



(12)发明专利申请

(10)申请公布号 CN 108269122 A

(43)申请公布日 2018.07.10

(21)申请号 201711480255.2

(22)申请日 2017.12.29

(71)申请人 广东神马搜索科技有限公司

地址 510627 广东省广州市天河区黄埔大道西平云路163号广电平云广场B塔13层自编01单元

(72)发明人 刘夏龙

(74)专利代理机构 北京同立钧成知识产权代理有限公司 11205

代理人 张子青 刘芳

(51)Int.Cl.

G06Q 30/02(2012.01)

G06K 9/62(2006.01)

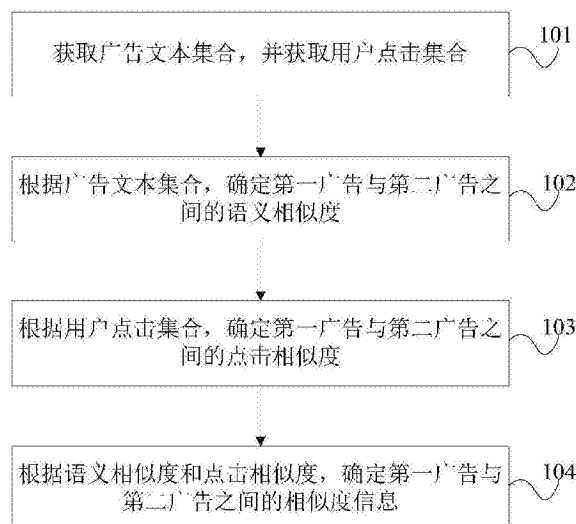
权利要求书4页 说明书16页 附图4页

(54)发明名称

广告的相似度处理方法和装置

(57)摘要

本发明提供一种广告的相似度处理方法和装置,该方法包括:获取广告文本集合,其中,广告文本集合中包括多个广告的广告文本整体的特征信息、多个广告中的每一个词语的特征信息,并获取用户点击集合,其中,用户点击集合中包括用户点击过的多个广告的特征信息,多个广告中具有第一广告和第二广告;根据广告文本集合,确定第一广告与第二广告之间的语义相似度;根据用户点击集合,确定第一广告与第二广告之间的点击相似度;根据语义相似度和点击相似度,确定第一广告与第二广告之间的相似度信息。确定出所有广告之间相似度,从而在向用户推送广告的时候,可以将相似的广告推送给用户。



1. 一种广告的相似度处理方法,其特征在于,包括:

获取广告文本集合,其中,所述广告文本集合中包括第一广告的广告文本整体的特征信息、所述第一广告中的每一个词语的特征信息、第二广告的广告文本整体的特征信息、所述第二广告中的每一个词语的特征信息、至少一个其他广告的广告文本整体的特征信息、以及所述至少一个其他广告的每一个其他广告中的每一个词语的特征信息,并获取用户点击集合,其中,所述用户点击集合中包括所述第一广告的特征信息、所述第二广告的特征信息、以及至少一个用户点击过的其他广告的特征信息;

根据所述广告文本集合,确定所述第一广告与所述第二广告之间的语义相似度;

根据所述用户点击集合,确定所述第一广告与所述第二广告之间的点击相似度;

根据所述语义相似度和所述点击相似度,确定所述第一广告与所述第二广告之间的相似度信息。

2. 根据权利要求1所述的方法,其特征在于,根据所述广告文本集合,确定所述第一广告与所述第二广告之间的语义相似度,包括:

根据所述广告文本集合,建立语义相似度目标函数;

对所述语义相似度目标函数进行求解,以确定在所述语义相似度目标函数的最优状态下的所述第一广告的广告文本整体的向量化表示、以及所述第二广告的广告文本整体的向量化表示;

根据所述第一广告的广告文本整体的向量化表示、以及所述第二广告的广告文本整体的向量化表示,确定所述语义相似度。

3. 根据权利要求2所述的方法,其特征在于,所述根据所述广告文本集合,建立语义相似度目标函数,包括:

根据所述广告文本集合,建立所述广告文本集合中的第 w_t 个特征信息的第一预设函数 $y_{w_t} = b + U * h(w_{t-k}, \dots, w_{t+k}; W)$, 其中, b 表示预设的偏差值, U 表示预设的参数向量, $h(w_{t-k}, \dots, w_{t+k}; W)$ 表示形式化函数, W 表示所述广告文本集合中的第 w_t 个特征信息, w_{t-k} 表示所述广告文本集合中第 $t-k$ 个特征信息, w_{t+k} 表示所述广告文本集合中第 $t+k$ 个特征信息, k 表示待建立的所述语义相似度目标函数的窗口大小, $t \in [k, T]$, T 表示所述广告文本集合中的特征信息的个数总和, k, t, T 都是正整数;

根据所述广告文本集合,建立第一概率分布函数 $p(w_i | w_{t-k}, \dots, w_{t+k}) = (e^{y_{w_i}}) / \left(\sum_i^{t+k} e^{y_{w_i}} \right)$,

其中, $i \in [t-k, t+k]$, i 是正整数; w_t 表示所述广告文本集合中第 t 个特征信息;

根据所述广告文本集合中的第 w_t 个特征信息的第一预设函数,以及所述第一概率分布函数,建立所述语义相似度目标函数 $x = \frac{1}{T} \sum_{t=k}^{T-k} \log p(w_i | w_{t-k}, \dots, w_{t+k})$ 。

4. 根据权利要求2所述的方法,其特征在于,所述语义相似度为 $\frac{A \bullet B}{\|A\|_2 \|B\|_2}$, 其中, A 表示所述第一广告的广告文本整体的向量化表示, B 表示所述第二广告的广告文本整体的向量化表示。

5. 根据权利要求1所述的方法,其特征在于,根据所述用户点击集合,确定所述第一广

告与所述第二广告之间的点击相似度,包括:

根据所述用户点击集合,建立点击相似度目标函数;

对所述点击相似度目标函数进行求解,以确定在所述点击相似度目标函数的最优状态下的所述第一广告的向量化表示、以及所述第二广告的向量化表示;

根据所述第一广告的向量化表示、以及所述第二广告的向量化表示,确定所述点击相似度。

6. 根据权利要求5所述的方法,其特征在于,根据所述用户点击集合,建立点击相似度目标函数,包括:

根据所述用户点击集合,建立所述用户点击集合中的第 $w't'$ 个广告的特征信息的第二预设函数 $y'_{w'_t} = b' + U' * h'(w'_{t'-k'}, \dots, w'_{t'+k'}; W')$,其中, b' 表示预设的第二偏差值, U' 表示预设的第二参数向量, $h'(w'_{t'-k'}, \dots, w'_{t'+k'}; W')$ 表示形式化函数, W' 表示所述用户点击集合中的第 $w't'$ 个广告的特征信息, $w'_{t'-k'}$ 表示所述用户点击集合中第 $t'-k'$ 个广告的特征信息, $w'_{t'+k'}$ 表示所述用户点击集合中第 $t'+k'$ 个广告的特征信息, k' 表示待建立的所述点击相似度目标函数的窗口大小, $t' \in [k', T']$, T' 表示所述用户点击集合中的广告个数总和, k' 、 t' 、 T' 都是正整数;

根据所述用户点击集合,建立第二概率分布函数 $p(w'_t | w'_{t'-k'}, \dots, w'_{t'+k'}) = (e^{y'_{w'_t}}) / (\sum_{t'=k'}^{t'+k'} e^{y'_{w'_t}})$,其中, $i' \in [t'-k', t'+k']$, i' 是正整数; $w'_{t'}$ 表示所述用户点击集合中第 t' 个广告的特征信息;

