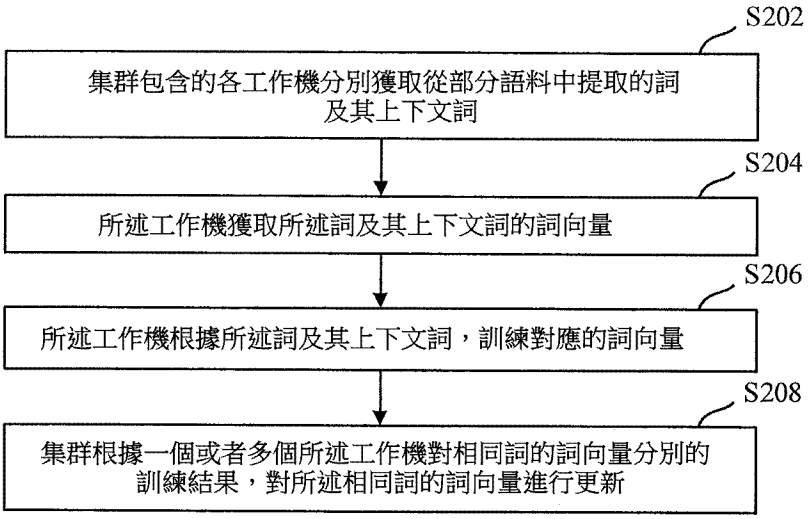

圖 2

【發明說明書】

【中文發明名稱】

基於集群的詞向量處理方法、裝置以及設備

【技術領域】

本說明書涉及電腦軟體技術領域，尤其涉及基於集群的詞向量處理方法、裝置以及設備。

【先前技術】

如今的自然語言處理的解決方案，大都採用基於神經網路的架構，而在這種架構下一個重要的基礎技術就是詞向量。詞向量是將詞映射到一個固定維度的向量，該向量表徵了該詞的語義資訊。

在現有技術中，常見的用於生成詞向量的演算法比如包括谷歌公司的單詞向量演算法、微軟公司的深度神經網路演算法等，往往在單機上運行。

基於現有技術，需要高效的大規模詞向量訓練方案。

【發明內容】

本說明書實施例提供基於集群的詞向量處理方法、裝置以及設備，用以解決如下技術問題：需要高效的大規模詞向量訓練方案。

為解決上述技術問題，本說明書實施例是這樣實現的：

本說明書實施例提供的一種基於集群的詞向量處理方法，所述集群包括多個工作機，所述方法包括：

各所述工作機分別執行：

獲取從部分語料中提取的詞及其上下文詞；

獲取所述詞及其上下文詞的詞向量；

根據所述詞及其上下文詞，訓練對應的詞向量；

根據一個或者多個所述工作機對相同詞的詞向量分別的訓練結果，對所述相同詞的詞向量進行更新。

本說明書實施例提供的一種基於集群的詞向量處理裝置，所述集群包括多個工作機，所述裝置位於所述集群，包括整合更新模組、位於所述工作機的訓練模組；

各所述工作機的訓練模組分別執行：

獲取從部分語料中提取的詞及其上下文詞；

獲取所述詞及其上下文詞的詞向量；

根據所述詞及其上下文詞，訓練對應的詞向量；

所述整合更新模組，根據一個或者多個所述工作機對相同詞的詞向量分別的訓練結果，對所述相同詞的詞向量進行更新。

本說明書實施例提供的一種基於集群的詞向量處理設備，所述設備屬於所述集群，包括：

至少一個處理器；以及，

與所述至少一個處理器通信連接的記憶體；其中，

所述記憶體儲存有可被所述至少一個處理器執行的指令，所述指令被所述至少一個處理器執行，以使所述至少

一個處理器能夠：

獲取從部分語料中提取的詞及其上下文詞；

獲取所述詞及其上下文詞的詞向量；

根據所述詞及其上下文詞，訓練對應的詞向量；

根據一個或者多個所述處理器對相同詞的詞向量分別的訓練結果，對所述相同詞的詞向量進行更新。

本說明書實施例採用的上述至少一個技術方案能夠達到以下有益效果：集群的分散式並行處理能力使得該方案能夠適用於大規模詞向量訓練且效率較高。

【圖式簡單說明】

為了更清楚地說明本說明書實施例或現有技術中的技術方案，下面將對實施例或現有技術描述中所需要使用的附圖作簡單地介紹，顯而易見地，下面描述中的附圖僅僅是本說明書中記載的一些實施例，對於本領域普通技術人員來講，在不付出進步性勞動性的前提下，還可以根據這些附圖獲得其他的附圖。

圖1為本說明書的方案在一種實際應用場景下涉及的一種整體架構示意圖；

圖2為本說明書實施例提供的一種基於集群的詞向量處理方法的流程示意圖；

圖3為本說明書實施例提供的一種實際應用場景下，基於集群的詞向量處理方法的原理示意圖；

圖4為本說明書實施例提供的對應於圖3的一種基於集

群的詞向量處理方法的詳細流程示意圖；

圖5為本說明書實施例提供的對應於圖2的一種基於集群的詞向量處理裝置的結構示意圖。

【實施方式】

本說明書實施例提供基於集群的詞向量處理方法、裝置以及設備。

為了使本技術領域的人員更好地理解本說明書中的技術方案，下面將結合本說明書實施例中的附圖，對本說明書實施例中的技術方案進行清楚、完整地描述，顯然，所描述的實施例僅僅是本申請一部分實施例，而不是全部的實施例。基於本說明書實施例，本領域普通技術人員在沒有作出進步性勞動前提下所獲得的所有其他實施例，都應當屬於本申請保護的範圍。

本說明書的方案適用於集群，在集群下對於大規模詞向量的處理效率更高，具體地：可以拆分訓練語料，進而由集群中的多個工作機分散式地分別根據拆分的部分語料，訓練所述部分語料對應的詞向量，由於各部分語料可能包含相同詞，因此，對於各工作機對相同詞的詞向量分別的訓練結果進行整合，以便於進一步地對訓練前保存的該相同詞的詞向量進行更新。

