

Claims

What is claimed is:

1. A cluster-based word vector processing method to be performed at a cluster comprising a plurality of worker computers and a server, the method comprising operations performed by each of the worker computers:

obtaining a word and at least one context word of the word that are extracted from a corpus;

obtaining word vectors for the word and the at least one context word;

calculating a gradient according to the word, the at least one context word, and the word vectors; and

asynchronously updating the gradient by sending, without waiting for a peer worker computer to send a gradient to the server, the updated gradient to the server to enable the server to update the word vectors for the word and the at least one context word according to the updated gradient, wherein the server updates the word vectors for the word and the at least one context word, by:

iteratively updating the word vectors for the word and the at least one context word, and a word vector of a negative sample word of the word according to the following formula:

$$\vec{w}_{t+1} = \vec{w}_t - \alpha \cdot \nabla \cdot \vec{c}_t, w \in B_k \text{ and}$$

$$\vec{c}_{t+1} = \vec{c}_t - \alpha \cdot \nabla \cdot \vec{w}_t, c \in \Gamma(w),$$

$$\text{wherein } \nabla = \sigma(\vec{w} \cdot \vec{c}) - y,$$

$$y = \begin{cases} 1, \{w, c\} \\ 0, \{w, c'\} \end{cases}$$

w represents the word, c represents a context word of the at least one context word of w , c' represents the negative sample word, $\vec{w}$ represents the word vector of w , $\vec{c}$ represents the word vector of c , $\vec{w}_t$ and $\vec{c}_t$ represent a t^{th} update on the server, B_k represents a k^{th} corpus on the worker computer, $\Gamma(w)$ represents a set of the context word and the negative sample word of w , α represents a learning rate, and σ is a sigmoid function.

2. The method according to claim 1, wherein before the obtaining the word and the at least one context word of the word that are extracted from the corpus, the method further comprises:

obtaining, by each of the worker computers, a respective portion of the corpus through distributed reading; and

wherein the obtaining the word and the at least one context word of the word that are extracted from the corpus comprises:

creating word pairs according to the respective portion of the corpus, wherein each of the word pairs comprises a current word and a context word of the current word.

3. The method according to claim 2, wherein the obtaining word vectors for the word and the at least one context word comprises:

obtaining a current word set and a context word set through extraction according to the word pairs; and

obtaining word vectors for words included in the current word set and the context word set from the server.

4. The method according to claim 1, wherein the calculating the gradient according to the word, the at least one context word, and the word vectors comprises:

calculating, by one or more threads on the worker computer, the gradient in a manner of asynchronous calculation and unlocked updating.

5. The method according to claim 1, wherein the asynchronously updating the gradient comprises:

sending, by the worker computer, the updated gradient to the server without performing model averaging.

6. A cluster comprising a server and a worker-computer cluster including a plurality of worker computers, wherein:

each of the worker computers includes at least one processor and a memory and is configured to perform operations by the at least one processor, the operations including:

obtaining a word and at least one context word of the word that are extracted from a corpus;

obtaining word vectors for the word and the at least one context word;

calculating a gradient according to the word, the at least one context word, and the word vectors;

asynchronously updating the gradient by sending, without waiting for a peer worker computer to send a gradient to the server, the updated gradient to the server to enable the server to update the word vectors for the word and the at least one context word according to

the updated gradient, wherein the server updates the word vectors for the word and the at least one context word, by:

iteratively updating the word vectors for the word and the at least one context word, and a word vector of a negative sample word of the word according to the following formula:

$$\vec{w}_{t+1} = \vec{w}_t - \alpha \cdot \nabla \cdot \vec{c}_t, w \in B_k \text{ and}$$

$$\vec{c}_{t+1} = \vec{c}_t - \alpha \cdot \nabla \cdot \vec{w}_t, c \in \Gamma(w),$$

$$\text{wherein } \nabla = \sigma(\vec{w} \cdot \vec{c}) - y,$$

$$y = \begin{cases} 1, \{w, c\} \\ 0, \{w, c'\} \end{cases}$$

w represents the word, c represents a context word of the at least one context word of w , c' represents the negative sample word, $\vec{w}$ represents the word vector of w , $\vec{c}$ represents the word vector of c , $\vec{w}_t$ and $\vec{c}_t$ represent a t^{th} update on the server, B_k represents a k^{th} corpus on the worker computer, $\Gamma(w)$ represents a set of the context word and the negative sample word of w , α represents a learning rate, and σ is a sigmoid function.