根据所述用户点击集合中的第 $w't'$ 个广告的特征信息的第二预设函数,以及所述第二概率分布函数,建立所述点击相似度目标函数 $x' = \frac{1}{T'} \sum_{t'=k'}^{T'-k'} \log p(w'_t | w'_{t'-k'}, \dots, w'_{t'+k'})$ 。

7. 根据权利要求5所述的方法,其特征在于,所述点击相似度为 $\frac{C \cdot D}{\|C\|_2 \|D\|_2}$,其中, C 表示所述第一广告的向量化表示, D 表示所述第二广告的向量化表示。

8. 根据权利要求1-7任一项所述的方法,其特征在于,根据所述语义相似度和所述点击相似度,确定所述第一广告与所述第二广告之间的相似度信息,包括:

获取所述第二广告的用户点击频次;

根据所述用户点击频次、所述语义相似度和所述点击相似度,确定所述相似度信息。

9. 根据权利要求8所述的方法,其特征在于,所述相似度信息为 $\text{Sim} = (1/\log(\text{TF})) * \text{Sim}_{\text{content}} + \text{Sim}_{\text{session}}$;

其中, TF 表示所述用户点击频次, $\text{Sim}_{\text{content}}$ 表示所述语义相似度, $\text{Sim}_{\text{session}}$ 表示所述点击相似度。

10. 一种广告的相似度处理装置,其特征在于,包括:

获取单元,用于获取广告文本集合,其中,所述广告文本集合中包括第一广告的广告文本整体的特征信息、所述第一广告中的每一个词语的特征信息、第二广告的广告文本整体的特征信息、所述第二广告中的每一个词语的特征信息、至少一个其他广告的广告文本整体的特征信息、以及所述至少一个其他广告的每一个其他广告中的每一个词语的特征信

息,并获取用户点击集合,其中,所述用户点击集合中包括所述第一广告的特征信息、所述第二广告的特征信息、以及至少一个用户点击过的其他广告的特征信息;

第一确定单元,用于根据所述广告文本集合,确定所述第一广告与所述第二广告之间的语义相似度;

第二确定单元,用于根据所述用户点击集合,确定所述第一广告与所述第二广告之间的点击相似度;

第三确定单元,用于根据所述语义相似度和所述点击相似度,确定所述第一广告与所述第二广告之间的相似度信息。

11. 根据权利要求10所述的装置,其特征在于,所述第一确定单元,包括:

第一建立模块,用于根据所述广告文本集合,建立语义相似度目标函数;

第一求解模块,用于对所述语义相似度目标函数进行求解,以确定在所述语义相似度目标函数的最优状态下的所述第一广告的广告文本整体的向量化表示、以及所述第二广告的广告文本整体的向量化表示;

第一确定模块,用于根据所述第一广告的广告文本整体的向量化表示、以及所述第二广告的广告文本整体的向量化表示,确定所述语义相似度。

12. 根据权利要求11所述的装置,其特征在于,所述第一建立模块,包括:

第一建立子模块,用于根据所述广告文本集合,建立所述广告文本集合中的第 w_t 个特征信息的第一预设函数 $y_{w_t} = b + U * h(w_{t-k}, \dots, w_{t+k}; W)$,其中, b 表示预设的偏差值, U 表示预设的参数向量, $h(w_{t-k}, \dots, w_{t+k}; W)$ 表示形式化函数, W 表示所述广告文本集合中的第 w_t 个特征信息, w_{t-k} 表示所述广告文本集合中第 $t-k$ 个特征信息, w_{t+k} 表示所述广告文本集合中第 $t+k$ 个特征信息, k 表示待建立的所述语义相似度目标函数的窗口大小, $t \in [k, T]$, T 表示所述广告文本集合中的特征信息的个数总和, k 、 t 、 T 都是正整数;

第二建立子模块,用于根据所述广告文本集合,建立第一概率分布函数

$p(w_i | w_{i-k}, \dots, w_{i+k}) = (e^{y_{w_i}}) / \left(\sum_i e^{y_{w_i}} \right)$,其中, $i \in [t-k, t+k]$, i 是正整数; w_t 表示所述广告文本集合中第 t 个特征信息;

第三建立子模块,用于根据所述广告文本集合中的第 w_t 个特征信息的第一预设函数,以及所述第一概率分布函数,建立所述语义相似度目标函数 $x = \frac{1}{T} \sum_{i=k}^{T-k} \log p(w_i | w_{i-k}, \dots, w_{i+k})$ 。

13. 根据权利要求11所述的装置,其特征在于,所述语义相似度为 $\frac{A \bullet B}{\|A\|_2 \|B\|_2}$,其中, A 表示所述第一广告的广告文本整体的向量化表示, B 表示所述第二广告的广告文本整体的向量化表示。

14. 根据权利要求10所述的装置,其特征在于,所述第二确定单元,包括:

第二建立模块,用于根据所述用户点击集合,建立点击相似度目标函数;

第二求解模块,用于对所述点击相似度目标函数进行求解,以确定在所述点击相似度目标函数的最优状态下的所述第一广告的向量化表示、以及所述第二广告的向量化表示;

第二确定模块,用于根据所述第一广告的向量化表示、以及所述第二广告的向量化表

示,确定所述点击相似度。

15. 根据权利要求14所述的装置,其特征在于,所述第二建立模块,包括:

第四建立子模块,用于根据所述用户点击集合,建立所述用户点击集合中的第 $w't'$ 个广告的特征信息的第二预设函数 $y'_{w'_t} = b' + U' * h'(w'_{t'-k'}, \dots, w'_{t'+k'}; W')$,其中, b' 表示预设的第二偏差值, U' 表示预设的第二参数向量, $h'(w'_{t'-k'}, \dots, w'_{t'+k'}; W')$ 表示形式化函数, W' 表示所述用户点击集合中的第 $w't'$ 个广告的特征信息, $w'_{t'-k'}$ 表示所述用户点击集合中第 $t'-k'$ 个广告的特征信息, $w'_{t'+k'}$ 表示所述用户点击集合中第 $t'+k'$ 个广告的特征信息, k' 表示待建立的所述点击相似度目标函数的窗口大小, $t' \in [k', T']$, T' 表示所述用户点击集合中的广告个数总和, k' 、 t' 、 T' 都是正整数;

第五建立子模块,用于根据所述用户点击集合,建立第二概率分布函数 $p(w'_t | w'_{t-k'}, \dots, w'_{t+k'}) = (e^{y'_{w'_t}}) / \left(\sum_{i'} e^{y'_{w'_i}} \right)$,其中, $i' \in [t'-k', t'+k']$, i' 是正整数; w'_t 表示所述用户点击集合中第 t' 个广告的特征信息;

第六建立子模块,用于根据所述用户点击集合中的第 $w't'$ 个广告的特征信息的第二预设函数,以及所述第二概率分布函数,建立所述点击相似度目标函数 $x' = \frac{1}{T'} \sum_{t'=k'}^{T'-k'} \log p(w'_t | w'_{t-k'}, \dots, w'_{t+k'})$ 。

16. 根据权利要求14所述的装置,其特征在于,所述点击相似度为 $\frac{C \cdot D}{\|C\|_2 \|D\|_2}$,其中, C 表示所述第一广告的向量化表示, D 表示所述第二广告的向量化表示。

17. 根据权利要求10-16任一项所述的装置,其特征在于,所述第三确定单元,包括:

获取模块,用于获取所述第二广告的用户点击频次;