方案涉及的集群可以有一個或者多個，以圖1為例，涉及了兩個集群。

圖1為本說明書的方案在一種實際應用場景下涉及的

一種整體架構示意圖。該整體架構中，主要涉及三部分：伺服器集群、工作機集群、資料庫。資料庫保存有用於訓練的語料，供工作機集群讀取，伺服器集群保存原始的詞向量，工作機集群與伺服器集群進行配合，實現對詞向量的訓練以及根據訓練結果對伺服器集群上的詞向量的更新。

圖1中的架構是示例性的，並非唯一。比如，方案也可以只涉及一個集群，該集群中包含至少一個調度機和多個工作機，由調度機完成上述伺服器集群的工作；再比如，方案也可以涉及一個工作機集群和一個伺服器；等等。

下面對本說明書的方案進行詳細說明。

圖2為本說明書實施例提供的一種基於集群的詞向量處理方法的流程示意圖，所述集群包括多個工作機。圖2中各步驟由集群中的至少一個機器(或者機器上的程式)執行，不同步驟的執行主體可以不同，圖2中的流程可以執行多輪，每輪可以使用不同組的語料。

圖2中的流程包括以下步驟：

S202：集群包含的各工作機分別獲取從部分語料中提取的詞及其上下文詞。

S204：所述工作機獲取所述詞及其上下文詞的詞向量。

S206：所述工作機根據所述詞及其上下文詞，訓練對應的詞向量。

S208：集群根據一個或者多個所述工作機對相同詞的詞向量分別的訓練結果，對所述相同詞的詞向量進行更新。

在本說明書實施例中，各工作機可以分散式地並存執行步驟S202~S206，其中，各工作機對應的部分語料通常是不同的，如此能夠高效利用大規模的訓練語料，也能夠提高詞向量的訓練效率。比如，對於當前用於訓練詞向量的語料，可以將語料拆分為多份，各工作機可以分別讀取一部分，進而基於自己讀取的部分語料執行步驟S202~S206。

為了便於描述，對於步驟S202~S204，以下各實施例主要從某一個工作機的角度進行說明。

在本說明書實施例中，若本輪流程是第一輪流程，步驟S204中獲取的詞向量可以是初始化得到的。比如，可以採用隨機初始化的方式或者按照指定概率分佈初始化的方式，初始化各詞的詞向量，以及各詞的上下文詞的詞向量，指定概率分佈比如是0-1分佈等。而若本輪流程並非第一輪流程，則步驟S204中獲取的詞向量可以是上輪流程執行完畢後更新並保存的詞向量。

在本說明書實施例中，一般地，步驟S208可以由工作機集群以外的伺服器集群執行，或者由與工作機屬於同一集群的調度機或伺服器執行，如此可以降低工作機的負擔。相應地，更新後的詞向量可以保存於伺服器上，以便下輪流程使用。

以此類推，進行多輪流程直至所有組的訓練語料全部使用完畢後，可以將最終更新得到的詞向量寫出到資料庫，以便用於需求詞向量的各種場景，或者也可以仍然保存於集群中。

通過圖2的方法，集群的分散式並行處理能力使得該方法能夠適用於大規模詞向量訓練且效率較高，不僅如此，也能夠高效地利用大規模的訓練資料。

基於圖2的方法，本說明書實施例還提供了該方法的一些具體實施方案，以及擴展方案，下面以圖1中的架構為例，進行說明。

在本說明書的實施例中，基於圖1的架構，圖2中的集群包括伺服器集群和工作機集群，由工作機集群執行步驟S202~S206，伺服器集群執行步驟S208。

圖1的架構也可以稱為參數伺服器，通過參數伺服器能夠實現常見的並行需求：資料並行、模型並行。資料並行指：每台機器載入不同的訓練資料，同步進行模型訓練，每隔一段時間，可能會進行一次全域資訊同步。模型並行指：每台機器僅載入部分模型參數，所有機器載入的模型參數放在一起為全量的模型參數。

伺服器集群主要是實現模型並行，即伺服器集群記憶體中維護一份全量的模型參數，而工作機集群讀入不同的訓練資料並行進行訓練。整個過程為：伺服器集群將參數分發給工作機集群(每個工作機可能讀入全量模型參數，也可以只是部分模型參數)；每個工作機讀入不同訓練資

料開始並行訓練、更新本機的模型參數；工作機集群將訓練好的模型參數回傳到伺服器集群；伺服器集群綜合所有更新資料做出匯總處理，即模型更新，然後將新的模型參數再傳給工作機集群；按照此過程，交互進行，直到所有訓練資料訓練完畢，或者達到最大訓練次數。具體到本說明書的場景，上述訓練資料即可以是語料，模型參數即可以是詞向量。

在本說明書實施例中，從語料中提取詞及其上下文詞可以由工作機執行，也可以由其他設備預先執行。以前一種方式為例，則對於步驟S202，所述獲取從部分語料中提取的詞及其上下文詞前，還可以執行：各所述工作機分散式地讀取得到部分語料。語料若保存於資料庫，則可以從資料庫讀取。

在本說明書實施例中。所述獲取從部分語料中提取的詞及其上下文詞，具體可以包括：根據自己所讀取得到的語料，建立相應的詞對，所述詞對包含當前詞及其上下詞。比如，可以掃描自己所讀取得到的語料中的詞，當前掃描的詞為當前詞記作w，根據設定的滑窗距離確定包含w的一個滑窗，將該滑窗內的其他每個詞分別作為w的一個上下文詞，記作c，如此構成詞對{w,c}。

進一步地，假定詞向量保存於伺服器集群包含的多個伺服器上。則對於步驟S204，所述獲取所述詞及其上下文詞的詞向量，具體可以包括：根據自己建立的各所述詞對，提取得到當前詞集合和上下文詞集合；從所述伺服器

獲取所述當前詞集合和上下文詞集合包含的詞的詞向量。當然，這並非唯一實施方式，比如，也可以在掃描語料時，同步地從伺服器獲取當前掃描到的詞的詞向量而未必要依賴於建立的詞對，等等。