7. The cluster according to claim 6, wherein before the obtaining a word and at least one context word of the word that are extracted from a corpus, each of the worker computers is configured to perform:

obtaining, by each of the worker computers, a respective portion of the corpus through distributed reading; and

wherein the obtaining a word and at least one context word of the word that are extracted from a corpus comprises:

creating word pairs according to the respective portion of the corpus, wherein each of the word pairs comprises a current word and a context word of the current word.

8. The cluster according to claim 7, wherein the obtaining word vectors for the word and the at least one context word comprises:

obtaining a current word set and a context word set through extraction according to the word pairs; and

obtaining word vectors for words included in the current word set and the context word set from the server.

9. The cluster according to claim 6, wherein the calculating the gradient according to the word, the at least one context word, and the word vectors comprises:

calculating, by one or more threads on the worker computer, the gradient in a manner of asynchronous calculation and unlocked updating.

10. The cluster according to claim 6, wherein the asynchronously updating the gradient comprises:

sending, by the worker computer, the updated gradient to the server without performing model averaging.

11. A non-transitory computer-readable storage medium storing instructions executable by one or more processors to cause the one or more processors to perform at a cluster comprising a server and a worker computer cluster including a plurality of worker computers:

operations at each of the worker computers:

obtaining a word and at least one context word of the word that are extracted from a corpus;

obtaining word vectors for the word and the at least one context word;

calculating a gradient according to the word, the at least one context word, and the word vectors;

asynchronously updating the gradient by sending, without waiting for a peer worker computer to send a gradient to the server, the updated gradient to the server to enable the server to update the word vectors for the word and the at least one context word according to the updated gradient, wherein the server updates the word vectors for the word and the at least one context word, by:

iteratively updating the word vectors for the word and the at least one context word, and a word vector of a negative sample word of the word according to the following formula:

$$\vec{w}_{t+1} = \vec{w}_t - \alpha \cdot \nabla \cdot \vec{c}_t, w \in B_k \text{ and}$$

$$\vec{c}_{t+1} = \vec{c}_t - \alpha \cdot \nabla \cdot \vec{w}_t, c \in \Gamma(w),$$

wherein $\nabla = \sigma(\vec{w} \cdot \vec{c}) - y$,

$$y = \begin{cases} 1, \{w, c\} \\ 0, \{w, c'\} \end{cases}$$

w represents the word, c represents a context word of the at least one context word of w , c' represents the negative sample word, $\vec{w}$ represents the word vector of w , $\vec{c}$ represents the word vector of c , $\vec{w}_t$ and $\vec{c}_t$ represent a t^{th} update on the server, B_k represents a k^{th} corpus on the worker computer, $\Gamma(w)$ represents a set of the context word and the negative sample word of w , α represents a learning rate, and σ is a sigmoid function.

12. The non-transitory computer-readable storage medium according to claim 11, wherein before the obtaining the word and the at least one context word of the word that are extracted from the corpus, the operations at each of worker computers further comprise:

obtaining, by each of the worker computers, a respective portion of the corpus through distributed reading; and

wherein the obtaining the word and the at least one context word of the word that are extracted from the corpus comprises:

creating word pairs according to the respective portion of the corpus, wherein each of the word pairs comprises a current word and a context word of the current word.

13. The non-transitory computer-readable storage medium according to claim 12, wherein the obtaining word vectors for the word and the at least one context word comprises:

obtaining a current word set and a context word set through extraction according to the word pairs; and

obtaining word vectors for words included in the current word set and the context word set from the server.

14. The non-transitory computer-readable storage medium according to claim 11, wherein the calculating a gradient according to the word, the at least one context word, and the word vectors comprises:

calculating, by one or more threads on the worker computer, the gradient in a manner of asynchronous calculation and unlocked updating.

15. The non-transitory computer-readable storage medium according to claim 11, wherein the asynchronously updating the gradient comprises:

sending, by the worker computer, the updated gradient to the server without performing model averaging.