第三确定模块,用于根据所述用户点击频次、所述语义相似度和所述点击相似度,确定所述相似度信息。

18. 根据权利要求17所述的装置,其特征在于,所述相似度信息为 $\text{Sim} = (1/\log(\text{TF})) * \text{Sim}_{\text{content}} + \text{Sim}_{\text{session}}$;

其中, TF 表示所述用户点击频次, $\text{Sim}_{\text{content}}$ 表示所述语义相似度, $\text{Sim}_{\text{session}}$ 表示所述点击相似度。

广告的相似度处理方法和装置

技术领域

[0001] 本发明涉及广告技术领域,尤其涉及一种广告的相似度处理方法和装置。

背景技术

[0002] 随着媒体技术的发展,广告也越来越多的应用到媒体技术中。广告作为推销产品的一种重要手段被广泛使用;在投放广告的时候,需要考虑到广告之间相似性,以便于向用户的终端推送相似产品的广告,进而便于用户获知更多的产品信息。

[0003] 现有技术中,在分析广告之间的相似性的时候,一般是获取到广告的关键词信息,然后根据广告的关键词信息确定广告之间是否相似,然后将相似的广告推送给适合的用户群体。

[0004] 然而现有技术中,由于广告用户的不断变化以及广告行文的复杂性,进而在分析广告之间的相似性的时候,容易提取出错误的关键词信息,进而分析出的广告之间的相似性并不准确,进一步的,推送给用户群体的广告并不是相似的广告,进而推送广告错误。

发明内容

[0005] 本发明提供一种广告的相似度处理方法和装置,用以解决分析出的广告之间的相似性并不准确的问题。

[0006] 一方面,本发明提供一种广告的相似度处理方法,包括:

[0007] 获取广告文本集合,其中,所述广告文本集合中包括第一广告的广告文本整体的特征信息、所述第一广告中的每一个词语的特征信息、第二广告的广告文本整体的特征信息、所述第二广告中的每一个词语的特征信息、至少一个其他广告的广告文本整体的特征信息、以及所述至少一个其他广告的每一个其他广告中的每一个词语的特征信息,并获取用户点击集合,其中,所述用户点击集合中包括所述第一广告的特征信息、所述第二广告的特征信息、以及至少一个用户点击过的其他广告的特征信息;

[0008] 根据所述广告文本集合,确定所述第一广告与所述第二广告之间的语义相似度;

[0009] 根据所述用户点击集合,确定所述第一广告与所述第二广告之间的点击相似度;

[0010] 根据所述语义相似度和所述点击相似度,确定所述第一广告与所述第二广告之间的相似度信息。

[0011] 进一步地,根据所述广告文本集合,确定所述第一广告与所述第二广告之间的语义相似度,包括:

[0012] 根据所述广告文本集合,建立语义相似度目标函数;

[0013] 对所述语义相似度目标函数进行求解,以确定在所述语义相似度目标函数的最优状态下的所述第一广告的广告文本整体的向量化表示、以及所述第二广告的广告文本整体的向量化表示;

[0014] 根据所述第一广告的广告文本整体的向量化表示、以及所述第二广告的广告文本整体的向量化表示,确定所述语义相似度。

[0015] 进一步地,所述根据所述广告文本集合,建立语义相似度目标函数,包括:

[0016] 根据所述广告文本集合,建立所述广告文本集合中的第 w_t 个特征信息的第一预设函数 $y_{w_t} = b + U * h(w_{t-k}, \dots, w_{t+k}; W)$,其中, b 表示预设的偏差值, U 表示预设的参数向量, $h(w_{t-k}, \dots, w_{t+k}; W)$ 表示形式化函数, W 表示所述广告文本集合中的第 w_t 个特征信息, w_{t-k} 表示所述广告文本集合中第 $t-k$ 个特征信息, w_{t+k} 表示所述广告文本集合中第 $t+k$ 个特征信息, k 表示待建立的所述语义相似度目标函数的窗口大小, $t \in [k, T]$, T 表示所述广告文本集合中的特征信息的个数总和, k 、 t 、 T 都是正整数;

[0017] 根据所述广告文本集合,建立第一概率分布函数 $p(w_t | w_{t-k}, \dots, w_{t+k}) = (e^{y_{w_t}}) / (\sum_i e^{y_{w_i}})$,

其中, $i \in [t-k, t+k]$, i 是正整数; w_t 表示所述广告文本集合中第 t 个特征信息;

[0018] 根据所述广告文本集合中的第 w_t 个特征信息的第一预设函数,以及所述第一概率分布函数,建立所述语义相似度目标函数 $x = \frac{1}{T} \sum_{t=k}^{T-k} \log p(w_t | w_{t-k}, \dots, w_{t+k})$ 。

[0019] 进一步地,所述语义相似度为 $\frac{A \cdot B}{\|A\|_2 \|B\|_2}$,其中, A 表示所述第一广告的广告文本整体的向量化表示, B 表示所述第二广告的广告文本整体的向量化表示。

[0020] 进一步地,根据所述用户点击集合,确定所述第一广告与所述第二广告之间的点击相似度,包括:

[0021] 根据所述用户点击集合,建立点击相似度目标函数;

[0022] 对所述点击相似度目标函数进行求解,以确定在所述点击相似度目标函数的最优状态下的所述第一广告的向量化表示、以及所述第二广告的向量化表示;

[0023] 根据所述第一广告的向量化表示、以及所述第二广告的向量化表示,确定所述点击相似度。

[0024] 进一步地,根据所述用户点击集合,建立点击相似度目标函数,包括:

[0025] 根据所述用户点击集合,建立所述用户点击集合中的第 w'_t 个广告的特征信息的第二预设函数 $y'_{w'_t} = b' + U' * h'(w'_{t'-k'}, \dots, w'_{t'+k'}; W')$,其中, b' 表示预设的第二偏差值, U' 表示预设的第二参数向量, $h'(w'_{t'-k'}, \dots, w'_{t'+k'}; W')$ 表示形式化函数, W' 表示所述用户点击集合中的第 w'_t 个广告的特征信息, $w'_{t'-k'}$ 表示所述用户点击集合中第 $t'-k'$ 个广告的特征信息, $w'_{t'+k'}$ 表示所述用户点击集合中第 $t'+k'$ 个广告的特征信息, k' 表示待建立的所述点击相似度目标函数的窗口大小, $t' \in [k', T']$, T' 表示所述用户点击集合中的广告的个数总和, k' 、 t' 、 T' 都是正整数;

[0026] 根据所述用户点击集合,建立第二概率分布函数 $p(w'_t | w'_{t'-k'}, \dots, w'_{t'+k'}) = (e^{y'_{w'_t}}) / (\sum_{i'} e^{y'_{w'_{i'}}})$,

其中, $i' \in [t'-k', t'+k']$, i' 是正整数; w'_t 表示所述用户点击集合中第 t' 个广告的特征信息;

[0027] 根据所述用户点击集合中的第 w'_t 个广告的特征信息的第二预设函数,以及所述第二概率分布函数,建立所述点击相似度目标函数 $x' = \frac{1}{T'} \sum_{t'=k'}^{T'-k'} \log p(w'_t | w'_{t'-k'}, \dots, w'_{t'+k'})$ 。

[0028] 进一步地,所述点击相似度为 $\frac{C \cdot D}{\|C\|_2 \|D\|_2}$,其中,C表示所述第一广告的向量化表示,D表示所述第二广告的向量化表示。

[0029] 进一步地,根据所述语义相似度和所述点击相似度,确定所述第一广告与所述第二广告之间的相似度信息,包括:

[0030] 获取所述第二广告的用户点击频次;