在本說明書實施例中，可以根據指定的損失函數和建立的詞對，訓練對應的詞向量。

為了獲得更好的訓練效果以及更快地收斂，還可以結合指定的負樣例詞對照地進行訓練，負樣例詞被視為：相比於上下文詞，與對應的當前詞相關性相對低的詞，一般可以在全部詞中隨機選擇若干個。在這種情況下，對於步驟S206，所述根據所述詞及其上下文詞，訓練對應的詞向量，具體可以包括：根據指定的損失函數、負樣例詞，以及自己建立的各所述詞對，訓練對應的詞向量。當前詞和每個負樣例詞也可以構成一個詞對，假定有λ個負樣例詞，相應的詞可以記作$\{w,c_1'\}$、$\{w,c_2'\}$、…、$\{w,c_\lambda'\}$，為了便於描述將負樣例詞對和上面的上下文詞對統一記作$\{w,c\}$，並用y來區分，對於上下文詞對，y=1，對於負樣例詞對，y=0。

為了便於理解，給出損失函數的一個實例如下：

$$L(w,c)=log\sigma\left(\vec{w}\cdot\vec{c}\right)+\lambda E_{c'\sim P(D)}\left[-log\sigma\left(\vec{w}\cdot\vec{c'}\right)\right]$$

其中，$L(w,c)$表示損失函數，c'表示負樣例詞，$\vec{w}$表示w的詞向量，$\vec{c}$表示c的詞向量，$\vec{c'}$表示c'的詞向量，λ為w的負樣例詞數量，σ是啟動函數，比如Sigmoid函數等。

當然，除了上例以外，損失函數也可以有其他實現形

式，訓練目標是使得$\vec{w}$與$\vec{c}$的相似度儘量大，以及$\vec{w}$與$\vec{c}'$的相似度儘量小，上例是用向量點乘度量相似度的，也可以採用其他方式度量相似度。

進一步地，若採用梯度下降法訓練詞向量，則所述根據指定的損失函數、負樣例詞，以及自己建立的各所述詞對，訓練對應的詞向量，具體可以包括：對自己所讀取得到的語料進行遍歷；根據指定的損失函數、負樣例詞，以及自己建立的各所述詞對，計算梯度，並根據所述梯度對遍歷的當前詞及其上下文詞的詞向量進行反覆運算更新。

在實際應用中，每個工作機上的一個或者多個執行緒可以以非同步計算且不加鎖的方式，所述對遍歷的當前詞及其上下文詞的詞向量進行反覆運算更新。從而，工作機內各執行緒也可以並行更新且不會相互妨礙，能夠進一步地提高訓練效率。

在本說明書實施例中，當採用不同的損失函數和不同的啟動函數時，梯度以及訓練結果也可能不同。沿用損失函數的上例，對訓練過程中的計算進行說明。

所述根據指定的損失函數、負樣例詞，以及自己建立的各所述詞對，計算梯度，並根據所述梯度對遍歷的當前詞及其上下文詞的詞向量進行反覆運算更新，具體可以包括：

按照以下公式，對遍歷的當前詞及其上下文詞的詞向量進行反覆運算更新：

$$\vec{w}_{i,t+1}=\vec{w}_{i,t}+g\vec{c}_{i,t},w\in B_{i,k}\quad\text{(公式一)}$$

$$\vec{c}_{i,t+1}=\vec{c}_{i,t}+g\vec{w}_{i,t},c\in\Gamma(w)\quad\text{(公式二)}$$

其中，$g=\alpha\left(y-\sigma\left(\vec{w}\cdot\vec{c}\right)\right)$，$y=\begin{cases}1,\{w,c\}\\0,\{w,c'\}\end{cases}$，w表示當前詞，c表示w的上下文詞，c'表示負樣例詞，$\vec{w}$表示w的詞向量，$\vec{c}$表示c的詞向量，$\vec{w}_{i,t}$和$\vec{c}_{i,t}$表示第t個工作機上第i次更新，$B_{i,k}$表示第i個工作機上第k組語料，$\Gamma(w)$表示w的上下文詞集合，α表示學習率，比如可以取0.025，σ為Sigmoid函數，也即$\sigma=\frac{1}{1+e^{-x}}$。

進一步地對梯度的計算進行說明：

$$\nabla\sigma(z)|_z=\frac{1}{\sigma(z)}\cdot\sigma(z)\cdot(1-\sigma(z))=1-\sigma(z)$$

$\nabla\sigma(-z)|_z=-(1-\sigma(-z))=-\sigma(z)$；則有：

$$\nabla L(w,c)|_{\vec{w}}=\left(y-\sigma\left(\vec{w}\cdot\vec{c}\right)\right)\vec{c}$$

$$\nabla L(w,c)|_{\vec{c}}=\left(y-\sigma\left(\vec{w}\cdot\vec{c}\right)\right)\vec{w}。$$

在本說明書實施例中，當由伺服器根據訓練結果更新詞向量時，對於步驟S208，所述根據一個或者多個所述工作機對相同詞的詞向量分別的訓練結果，對所述相同詞的詞向量進行更新，具體可以包括：所述伺服器獲取一個或者多個所述工作機對相同詞的詞向量分別的訓練結果；根據各所述訓練結果，以及訓練前保存的所述相同詞的詞向量，進行平均計算，得到向量增量值，並根據所述向量增量值對所述相同詞的詞向量進行更新。

該更新過程即是模型平均過程，平均計算可以有多種

實現方式，比如，以詞的在各工作機的出現次數作為權重對各工作機的訓練結果進行平均；再比如，直接對各工作機的訓練結果進行平均；等等。以前一種方式為例，比如可以按照以下公式，計算得到上述的向量增量值：

$$\Delta\left(\vec{w}\right)=\frac{\sum_{i=0}^{I}\lambda_i(w)\left(\vec{w}_{i,T}-\vec{w}_{srv}\right)}{\sum_{i=0}^{I}\lambda_i(w)} \quad \text{(公式三)}$$

$$\Delta\left(\vec{c}\right)=\frac{\sum_{i=0}^{I}\lambda_i(c)\left(\vec{c}_{i,T}-\vec{w}_{srv}\right)}{\sum_{i=0}^{I}\lambda_i(c)} \quad \text{(公式四)}$$

其中，$\vec{w}_{i,T}$和$\vec{c}_{i,T}$表示第i個工作機上反覆運算更新結果，$\lambda_i(w)$表示w在第i個工作機上出現的次數，$\vec{w}_{srv}$表示伺服器訓練前保存的$\vec{w}$。

更新前的詞向量加上計算出的對應的向量增量值，即可以得到更新後的詞向量。

根據上面的說明，本說明書實施例還提供了一種實際應用場景下，基於集群的詞向量處理方法的原理示意圖，如圖3所示，進一步地，本說明書實施例還提供了對應於圖3的一種基於集群的詞向量處理方法的詳細流程示意圖，如圖4所示。

在圖3中，示例性地示出了工作機0~2、伺服器0~2，主要針對工作機0進行說明，而工作機1和2簡略地進行了表示，工作方式與工作機0是一致的。“wid”、“cid”為標識，分別表示當前詞和上下文詞，“wid list”、“cid list”是識別欄位表，分別表示當前詞集合和上下文詞集合。圖3

中的簡略工作流程包括：各工作機分散式地讀取語料，建立詞對；各工作機從伺服器集群獲取相應的詞向量；各工作機利用讀取的語料訓練詞向量；伺服器集群根據各工作機的訓練結果進行模型平均。

圖4中示出了更詳細的流程，主要包括以下步驟：

S402：各工作機分散式地讀取部分語料，建立詞對{w,c}，從詞對中提取wid list和cid list，如圖4中的工作機0所示。

S404：工作機根據wid list和cid list，從伺服器集群獲取相應的詞向量。

S406：工作機根據詞對，計算梯度，進而反覆運算更新詞向量，具體採用上述的公式一和公式二進行計算。

S408：在各工作機反覆運算更新完畢後，伺服器集群進行模型平均，以對反覆運算更新結果進行整合，具體採用上述的公式三和公式四進行計算。

基於同樣的思路，本說明書實施例還提供了上述方法的對應裝置，如圖5所示。

圖5為本說明書實施例提供的對應於圖2的一種基於集群的詞向量處理裝置的結構示意圖，所述集群包括多個工作機，所述裝置位於所述集群，包括整合更新模組501、位於所述工作機的訓練模組502；

各所述工作機的訓練模組502分別執行：

獲取從部分語料中提取的詞及其上下文詞；

獲取所述詞及其上下文詞的詞向量；

根據所述詞及其上下文詞，訓練對應的詞向量；

所述整合更新模組501，根據一個或者多個所述工作機對相同詞的詞向量分別的訓練結果，對所述相同詞的詞向量進行更新。

可選地，所述工作機還具有讀取模組503，在所述訓練模組502獲取從部分語料中提取的詞及其上下文詞前，各所述工作機的讀取模組503分散式地讀取得到部分語料；

所述訓練模組502獲取從部分語料中提取的詞及其上下文詞，具體包括：

所述訓練模組502根據自己所在工作機的讀取模組503所讀取得到的語料，建立相應的詞對，所述詞對包含當前詞及其上下詞。

可選地，所述集群還包括多個伺服器，所述訓練模組502獲取所述詞及其上下文詞的詞向量，具體包括：

所述訓練模組502根據自己建立的各所述詞對，提取得到當前詞集合和上下文詞集合；

從所述伺服器獲取所述當前詞集合和上下文詞集合包含的詞的詞向量。

可選地，所述訓練模組502根據所述詞及其上下文詞，訓練對應的詞向量，具體包括：

所述訓練模組502根據指定的損失函數、負樣例詞，以及自己建立的各所述詞對，訓練對應的詞向量。

可選地，所述訓練模組502根據指定的損失函數、負

樣例詞，以及自己建立的各所述詞對，訓練對應的詞向量，具體包括：

所述訓練模組502對自己所讀取得到的語料進行遍歷；

根據指定的損失函數、負樣例詞，以及自己建立的各所述詞對，計算梯度，並根據所述梯度對遍歷的當前詞及其上下文詞的詞向量進行反覆運算更新。

可選地，所述訓練模組502根據指定的損失函數、負樣例詞，以及自己建立的各所述詞對，計算梯度，並根據所述梯度對遍歷的當前詞及其上下文詞的詞向量進行反覆運算更新，具體包括：

所述訓練模組502按照以下公式，對遍歷的當前詞及其上下文詞的詞向量進行反覆運算更新：

$$\vec{w}_{i,t+1}=\vec{w}_{i,t}+g\vec{c}_{i,t},w\in B_{i,k}$$

$$\vec{c}_{i,t+1}=\vec{c}_{i,t}+g\vec{w}_{i,t},c\in\Gamma(w)$$

其中，$g=\alpha\left(y-\sigma\left(\vec{w}\cdot\vec{c}\right)\right)$，$y=\begin{cases}1,\{w,c\}\\0,\{w,c'\}\end{cases}$，w表示當前詞，c表示w的上下文詞，c'表示負樣例詞，$\vec{w}$表示w的詞向量，$\vec{c}$表示c的詞向量，$\vec{w}_{i,t}$和$\vec{c}_{i,t}$表示第t個工作機上第i次更新，$B_{i,k}$表示第i個工作機上第k組語料，$\Gamma(w)$表示w的上下文詞集合，α表示學習率，σ為Sigmoid函數。

可選地，所述訓練模組502對遍歷的當前詞及其上下文詞的詞向量進行反覆運算更新，具體包括：

所述訓練模組502通過所在工作機上的一個或者多個

執行緒，以非同步計算且不加鎖的方式，所述對遍歷的當前詞及其上下文詞的詞向量進行反覆運算更新。

可選地，所述整合更新模組501位於所述伺服器，所述整合更新模組501根據一個或者多個所述工作機對相同詞的詞向量分別的訓練結果，對所述相同詞的詞向量進行更新，具體包括：