[0031] 根据所述用户点击频次、所述语义相似度和所述点击相似度,确定所述相似度信息。

[0032] 进一步地,所述相似度信息为 $\text{Sim} = (1/\log(\text{TF})) * \text{Sim}_{\text{content}} + \text{Sim}_{\text{session}}$;

[0033] 其中,TF表示所述用户点击频次, $\text{Sim}_{\text{content}}$ 表示所述语义相似度, $\text{Sim}_{\text{session}}$ 表示所述点击相似度。

[0034] 另一方面,本发明提供一种广告的相似度处理装置,包括:

[0035] 获取单元,用于获取广告文本集合,其中,所述广告文本集合中包括第一广告的广告文本整体的特征信息、所述第一广告中的每一个词语的特征信息、第二广告的广告文本整体的特征信息、所述第二广告中的每一个词语的特征信息、至少一个其他广告的广告文本整体的特征信息、以及所述至少一个其他广告的每一个其他广告中的每一个词语的特征信息,并获取用户点击集合,其中,所述用户点击集合中包括所述第一广告的特征信息、所述第二广告的特征信息、以及至少一个用户点击过的其他广告的特征信息;

[0036] 第一确定单元,用于根据所述广告文本集合,确定所述第一广告与所述第二广告之间的语义相似度;

[0037] 第二确定单元,用于根据所述用户点击集合,确定所述第一广告与所述第二广告之间的点击相似度;

[0038] 第三确定单元,用于根据所述语义相似度和所述点击相似度,确定所述第一广告与所述第二广告之间的相似度信息。

[0039] 进一步地,所述第一确定单元,包括:

[0040] 第一建立模块,用于根据所述广告文本集合,建立语义相似度目标函数;

[0041] 第一求解模块,用于对所述语义相似度目标函数进行求解,以确定在所述语义相似度目标函数的最优状态下的所述第一广告的广告文本整体的向量化表示、以及所述第二广告的广告文本整体的向量化表示;

[0042] 第一确定模块,用于根据所述第一广告的广告文本整体的向量化表示、以及所述第二广告的广告文本整体的向量化表示,确定所述语义相似度。

[0043] 进一步地,所述第一建立模块,包括:

[0044] 第一建立子模块,用于根据所述广告文本集合,建立所述广告文本集合中的第 w_t 个特征信息的第一预设函数 $y_{w_t} = b + U * h(w_{t-k}, \dots, w_{t+k}; W)$,其中,b表示预设的偏差值,U表示预设的参数向量, $h(w_{t-k}, \dots, w_{t+k}; W)$ 表示形式化函数,W表示所述广告文本集合中的第 w_t 个特征信息, w_{t-k} 表示所述广告文本集合中第 $t-k$ 个特征信息, w_{t+k} 表示所述广告文本集合中第 $t+k$ 个特征信息,k表示待建立的所述语义相似度目标函数的窗口大小, $t \in [k, T]$,T表示所述广告文本集合中的特征信息的个数总和,k、t、T都是正整数;

[0045] 第二建立子模块,用于根据所述广告文本集合,建立第一概率分布函数

$p(w_t | w_{t-k}, \dots, w_{t+k}) = (e^{y_{w_t}}) / \left(\sum_i^{t+k} e^{y_{w_i}} \right)$, 其中, $i \in [t-k, t+k]$, i 是正整数; w_t 表示所述广告文本集合中第 t 个特征信息;

[0046] 第三建立子模块, 用于根据所述广告文本集合中的第 w_t 个特征信息的第一预设函数, 以及所述第一概率分布函数, 建立所述语义相似度目标函数 $x = \frac{1}{T} \sum_{t=k}^{T-k} \log p(w_t | w_{t-k}, \dots, w_{t+k})$ 。

[0047] 进一步地, 所述语义相似度为 $\frac{A \cdot B}{\|A\|_2 \|B\|_2}$, 其中, A 表示所述第一广告的广告文本整体的向量化表示, B 表示所述第二广告的广告文本整体的向量化表示。

[0048] 进一步地, 所述第二确定单元, 包括:

[0049] 第二建立模块, 用于根据所述用户点击集合, 建立点击相似度目标函数;

[0050] 第二求解模块, 用于对所述点击相似度目标函数进行求解, 以确定在所述点击相似度目标函数的最优状态下的所述第一广告的向量化表示、以及所述第二广告的向量化表示;

[0051] 第二确定模块, 用于根据所述第一广告的向量化表示、以及所述第二广告的向量化表示, 确定所述点击相似度。

[0052] 进一步地, 所述第二建立模块, 包括:

[0053] 第四建立子模块, 用于根据所述用户点击集合, 建立所述用户点击集合中的第 w'_t 个广告的特征信息的第二预设函数 $y_{w'_t} = b' + U' * h'(w'_{t'-k'}, \dots, w'_{t'+k'}; W')$, 其中, b' 表示预设的第二偏差值, U' 表示预设的第二参数向量, $h'(w'_{t'-k'}, \dots, w'_{t'+k'}; W')$ 表示形式化函数, W' 表示所述用户点击集合中的第 w'_t 个广告的特征信息, $w'_{t'-k'}$ 表示所述用户点击集合中第 $t'-k'$ 个广告的特征信息, $w'_{t'+k'}$ 表示所述用户点击集合中第 $t'+k'$ 个广告的特征信息, k' 表示待建立的所述点击相似度目标函数的窗口大小, $t' \in [k', T']$, T' 表示所述用户点击集合中的广告个数总和, k' 、 t' 、 T' 都是正整数;

[0054] 第五建立子模块, 用于根据所述用户点击集合, 建立第二概率分布函数

$p(w'_t | w'_{t'-k'}, \dots, w'_{t'+k'}) = (e^{y_{w'_t}}) / \left(\sum_i^{t'+k'} e^{y_{w'_i}} \right)$, 其中, $i' \in [t'-k', t'+k']$, i' 是正整数; w'_t 表示所述用户点击集合中第 t' 个广告的特征信息;

[0055] 第六建立子模块, 用于根据所述用户点击集合中的第 w'_t 个广告的特征信息的第二预设函数, 以及所述第二概率分布函数, 建立所述点击相似度目标函数

$x' = \frac{1}{T'} \sum_{t'=k'}^{T'-k'} \log p(w'_t | w'_{t'-k'}, \dots, w'_{t'+k'})$ 。

[0056] 进一步地, 所述点击相似度为 $\frac{C \cdot D}{\|C\|_2 \|D\|_2}$, 其中, C 表示所述第一广告的向量化表示, D 表示所述第二广告的向量化表示。

[0057] 进一步地, 所述第三确定单元, 包括:

[0058] 获取模块, 用于获取所述第二广告的用户点击频次;

[0059] 第三确定模块, 用于根据所述用户点击频次、所述语义相似度和所述点击相似度,

确定所述相似度信息。

[0060] 进一步地,所述相似度信息为 $\text{Sim} = (1/\log(\text{TF})) * \text{Sim}_{\text{content}} + \text{Sim}_{\text{session}}$;

[0061] 其中,TF表示所述用户点击频次, $\text{Sim}_{\text{content}}$ 表示所述语义相似度, $\text{Sim}_{\text{session}}$ 表示所述点击相似度。