所述整合更新模組501獲取一個或者多個所述工作機對相同詞的詞向量分別的訓練結果；

根據各所述訓練結果，以及訓練前保存的所述相同詞的詞向量，進行平均計算，得到向量增量值，並根據所述向量增量值對所述相同詞的詞向量進行更新。

可選地，所述整合更新模組501根據各所述訓練結果，以及訓練前保存的所述相同詞的詞向量，進行平均計算，得到向量增量值，具體包括：

所述整合更新模組501按照以下公式，計算得到向量增量值：

$$\Delta\left(\vec{w}\right)=\frac{\sum_{i=0}^{I}\lambda_i(w)\left(\vec{w}_{i,T}-\vec{w}_{srv}\right)}{\sum_{i=0}^{I}\lambda_i(w)}$$

$$\Delta\left(\vec{c}\right)=\frac{\sum_{i=0}^{I}\lambda_i(c)\left(\vec{c}_{i,T}-\vec{w}_{srv}\right)}{\sum_{i=0}^{I}\lambda_i(c)}$$

其中，w表示當前詞，c表示w的上下文詞，$\vec{w}$表示w的詞向量，$\vec{c}$表示c的詞向量，$\vec{w}_{i,T}$和$\vec{c}_{i,T}$表示第i個工作機上反覆運算更新結果，$\lambda_i(w)$表示w在第i個工作機上出現的次數，$\vec{w}_{srv}$表示伺服器訓練前保存的$\vec{w}$。

基於同樣的思路，本說明書實施例還提供了對應於圖2的一種基於集群的詞向量處理設備，該設備屬於所述集群，包括：

至少一個處理器；以及，

與所述至少一個處理器通信連接的記憶體；其中，

所述記憶體儲存有可被所述至少一個處理器執行的指令，所述指令被所述至少一個處理器執行，以使所述至少一個處理器能夠：

獲取從部分語料中提取的詞及其上下文詞；

獲取所述詞及其上下文詞的詞向量；

根據所述詞及其上下文詞，訓練對應的詞向量；

根據一個或者多個對相同詞的詞向量分別的訓練結果，對所述相同詞的詞向量進行更新。

基於同樣的思路，本說明書實施例還提供了對應於圖2的一種非易失性電腦儲存媒體，儲存有電腦可執行指令，所述電腦可執行指令設置為：

獲取從部分語料中提取的詞及其上下文詞；

獲取所述詞及其上下文詞的詞向量；

根據所述詞及其上下文詞，訓練對應的詞向量；

根據一個或者多個對相同詞的詞向量分別的訓練結果，對所述相同詞的詞向量進行更新。

上述對本說明書特定實施例進行了描述。其它實施例在所附申請專利範圍的範圍內。在一些情況下，在申請專利範圍中記載的動作或步驟可以按照不同於實施例中的順

序來執行並且仍然可以實現期望的結果。另外，在附圖中描繪的過程不一定要求示出的特定順序或者連續順序才能實現期望的結果。在某些實施方式中，多工處理和並行處理也是可以的或者可能是有利的。

本說明書中的各個實施例均採用遞進的方式描述，各個實施例之間相同相似的部分互相參見即可，每個實施例重點說明的都是與其他實施例的不同之處。尤其，對於裝置、設備、非易失性電腦儲存媒體實施例而言，由於其基本相似於方法實施例，所以描述的比較簡單，相關之處參見方法實施例的部分說明即可。

本說明書實施例提供的裝置、設備、非易失性電腦儲存媒體與方法是對應的，因此，裝置、設備、非易失性電腦儲存媒體也具有與對應方法類似的有益技術效果，由於上面已經對方法的有益技術效果進行了詳細說明，因此，這裡不再贅述對應裝置、設備、非易失性電腦儲存媒體的有益技術效果。

在20世紀90年代，對於一個技術的改進可以很明顯地區分是硬體上的改進(例如，對二極體、電晶體、開關等電路結構的改進)還是軟體上的改進(對於方法流程的改進)。然而，隨著技術的發展，當今的很多方法流程的改進已經可以視為硬體電路結構的直接改進。設計人員幾乎都通過將改進的方法流程程式設計到硬體電路中來得到相應的硬體電路結構。因此，不能說一個方法流程的改進就不能用硬體實體模組來實現。例如，可程式設計邏輯裝置

(Programmable Logic Device, PLD)(例如現場可程式設計閘陣列(Field Programmable Gate Array，FPGA))就是這樣一種積體電路，其邏輯功能由使用者對裝置程式設計來確定。由設計人員自行程式設計來把一個數位系統“集成”在一片PLD上，而不需要請晶片製造廠商來設計和製作專用的積體電路晶片。而且，如今，取代手工地製作積體電路晶片，這種程式設計也多半改用“邏輯編譯器(logic compiler)”軟體來實現，它與程式開發撰寫時所用的軟體編譯器相類似，而要編譯之前的原始代碼也得用特定的程式設計語言來撰寫，此稱之為硬體描述語言(Hardware Description Language，HDL)，而HDL也並非僅有一種，而是有許多種，如ABEL(Advanced Boolean Expression Language)、AHDL(Altera Hardware Description Language)、Confluence、CUPL(Cornell University Programming Language)、HDCal、JHDL(Java Hardware Description Language)、Lava、Lola、MyHDL、PALASM、RHDL(Ruby Hardware Description Language)等，目前最普遍使用的是VHDL(Very-High-Speed Integrated Circuit Hardware Description Language)與Verilog。本領域技術人員也應該清楚，只需要將方法流程用上述幾種硬體描述語言稍作邏輯程式設計並程式設計到積體電路中，就可以很容易得到實現該邏輯方法流程的硬體電路。