[0062] 本发明提供的广告的相似度处理方法和装置,通过获取广告文本集合,其中,广告文本集合中包括第一广告的广告文本整体的特征信息、第一广告中的每一个词语的特征信息、第二广告的广告文本整体的特征信息、第二广告中的每一个词语的特征信息、至少一个其他广告的广告文本整体的特征信息、以及至少一个其他广告的一个其他广告中的每一个词语的特征信息,并获取用户点击集合,其中,用户点击集合中包括第一广告的特征信息、第二广告的特征信息、以及至少一个用户点击过的其他广告的特征信息;根据广告文本集合,确定第一广告与第二广告之间的语义相似度;根据用户点击集合,确定第一广告与第二广告之间的点击相似度;根据语义相似度和点击相似度,确定第一广告与第二广告之间的相似度信息。从而通过对海量的广告中的词语进行提取,根据神经网络模型对海量的广告中的词语进行分析,可以对短文本的广告和长文本的广告都进行分析,便于提取广告中的主题和关键信息;并且,可以从用户点击广告的行为的角度进行分析,去获取到属于同一群体的用户所点击的海量的广告,然后,将属于同一群体的用户所点击过的广告构成一个用户点击集合,去对用户点击集合中的所有广告的特征进行分析,有利于进行广告归类;并且以上过程都是对海量的广告数据进行的分析,可以更准确的确定出广告之间相似性;然后将根据广告文本集合计算得到的语义相似度、以及根据用户点击集合计算得到的点击相似度进行计算,计算得到第一广告与第二广告之间的相似度信息,即第二广告在多大程度上是与第一广告相似的,可以准确确定的确定出广告之间的相似度。进而可以根据以上过程确定出所有广告之间相似度,从而在向用户推送广告的时候,可以将相似的广告推送给用户。

附图说明

[0063] 此处的附图被并入说明书中并构成本说明书的一部分,示出了符合本公开的实施例,并与说明书一起用于解释本公开的原理。

[0064] 图1为本申请实施例提供的一种广告的相似度处理方法的流程示意图;

[0065] 图2为本申请实施例提供的一种广告的相似度处理方法中的点击会话日志的示意图;

[0066] 图3为本申请实施例提供的一种广告的相似度处理方法中的神经网络模型的结构示意图;

[0067] 图4为本申请实施例提供的另一种广告的相似度处理方法的流程示意图;

[0068] 图5为本发明实施例提供的一种广告的相似度处理装置的结构示意图;

[0069] 图6为本发明实施例提供的另一种广告的相似度处理装置的结构示意图。

[0070] 通过上述附图,已示出本公开明确的实施例,后文中将有更详细的描述。这些附图和文字描述并不是为了通过任何方式限制本公开构思的范围,而是通过参考特定实施例为本领域技术人员说明本公开的概念。

具体实施方式

[0071] 这里将详细地对示例性实施例进行说明,其示例表示在附图中。下面的描述涉及附图时,除非另有表示,不同附图中的相同数字表示相同或相似的要素。以下示例性实施例中所描述的实施方式并不代表与本公开相一致的所有实施方式。相反,它们仅是与如所附权利要求书中所详述的、本公开的一些方面相一致的装置和方法的例子。

[0072] 首先对本发明所涉及的名词进行解释:

[0073] Word Embedding:指的是词嵌入技术;具体来说,将词进行向量化表示,实体的抽象成了数学描述,就可以进行建模,应用到很多任务中,例如比较词语词之间的相似性,可以直接通过向量之间的余弦距离度量来确定。

[0074] 多层神经网络深度学习模型(Deep Structured Semantic Models,简称DSSM):这是一种神经网络模型,又称为sent2vec。

[0075] 随机梯度下降法(Stochastic Gradient Descent,简称SGD):是求解无约束最优化问题的一种常用方法,有实现简单的优点;随机梯度下降法是迭代算法,每一步需要求解目标函数的梯度向量。

[0076] 本发明具体的应用场景如下。随着媒体技术以及终端技术的发展,越来越多的广告需要投放到媒体技术中;可以向用户推送广告,可以根据用户特性将用户进行划分得到多个用户群体,进而向每一种用户群体推送相似的广告;或者直接向用户推送一系列相似的广告。那么如何准确的哪些广告是相似的,即广告之间的相似度,是一个需要解决的问题。

[0077] 本发明提供的广告的相似度处理方法和装置,旨在解决现有技术的如上技术问题。

[0078] 下面以具体地实施例对本发明的技术方案以及本申请的技术方案如何解决上述技术问题进行详细说明。下面这几个具体的实施例可以相互结合,对于相同或相似的概念或过程可能在某些实施例中不再赘述。下面将结合附图,对本发明的实施例进行描述。

[0079] 图1为本申请实施例提供的一种广告的相似度处理方法的流程示意图。如图1所示,该方法包括:

[0080] 步骤101、获取广告文本集合,其中,广告文本集合中包括第一广告的广告文本整体的特征信息、第一广告中的每一个词语的特征信息、第二广告的广告文本整体的特征信息、第二广告中的每一个词语的特征信息、至少一个其他广告的广告文本整体的特征信息、以及至少一个其他广告中的每一个其他广告中的每一个词语的特征信息,并获取用户点击集合,其中,用户点击集合中包括第一广告的特征信息、第二广告的特征信息、以及至少一个用户点击过的其他广告的特征信息。

[0081] 在本实施例中,具体的,本实施例的执行主体可以为广告的相似度处理装置、服务器或者其他可以执行本实施例方法的设备。

[0082] 首先,需要获取到广告主所提供的每一个广告;然后对每一个广告进行分析,进而可以将每一个广告拆分为多个词语;进而得到一个广告文本集合。在该广告文本集合中包括了多个广告中的每一个广告的广告文本整体的特征信息、以及多个广告中每一个广告的每一个词语的特征信息;其中,这多个广告中就有待分析的第一广告和第二广告。并且,每

一个广告的广告文本整体的特征信息是一个向量,并且每一个词语的特征信息也是一个向量。

[0083] 举例来说,根据一万个广告生成一个广告文本集合,广告文本集合中包括了广告1的广告文本整体的特征信息、广告1的词语1的特征信息、广告1的词语2的特征信息、广告1的词语3的特征信息、广告2的广告文本整体的特征信息、广告2的词语2的特征信息、广告2的词语3的特征信息、广告2的词语4的特征信息、广告3的广告文本整体的特征信息、广告3的词语2的特征信息、广告3的词语3的特征信息、广告3的词语4的特征信息、广告4的广告文本整体的特征信息、广告4的词语4的特征信息、广告4的词语5的特征信息、广告4的词语6的特征信息,以此类推;其中不同词语的标号代表了不同给的词语,广告1为第一广告,广告2为第二广告。需要分析广告1与广告2之间相似度。

[0084] 并且,需要获取到多个用户所点击过的广告,将多个用户所点击过的广告构成一个用户点击集合。具体来说,首先,获取每一个用户的点击会话(Click Session)日志,根据在每一个用户的Click Session日志,确定出每一个用户所点击过的广告;然后将每一个用户所点击过的广告放入到一个用户点击集合中;该用户点击集合中包括了多个用户所点击过的广告中的每一广告的特征信息,其中,这多个用户所点击过的广告中就有待分析的第一广告和第二广告。可知,第一广告和第二广告也是用户所点击过的广告。其中,每一个广告的特征信息为一个向量。例如,由于具有同样兴趣的用户群体对广告的点击也是有偏好的,进而被属于同一群体的用户点击过的广告也反映了广告本身的相似性;进而可以去获取属于同一群体的用户所点击过的广告,将属于同一群体的用户所点击过的广告构成一个用户点击集合,然后,对这些广告进行画像和归类。图2为本申请实施例提供的一种广告的相似度处理方法中的点击会话日志的示意图,如图2所示,通过分析用户的电机行为,可以在获取到用户曾经点击过的广告的内容。并且,获取用户海量的广告点击行为,其中,每一次广告点击行为对应着一个广告,这些海量的点击行为可以避免广告之间的噪声偏差问题。