控制器可以按任何適當的方式實現，例如，控制器可

以採取例如微處理器或處理器以及儲存可由該(微)處理器執行的電腦可讀程式碼(例如軟體或固件)的電腦可讀媒體、邏輯閘、開關、專用積體電路(Application Specific Integrated Circuit，ASIC)、可程式設計邏輯控制器和嵌入微控制器的形式，控制器的例子包括但不限於以下微控制器：ARC 625D、Atmel AT91SAM、Microchip PIC18F26K20以及Silicone Labs C8051F320，記憶體控制器還可以被實現為記憶體的控制邏輯的一部分。本領域技術人員也知道，除了以純電腦可讀程式碼方式實現控制器以外，完全可以通過將方法步驟進行邏輯程式設計來使得控制器以邏輯閘、開關、專用積體電路、可程式設計邏輯控制器和嵌入微控制器等的形式來實現相同功能。因此這種控制器可以被認為是一種硬體部件，而對其內包括的用於實現各種功能的裝置也可以視為硬體部件內的結構。或者甚至，可以將用於實現各種功能的裝置視為既可以是實現方法的軟體模組又可以是硬體部件內的結構。

上述實施例闡明的系統、裝置、模組或單元，具體可以由電腦晶片或實體實現，或者由具有某種功能的產品來實現。一種典型的實現設備為電腦。具體的，電腦例如可以為個人電腦、膝上型電腦、蜂巢式電話、相機電話、智慧型電話、個人數位助理、媒體播放機、導航設備、電子郵件設備、遊戲控制台、平板電腦、可穿戴設備或者這些設備中的任何設備的組合。

為了描述的方便，描述以上裝置時以功能分為各種單

元分別描述。當然，在實施本說明書時可以把各單元的功能在同一個或多個軟體和/或硬體中實現。

本領域內的技術人員應明白，本說明書實施例可提供為方法、系統、或電腦程式產品。因此，本說明書實施例可採用完全硬體實施例、完全軟體實施例、或結合軟體和硬體方面的實施例的形式。而且，本說明書實施例可採用在一個或多個其中包含有電腦可用程式碼的電腦可用儲存媒體(包括但不限於磁碟記憶體、CD-ROM、光學記憶體等)上實施的電腦程式產品的形式。

本說明書是參照根據本說明書實施例的方法、設備(系統)、和電腦程式產品的流程圖和/或方塊圖來描述的。應理解可由電腦程式指令實現流程圖和/或方塊圖中的每一流程和/或方塊、以及流程圖和/或方塊圖中的流程和/或方塊的結合。可提供這些電腦程式指令到通用電腦、專用電腦、嵌入式處理機或其他可程式設計資料處理設備的處理器以產生一個機器，使得通過電腦或其他可程式設計資料處理設備的處理器執行的指令產生用於實現在流程圖一個流程或多個流程和/或方塊圖一個方塊或多個方塊中指定的功能的裝置。

這些電腦程式指令也可儲存在能引導電腦或其他可程式設計資料處理設備以特定方式工作的電腦可讀記憶體中，使得儲存在該電腦可讀記憶體中的指令產生包括指令裝置的製造品，該指令裝置實現在流程圖一個流程或多個流程和/或方塊圖一個方塊或多個方塊中指定的功能。

這些電腦程式指令也可裝載到電腦或其他可程式設計資料處理設備上，使得在電腦或其他可程式設計設備上執行一系列操作步驟以產生電腦實現的處理，從而在電腦或其他可程式設計設備上執行的指令提供用於實現在流程圖一個流程或多個流程和/或方塊圖一個方塊或多個方塊中指定的功能的步驟。

在一個典型的配置中，計算設備包括一個或多個處理器(CPU)、輸入/輸出介面、網路介面和記憶體。

記憶體可能包括電腦可讀媒體中的非永久性記憶體，隨機存取記憶體(RAM)和/或非易失性記憶體等形式，如唯讀記憶體(ROM)或快閃記憶體(flash RAM)。記憶體是電腦可讀媒體的示例。

電腦可讀媒體包括永久性和非永久性、可移動和非可移動媒體可以由任何方法或技術來實現資訊儲存。資訊可以是電腦可讀指令、資料結構、程式的模組或其他資料。電腦的儲存媒體的例子包括，但不限於相變記憶體(PRAM)、靜態隨機存取記憶體(SRAM)、動態隨機存取記憶體(DRAM)、其他類型的隨機存取記憶體(RAM)、唯讀記憶體(ROM)、電可擦除可程式設計唯讀記憶體(EEPROM)、快閃記憶體或其他記憶體技術、唯讀光碟唯讀記憶體(CD-ROM)、數位多功能光碟(DVD)或其他光學儲存、磁盒式磁帶，磁帶磁磁片儲存或其他磁性存放裝置或任何其他非傳輸媒體，可用於儲存可以被計算設備存取的資訊。按照本文中的界定，電腦可讀媒體不包括暫存電

腦可讀媒體(transitory media)，如調變的資料信號和載波。

還需要說明的是，術語“包括”、“包含”或者其任何其他變體意在涵蓋非排他性的包含，從而使得包括一系列要素的過程、方法、商品或者設備不僅包括那些要素，而且還包括沒有明確列出的其他要素，或者是還包括為這種過程、方法、商品或者設備所固有的要素。在沒有更多限制的情況下，由語句“包括一個……”限定的要素，並不排除在包括所述要素的過程、方法、商品或者設備中還存在另外的相同要素。

本說明書可以在由電腦執行的電腦可執行指令的一般上下文中描述，例如程式模組。一般地，程式模組包括執行特定任務或實現特定抽象資料類型的常式、程式、物件、組件、資料結構等等。也可以在分散式運算環境中實踐本說明書，在這些分散式運算環境中，由通過通信網路而被連接的遠端處理設備來執行任務。在分散式運算環境中，程式模組可以位於包括存放裝置在內的本地和遠端電腦儲存媒體中。

本說明書中的各個實施例均採用遞進的方式描述，各個實施例之間相同相似的部分互相參見即可，每個實施例重點說明的都是與其他實施例的不同之處。尤其，對於系統實施例而言，由於其基本相似於方法實施例，所以描述的比較簡單，相關之處參見方法實施例的部分說明即可。

以上所述僅為本說明書實施例而已，並不用於限制本

申請。對於本領域技術人員來說，本申請可以有各種更改和變化。凡在本申請的精神和原理之內所作的任何修改、等同替換、改進等，均應包含在本申請的申請專利範圍之內。

【符號說明】

501：整合更新模組

502：訓練模組

503：讀取模組

201917603

【發明摘要】

【中文發明名稱】

基於集群的詞向量處理方法、裝置以及設備

【中文】

本說明書實施例公開了基於集群的詞向量處理方法、裝置以及設備，方案包括：集群包括伺服器集群和工作機集群，工作機集群中的各工作機並行地分別讀取部分語料，並從讀取的語料中提取詞及其上下文詞，從伺服器集群中的伺服器獲取對應的詞向量並進行訓練，由伺服器根據一個或者多個工作機對相同詞的詞向量分別的訓練結果，對訓練前保存的相同詞的詞向量進行更新。

【指定代表圖】第(２)圖。
【代表圖之符號簡單說明】無

【特徵化學式】無

【發明申請專利範圍】

【第1項】

一種基於集群的詞向量處理方法，所述集群包括多個工作機，所述方法包括：

各所述工作機分別執行：

獲取從部分語料中提取的詞及其上下文詞；

獲取所述詞及其上下文詞的詞向量；

根據所述詞及其上下文詞，訓練對應的詞向量；

根據一個或者多個所述工作機對相同詞的詞向量分別的訓練結果，對所述相同詞的詞向量進行更新。

【第2項】

如申請專利範圍第1項所述的方法，所述獲取從部分語料中提取的詞及其上下文詞前，所述方法還包括：

各所述工作機分散式地讀取得到部分語料；

所述獲取從部分語料中提取的詞及其上下文詞，具體包括：

根據自己所讀取得到的語料，建立相應的詞對，所述詞對包含當前詞及其上下詞。

【第3項】

如申請專利範圍第2項所述的方法，所述集群還包括多個伺服器，所述獲取所述詞及其上下文詞的詞向量，具體包括：

根據自己建立的各所述詞對，提取得到當前詞集合和上下文詞集合；

從所述伺服器獲取所述當前詞集合和上下文詞集合包含的詞的詞向量。

【第4項】

如申請專利範圍第2項所述的方法，所述根據所述詞及其上下文詞，訓練對應的詞向量，具體包括：

根據指定的損失函數、負樣例詞，以及自己建立的各所述詞對，訓練對應的詞向量。

【第5項】

如申請專利範圍第4項所述的方法，所述根據指定的損失函數、負樣例詞，以及自己建立的各所述詞對，訓練對應的詞向量，具體包括：

對自己所讀取得到的語料進行遍歷；

根據指定的損失函數、負樣例詞，以及自己建立的各所述詞對，計算梯度，並根據所述梯度對遍歷的當前詞及其上下文詞的詞向量進行反覆運算更新。

【第6項】

如申請專利範圍第5項所述的方法，所述根據指定的損失函數、負樣例詞，以及自己建立的各所述詞對，計算梯度，並根據所述梯度對遍歷的當前詞及其上下文詞的詞向量進行反覆運算更新，具體包括：

按照以下公式，對遍歷的當前詞及其上下文詞的詞向量進行反覆運算更新：

$$\vec{w}_{i,t+1} = \vec{w}_{i,t} + g\vec{c}_{i,t}, w \in B_{i,k}$$

$$\vec{c}_{i,t+1} = \vec{c}_{i,t} + g\vec{w}_{i,t}, c \in \Gamma(w)$$

其中，$g=\alpha\left(y-\sigma\left(\vec{w}\cdot\vec{c}\right)\right)$，$y=\begin{cases}1,\{w,c\}\\0,\{w,c'\}\end{cases}$，w表示當前詞，c表示w的上下文詞，c'表示負樣例詞，$\vec{w}$表示w的詞向量，$\vec{c}$表示c的詞向量，$\vec{w}_{i,t}$和$\vec{c}_{i,t}$表示第t個工作機上第i次更新，$B_{i,k}$表示第i個工作機上第k組語料，$\Gamma(w)$表示w的上下文詞集合，α表示學習率，σ為Sigmoid函數。

【第7項】

如申請專利範圍第6項所述的方法，所述對遍歷的當前詞及其上下文詞的詞向量進行反覆運算更新，具體包括：

所述工作機上的一個或者多個執行緒以非同步計算且不加鎖的方式，所述對遍歷的當前詞及其上下文詞的詞向量進行反覆運算更新。

【第8項】

如申請專利範圍第3項所述的方法，所述根據一個或者多個所述工作機對相同詞的詞向量分別的訓練結果，對所述相同詞的詞向量進行更新，具體包括：

所述伺服器獲取一個或者多個所述工作機對相同詞的詞向量分別的訓練結果；

根據各所述訓練結果，以及訓練前保存的所述相同詞的詞向量，進行平均計算，得到向量增量值，並根據所述向量增量值對所述相同詞的詞向量進行更新。

【第9項】

如申請專利範圍第8項所述的方法，所述根據各所述

訓練結果，以及訓練前保存的所述相同詞的詞向量，進行平均計算，得到向量增量值，具體包括：

按照以下公式，計算得到向量增量值：

$$\Delta(\vec{w})=\frac{\sum_{i=0}^{I}\lambda_i(w)\left(\vec{w}_{i,T}-\vec{w}_{srv}\right)}{\sum_{i=0}^{I}\lambda_i(w)}$$

$$\Delta(\vec{c})=\frac{\sum_{i=0}^{I}\lambda_i(c)\left(\vec{c}_{i,T}-\vec{w}_{srv}\right)}{\sum_{i=0}^{I}\lambda_i(c)}$$

其中，w表示當前詞，c表示w的上下文詞，$\vec{w}$表示w的詞向量，$\vec{c}$表示c的詞向量，$\vec{w}_{i,T}$和$\vec{c}_{i,T}$表示第i個工作機上反覆運算更新結果，$\lambda_i(w)$表示w在第i個工作機上出現的次數，$\vec{w}_{srv}$表示伺服器訓練前保存的$\vec{w}$。