[0085] 举例来说,可以获取到属于同一群体的用户所点击过的一万个广告,将这一万个广告构成一个用户点击集合,用户点击集合中包括了广告1的特征信息、广告2的特征信息、广告3的特征信息、广告4的特征信息,以此类推;其中,广告1为第一广告,广告2为第二广告,需要分析广告1与广告2之间相似度。

[0086] 步骤102、根据广告文本集合,确定第一广告与第二广告之间的语义相似度。

[0087] 在本实施例中,具体的,根据神经网络模型以及Word Embedding技术,对广告文本集合中每一广告的广告文本整体的特征信息、以及每一个广告中的每一个词语的特征信息进行分析,由于广告文本集合中包括了待分析的第一广告与第二广告,进而可以确定出第一广告与第二广告之间的语义相似度。在本实施例中,该语义相似度表征了第二广告像第一广告的程度是多少。

[0088] 图3为本申请实施例提供的一种广告的相似度处理方法中的神经网络模型的结构示意图,如图3所示,该神经网络模型中第一层为分类器(Classifier);该神经网络模型中第二层为平均化/联接(Average/Concatenate)层,该平均化/联接层代表的是下层网络到上层网络的一种连接形式;该神经网络模型中最后一层表征的是广告矩阵(Paragraph matrix),即所有广告的向量化表示,例如,D表示某一个广告,Paragraph是段落的意思,

Paragraph代指一个广告,W是每一个广告里面的词(Word)的前缀。

[0089] 步骤103、根据用户点击集合,确定第一广告与第二广告之间的点击相似度。

[0090] 在本实施例中,具体的,采用神经网络算法以及Word Embedding技术,对用户点击集合进行建模处理,其中,神经网络算法中具有连续词袋模型(Continuous Bag of Words,简称Cbow)以及skip-gram结构,在这里,神经网络算法可以采用skip-gram结构;进而对每一广告的特征信息进行分析,得到第一广告与第二广告之间的点击相似度。在本实施例中,该点击相似度表征了第二广告像第一广告的程度是多少。

[0091] 步骤104、根据语义相似度和点击相似度,确定第一广告与第二广告之间的相似度信息。

[0092] 在可选的一种实施方式中,步骤104具体包括:根据语义相似度和点击相似度,确定第一广告与第二广告之间的相似度信息,包括:获取第二广告的用户点击频次;根据用户点击频次、语义相似度和点击相似度,确定相似度信息。

[0093] 在可选的一种实施方式中,相似度信息为 $\text{Sim} = (1/\log(\text{TF})) * \text{Sim}_{\text{content}} + \text{Sim}_{\text{session}}$;其中,TF表示用户点击频次, $\text{Sim}_{\text{content}}$ 表示语义相似度, $\text{Sim}_{\text{session}}$ 表示点击相似度。

[0094] 在本实施例中,具体的,可以依据计算得到的语义相似度和点击相似度,去计算相似度信息。具体来说,由于需要计算第二广告有多像第一广告,进而首先获取到第二广告的用户点击频次TF,该用户点击频次TF为第二广告被用户点击的次数;然后,根据用户点击频次TF、语义相似度 $\text{Sim}_{\text{content}}$ 以及点击相似度 $\text{Sim}_{\text{session}}$,计算出第一广告与第二广告之间的相似度信息,相似度信息的计算公式可以有多种,本实施例提供了一种优选的计算方式,可以得到相似度信息为 $\text{Sim} = (1/\log(\text{TF})) * \text{Sim}_{\text{content}} + \text{Sim}_{\text{session}}$ 。

[0095] 本实施例通过获取广告文本集合,其中,广告文本集合中包括第一广告的广告文本整体的特征信息、第一广告中的每一个词语的特征信息、第二广告的广告文本整体的特征信息、第二广告中的每一个词语的特征信息、至少一个其他广告的广告文本整体的特征信息、以及至少一个其他广告的每一个其他广告中的每一个词语的特征信息,并获取用户点击集合,其中,用户点击集合中包括第一广告的特征信息、第二广告的特征信息、以及至少一个用户点击过的其他广告的特征信息;根据广告文本集合,确定第一广告与第二广告之间的语义相似度;根据用户点击集合,确定第一广告与第二广告之间的点击相似度;根据语义相似度和点击相似度,确定第一广告与第二广告之间的相似度信息。从而通过对海量的广告中的词语进行提取,根据神经网络模型对海量的广告中的词语进行分析,可以对短文本的广告和长文本的广告都进行分析,便于提取广告中的主题和关键信息;并且,可以从用户点击广告的行为的角度进行分析,去获取到属于同一群体的用户所点击的海量的广告,然后,将属于同一群体的用户所点击过的广告构成一个用户点击集合,去对用户点击集合中的所有广告的特征进行分析,有利于进行广告归类;并且以上过程都是对海量的广告数据进行的分析,可以更准确的确定出广告之间相似性;然后将根据广告文本集合计算得到的语义相似度、以及根据用户点击集合计算得到的点击相似度进行计算,计算得到第一广告与第二广告之间的相似度信息,即第二广告在多大程度上是与第一广告相似的,可以准确确定的确定出广告之间的相似度。进而可以根据以上过程确定出所有广告之间相似度,从而在向用户推送广告的时候,可以将相似的广告推送给用户。

[0096] 图4为本申请实施例提供的另一种广告的相似度处理方法的流程示意图。如图4所

示,该方法包括:

[0097] 步骤201、获取广告文本集合,其中,广告文本集合中包括第一广告的广告文本整体的特征信息、第一广告中的每一个词语的特征信息、第二广告的广告文本整体的特征信息、第二广告中的每一个词语的特征信息、至少一个其他广告的广告文本整体的特征信息、以及至少一个其他广告的每一个其他广告中的每一个词语的特征信息,并获取用户点击集合,其中,用户点击集合中包括第一广告的特征信息、第二广告的特征信息、以及至少一个用户点击过的其他广告的特征信息。

[0098] 在本实施例中,具体的,本实施例的执行主体可以为广告的相似度处理装置、服务器或者其他可以执行本实施例方法的设备。本步骤可以参见图1的步骤101不再赘述。

[0099] 步骤202、根据广告文本集合,建立语义相似度目标函数。

[0100] 在可选的一种实施方式中,步骤202具体包括以下步骤:

[0101] 步骤2021、根据广告文本集合,建立广告文本集合中的第 w_t 个特征信息的第一预设函数 $y_{w_t} = b + U * h(w_{t-k}, \dots, w_{t+k}; W)$,其中, b 表示预设的偏差值, U 表示预设的参数向量, $h(w_{t-k}, \dots, w_{t+k}; W)$ 表示形式化函数, W 表示广告文本集合中的第 w_t 个特征信息, w_{t-k} 表示广告文本集合中第 $t-k$ 个特征信息, w_{t+k} 表示广告文本集合中第 $t+k$ 个特征信息, k 表示待建立的语义相似度目标函数的窗口大小, $t \in [k, T]$, T 表示广告文本集合中的特征信息的个数总和, k 、 t 、 T 都是正整数。

[0102] 步骤2022、根据广告文本集合,建立第一概率分布函数 $p(w_i | w_{t-k}, \dots, w_{t+k}) = (e^{y_{w_i}}) / \left(\sum_i e^{y_{w_i}} \right)$,

其中, $i \in [t-k, t+k]$, i 是正整数; w_t 表示广告文本集合中第 t 个特征信息。

[0103] 步骤2023、根据广告文本集合中的第 w_t 个特征信息的第一预设函数,以及第一概率分布函数,建立语义相似度目标函数 $x = \frac{1}{T} \sum_{t=k}^{T-k} \log p(w_i | w_{t-k}, \dots, w_{t+k})$ 。

[0104] 在本实施例中,具体的,在步骤201之后,针对于广告文本集合,需要建立待求解的语义相似度目标函数。

[0105] 具体来说,对于广告文本集合中包括的句子和词语,可以采用DSSM对句子和词语等文本进行bi-char方式预处理,例如,以字为单位直接进行文本预处理。

[0106] 然后,依据广告文本集合中的所有特征信息,针对于广告文本集合中的第 w_t 个特征信息建立一个第一预设函数 $y_{w_t} = b + U * h(w_{t-k}, \dots, w_{t+k}; W)$,可知,对于广告文本集合中的每一个特征信息,后续都会建立一个第一预设函数。在第一预设函数的公式中, b 表示了一个预设的偏差值, U 表示了一个预设的参数向量; $h(w_{t-k}, \dots, w_{t+k}; W)$ 表示形式化函数,其中, W 表示广告文本集合中的第 w_t 个特征信息, w_{t-k} 表示广告文本集合中第 $t-k$ 个特征信息, w_{t+k} 表示广告文本集合中第 $t+k$ 个特征信息, k 表示待建立的语义相似度目标函数的窗口大小, $t \in [k, T]$, T 表示广告文本集合中的特征信息的个数总和, k 、 t 、 T 都是正整数;并且广告文本集合中每一个特征信息为一个向量。

[0107] 然后,根据第 w_t 个特征信息的第一预设函数 $y_{w_t} = b + U * h(w_{t-k}, \dots, w_{t+k}; W)$,以及广告文本集合中的所有特征信息,建立一个第一概率分布函数 $p(w_i | w_{t-k}, \dots, w_{t+k}) = (e^{y_{w_i}}) / \left(\sum_i e^{y_{w_i}} \right)$ 。

在该第一概率分布函数中, $i \in [t-k, t+k]$, i 是正整数; w_t 表示广告文本集合中第 t 个特征信息。

[0108] 然后, 将第 w_t 个特征信息的第一预设函数 $y_{w_t} = b + U * h(w_{t-k}, \dots, w_{t+k}; W)$, 代入到第一概率分布函数 $p(w_t | w_{t-k}, \dots, w_{t+k}) = (e^{y_{w_t}}) / \left(\sum_i e^{y_{w_i}} \right)$ 中; 然后, 由于可以得到广告文本集合中每一个特征信息的第一预设函数, 进而可以将每一个特征信息的第一预设函数都分别代入到第一概率分布函数 $p(w_t | w_{t-k}, \dots, w_{t+k}) = (e^{y_{w_t}}) / \left(\sum_i e^{y_{w_i}} \right)$ 中, 从而可以得到语义相似

度目标函数 $x = \frac{1}{T} \sum_{t=k}^{T-k} \log p(w_t | w_{t-k}, \dots, w_{t+k})$ 。

[0109] 步骤203、对语义相似度目标函数进行求解, 以确定在语义相似度目标函数的最优状态下的第一广告的广告文本整体的向量化表示、以及第二广告的广告文本整体的向量化表示。

[0110] 在本实施例中, 具体的, 对于步骤202中得到的语义相似度目标函数, 采用交叉熵方法进行求解, 去确定出在语义相似度目标函数的最优状态下, 广告文本集合中的每一个特征信息的向量化表示, 即得到第一广告的广告文本整体的向量化表示、第一广告中的每一个词语的向量化表示、第二广告的广告文本整体的向量化表示、第二广告中的每一个词语的向量化表示、至少一个其他广告的广告文本整体的向量化表示、以及至少一个其他广告的一个其他广告中的每一个词语的向量化表示。

[0111] 其中, 语义相似度目标函数的最优状态可以为语义相似度目标函数的值为最大值, 或者, 语义相似度目标函数的最优状态可以为语义相似度目标函数的值在预设范围内。

[0112] 步骤204、根据第一广告的广告文本整体的向量化表示、以及第二广告的广告文本整体的向量化表示, 确定语义相似度。

[0113] 在可选的一种实施方式中, 语义相似度为 $\frac{A \bullet B}{\|A\|_2 \|B\|_2}$, 其中, A 表示第一广告的广告文本整体的向量化表示, B 表示第二广告的广告文本整体的向量化表示。

[0114] 在本实施例中, 具体的, 在步骤203之后, 根据第一广告的广告文本整体的向量化表示 A 、以及第二广告的广告文本整体的向量化表示 B , 求取两者的 $\cos$ 的值, 进而计算出

第一广告与第二广告之间的语义相似度为 $\frac{A \bullet B}{\|A\|_2 \|B\|_2} = \left(\sum_{j=1}^J a_j b_j \right) / \left(\sqrt{\sum_{j=1}^J a_j^2} \sqrt{\sum_{j=1}^J b_j^2} \right)$, 其中, J 表

示向量 A 的维度, 且向量 A 的维度与向量 B 的维度相同, $j \in [1, J]$, j, J 都是正整数, a_j 为向量 A 的第 j 个值, b_j 为向量 B 的第 j 个值。

[0115] 步骤205、根据用户点击集合, 建立点击相似度目标函数。

[0116] 在可选的一种实施方式中, 步骤205具体包括以下步骤:

[0117] 步骤2051、根据用户点击集合, 建立用户点击集合中的第 w'_t 个广告的特征信息的第二预设函数 $y'_{w'_t} = b' + U' * h'(w'_{t-k}, \dots, w'_{t+k}; W')$, 其中, b' 表示预设的第二偏差值, U' 表示预设的第二参数向量, $h'(w'_{t-k}, \dots, w'_{t+k}; W')$ 表示形式化函数, W' 表示用户点击集合中的

第 $w'_{t'}$ 个广告的特征信息, $w'_{t'-k'}$ 表示用户点击集合中第 $t'-k'$ 个广告的特征信息, $w'_{t'+k'}$ 表示用户点击集合中第 $t'+k'$ 个广告的特征信息, k' 表示待建立的点击相似度目标函数的窗口大小, $t' \in [k', T']$, T' 表示用户点击集合中的广告个数总和, k' 、 t' 、 T' 都是正整数。

[0118] 步骤2052、根据用户点击集合,建立第二概率分布函数

$p(w'_i | w'_{t'-k'}, \dots, w'_{t'+k'}) = (e^{y'_{w'_i}}) / \left(\sum_{i'}^{t'+k'} e^{y'_{w'_{i'}}} \right)$, 其中, $i' \in [t'-k', t'+k']$, i' 是正整数; $w'_{t'}$ 表示用户点击集合中第 t' 个广告的特征信息。

[0119] 步骤2053、根据用户点击集合中的第 $w'_{t'}$ 个广告的特征信息的第二预设函数,以及第二概率分布函数,建立点击相似度目标函数 $x' = \frac{1}{T'} \sum_{i'=k'}^{T'-k'} \log p(w'_i | w'_{t'-k'}, \dots, w'_{t'+k'})$ 。

[0120] 在本实施例中,具体的,对于用户点击集合中的特征信息,可以先进行归一化的预处理。

[0121] 然后,依据用户点击集合中的所有特征信息,针对于用户点击集合中的第 w'_t 个广告的特征信息建立一个第二预设函数 $y'_{w'_t} = b' + U' * h'(w'_{t'-k'}, \dots, w'_{t'+k'}; W')$, 可知,对于用户点击集合中的每一个特征信息,后续都会建立一个第二预设函数。在第二预设函数的公式中, b' 表示了一个预设的第二偏差值, U' 表示了一个预设的第二参数向量; $h'(w'_{t'-k'}, \dots, w'_{t'+k'}; W')$ 表示形式化函数,其中, W' 表示用户点击集合中的第 w'_t 个广告的特征信息, $w'_{t'-k'}$ 表示用户点击集合中第 $t'-k'$ 个广告的特征信息, $w'_{t'+k'}$ 表示用户点击集合中第 $t'+k'$ 个广告的特征信息, k' 表示待建立的点击相似度目标函数的窗口大小, $t' \in [k', T']$, T' 表示用户点击集合中的广告个数总和, k' 、 t' 、 T' 都是正整数;并且用户点击集合中每一个特征信息为一个向量。