【第10項】

一種基於集群的詞向量處理裝置，所述集群包括多個工作機，所述裝置位於所述集群，包括整合更新模組、位於所述工作機的訓練模組；

各所述工作機的訓練模組分別執行：

獲取從部分語料中提取的詞及其上下文詞；

獲取所述詞及其上下文詞的詞向量；

根據所述詞及其上下文詞，訓練對應的詞向量；

所述整合更新模組，根據一個或者多個所述工作機對相同詞的詞向量分別的訓練結果，對所述相同詞的詞向量進行更新。

【第11項】

如申請專利範圍第10項所述的裝置，所述工作機還具

有讀取模組，在所述訓練模組獲取從部分語料中提取的詞及其上下文詞前，各所述工作機的讀取模組分散式地讀取得到部分語料；

所述訓練模組獲取從部分語料中提取的詞及其上下文詞，具體包括：

所述訓練模組根據自己所在工作機的讀取模組所讀取得到的語料，建立相應的詞對，所述詞對包含當前詞及其上下詞。

【第12項】

如申請專利範圍第11項所述的裝置，所述集群還包括多個伺服器，所述訓練模組獲取所述詞及其上下文詞的詞向量，具體包括：

所述訓練模組根據自己建立的各所述詞對，提取得到當前詞集合和上下文詞集合；

從所述伺服器獲取所述當前詞集合和上下文詞集合包含的詞的詞向量。

【第13項】

如申請專利範圍第11項所述的裝置，所述訓練模組根據所述詞及其上下文詞，訓練對應的詞向量，具體包括：

所述訓練模組根據指定的損失函數、負樣例詞，以及自己建立的各所述詞對，訓練對應的詞向量。

【第14項】

如申請專利範圍第13項所述的裝置，所述訓練模組根據指定的損失函數、負樣例詞，以及自己建立的各所述詞

對，訓練對應的詞向量，具體包括：

所述訓練模組對自己所讀取得到的語料進行遍歷；

根據指定的損失函數、負樣例詞，以及自己建立的各所述詞對，計算梯度，並根據所述梯度對遍歷的當前詞及其上下文詞的詞向量進行反覆運算更新。

【第15項】

如申請專利範圍第14項所述的裝置，所述訓練模組根據指定的損失函數、負樣例詞，以及自己建立的各所述詞對，計算梯度，並根據所述梯度對遍歷的當前詞及其上下文詞的詞向量進行反覆運算更新，具體包括：

所述訓練模組按照以下公式，對遍歷的當前詞及其上下文詞的詞向量進行反覆運算更新：

$$\vec{w}_{i,t+1} = \vec{w}_{i,t} + g\vec{c}_{i,t}, w \in B_{i,k}$$

$$\vec{c}_{i,t+1} = \vec{c}_{i,t} + g\vec{w}_{i,t}, c \in \Gamma(w)$$

其中，$g=\alpha\left(y-\sigma\left(\vec{w}\cdot\vec{c}\right)\right)$，$y=\begin{cases}1,\{w,c\}\\0,\{w,c'\}\end{cases}$，w表示當前詞，c表示w的上下文詞，c'表示負樣例詞，$\vec{w}$表示w的詞向量，$\vec{c}$表示c的詞向量，$\vec{w}_{i,t}$和$\vec{c}_{i,t}$表示第t個工作機上第i次更新，$B_{i,k}$表示第i個工作機上第k組語料，$\Gamma(w)$表示w的上下文詞集合，α表示學習率，σ為Sigmoid函數。

【第16項】

如申請專利範圍第15項所述的裝置，所述訓練模組對遍歷的當前詞及其上下文詞的詞向量進行反覆運算更新，具體包括：

所述訓練模組通過所在工作機上的一個或者多個執行緒，以非同步計算且不加鎖的方式，所述對遍歷的當前詞及其上下文詞的詞向量進行反覆運算更新。

【第17項】

如申請專利範圍第12項所述的裝置，所述整合更新模組位於所述伺服器，所述整合更新模組根據一個或者多個所述工作機對相同詞的詞向量分別的訓練結果，對所述相同詞的詞向量進行更新，具體包括：

所述整合更新模組獲取一個或者多個所述工作機對相同詞的詞向量分別的訓練結果；

根據各所述訓練結果，以及訓練前保存的所述相同詞的詞向量，進行平均計算，得到向量增量值，並根據所述向量增量值對所述相同詞的詞向量進行更新。

【第18項】

如申請專利範圍第17項所述的裝置，所述整合更新模組根據各所述訓練結果，以及訓練前保存的所述相同詞的詞向量，進行平均計算，得到向量增量值，具體包括：

所述整合更新模組按照以下公式，計算得到向量增量值：

$$\Delta\left(\vec{w}\right)=\frac{\sum_{i=0}^{I}\lambda_i(w)\left(\vec{w}_{i,T}-\vec{w}_{srv}\right)}{\sum_{i=0}^{I}\lambda_i(w)}$$

$$\Delta\left(\vec{c}\right)=\frac{\sum_{i=0}^{I}\lambda_i(c)\left(\vec{c}_{i,T}-\vec{w}_{srv}\right)}{\sum_{i=0}^{I}\lambda_i(c)}$$

其中，w表示當前詞，c表示w的上下文詞，$\vec{w}$表示w

的詞向量，$\vec{c}$表示c的詞向量，$\vec{w}_{i,T}$和$\vec{c}_{i,T}$表示第i個工作機上反覆運算更新結果，$\lambda_i(w)$表示w在第i個工作機上出現的次數，$\vec{w}_{srv}$表示伺服器訓練前保存的$\vec{w}$。

【第19項】

一種基於集群的詞向量處理設備，所述設備屬於所述集群，包括：

至少一個處理器；以及，

與所述至少一個處理器通信連接的記憶體；其中，

所述記憶體儲存有可被所述至少一個處理器執行的指令，所述指令被所述至少一個處理器執行，以使所述至少一個處理器能夠：

獲取從部分語料中提取的詞及其上下文詞；

獲取所述詞及其上下文詞的詞向量；

根據所述詞及其上下文詞，訓練對應的詞向量；

根據一個或者多個對相同詞的詞向量分別的訓練結果，對所述相同詞的詞向量進行更新。