[0122] 然后,根据第 w'_t 个广告的特征信息的第二预设函数 $y'_{w'_t} = b' + U' * h'(w'_{t'-k'}, \dots, w'_{t'+k'}; W')$,以及用户点击集合中的所有特征信息,建立一个第二概率分布函数

$p(w'_i | w'_{t'-k'}, \dots, w'_{t'+k'}) = (e^{y'_{w'_i}}) / \left(\sum_{i'}^{t'+k'} e^{y'_{w'_{i'}}} \right)$ 。在该第二概率分布函数中, $i' \in [t'-k', t'+k']$, i' 是正整数; $w'_{t'}$ 表示用户点击集合中第 t' 个广告的特征信息。

[0123] 然后,将第 w'_t 个广告的特征信息的第二预设函数 $y'_{w'_t} = b' + U' * h'(w'_{t'-k'}, \dots, w'_{t'+k'}; W')$,

代入到第二概率分布函数 $p(w'_i | w'_{t'-k'}, \dots, w'_{t'+k'}) = (e^{y'_{w'_i}}) / \left(\sum_{i'}^{t'+k'} e^{y'_{w'_{i'}}} \right)$ 中;然后,由于可以得到用户点击集合中每一个特征信息的第二预设函数,进而可以将每一个特征信息的第二预设函数都分别代入到第二概率分布函数中 $p(w'_i | w'_{t'-k'}, \dots, w'_{t'+k'}) = (e^{y'_{w'_i}}) / \left(\sum_{i'}^{t'+k'} e^{y'_{w'_{i'}}} \right)$,从而可以得到

点击相似度目标函数 $x' = \frac{1}{T'} \sum_{i'=k'}^{T'-k'} \log p(w'_i | w'_{t'-k'}, \dots, w'_{t'+k'})$ 。

[0124] 步骤206、对点击相似度目标函数进行求解,以确定在点击相似度目标函数的最优状态下的第一广告的向量化表示、以及第二广告的向量化表示。

[0125] 在本实施例中,具体的,对于步骤205中得到的点击相似度目标函数,可以采用SGD

方法进行求解,去确定出在点击相似度目标函数的最优状态下,用户点击集合中的每一个特征信息的向量化表示,即可以得到第一广告的向量化表示、第二广告的向量化表示、第三广告的向量化表示、以此类推。其中,用户点击集合中的每一个广告为用户点击过的广告,优选的,用户点击集合中的每一个广告为属于同一群体的用户所点击过的广告。

[0126] 其中,点击相似度目标函数的最优状态可以为点击相似度目标函数的值为最大值,或者,点击相似度目标函数的最优状态可以为点击相似度目标函数的值在预设范围内。

[0127] 步骤207、根据第一广告的向量化表示、以及第二广告的向量化表示,确定点击相似度。

[0128] 在可选的一种实施方式中,点击相似度为 $\frac{C \cdot D}{\|C\|_2 \|D\|_2}$,其中,C表示第一广告的向量化表示,D表示第二广告的向量化表示。

[0129] 在本实施例中,具体的,在步骤206之后,根据第一广告的向量化表示C、以及第二广告的向量化表示D,求取两者的cosine的值,进而计算出第一广告与第二广告之间的点击

相似度为 $\frac{C \cdot D}{\|C\|_2 \|D\|_2} = \left(\sum_{j'=1}^{J'} c_{j'} d_{j'} \right) / \left(\sqrt{\sum_{j'=1}^{J'} d_{j'}^2} \sqrt{\sum_{j'=1}^{J'} d_{j'}^2} \right)$,其中,J'表示向量C的维度,且向量C的维度与向量D的维度相同, $j' \in [1, J']$, j' 、 J' 都是正整数, $c_{j'}$ 为向量C的第 j' 个值, $d_{j'}$ 为向量D的第 j' 个值。

[0130] 步骤208、根据语义相似度和点击相似度,确定第一广告与第二广告之间的相似度信息。

[0131] 在可选的一种实施方式中,步骤208具体包括:根据语义相似度和点击相似度,确定第一广告与第二广告之间的相似度信息,包括:获取第二广告的用户点击频次;根据用户点击频次、语义相似度和点击相似度,确定相似度信息。

[0132] 在可选的一种实施方式中,相似度信息为 $\text{Sim} = (1/\log(\text{TF})) * \text{Sim}_{\text{content}} + \text{Sim}_{\text{session}}$;其中,TF表示用户点击频次, $\text{Sim}_{\text{content}}$ 表示语义相似度, $\text{Sim}_{\text{session}}$ 表示点击相似度。

[0133] 在本实施例中,具体的,本步骤可以参见图1的步骤104不再赘述。

[0134] 本实施例通过获取广告文本集合,其中,广告文本集合中包括第一广告的广告文本整体的特征信息、第一广告中的每一个词语的特征信息、第二广告的广告文本整体的特征信息、第二广告中的每一个词语的特征信息、至少一个其他广告的广告文本整体的特征信息、以及至少一个其他广告的每一个其他广告中的每一个词语的特征信息,并获取用户点击集合,其中,用户点击集合中包括第一广告的特征信息、第二广告的特征信息、以及至少一个用户点击过的其他广告的特征信息;根据广告文本集合,确定第一广告与第二广告之间的语义相似度;根据用户点击集合,确定第一广告与第二广告之间的点击相似度;根据语义相似度和点击相似度,确定第一广告与第二广告之间的相似度信息。从而通过对海量的广告中的词语进行提取,根据神经网络模型对海量的广告中的词语进行分析,可以对短文本的广告和长文本的广告都进行分析,便于提取广告中的主题和关键信息;并且,可以从用户点击广告的行为的角度进行分析,去获取到属于同一群体的用户所点击的海量的广告,然后,将属于同一群体的用户所点击过的广告构成一个用户点击集合,去对用户点击集合中的所有广告的特征进行分析,有利于进行广告归类;并且以上过程都是对海量的广告

数据进行的分析,可以更准确的确定出广告之间相似性;然后将根据广告文本集合计算得到的语义相似度、以及根据用户点击集合计算得到的点击相似度进行计算,计算得到第一广告与第二广告之间的相似度信息,即第二广告在多大程度上是与第一广告相似的,可以准确确定的确定出广告之间的相似度。进而可以根据以上过程确定出所有广告之间相似度,从而在向用户推送广告的时候,可以将相似的广告推送给用户。

[0135] 图5为本发明实施例提供的一种广告的相似度处理装置的结构示意图,如图5所示,本实施例的装置可以包括:

[0136] 获取单元31,用于获取广告文本集合,其中,广告文本集合中包括第一广告的广告文本整体的特征信息、第一广告中的每一个词语的特征信息、第二广告的广告文本整体的特征信息、第二广告中的每一个词语的特征信息、至少一个其他广告的广告文本整体的特征信息、以及至少一个其他广告的每一个其他广告中的每一个词语的特征信息,并获取用户点击集合,其中,用户点击集合中包括第一广告的特征信息、第二广告的特征信息、以及至少一个用户点击过的其他广告的特征信息。

[0137] 第一确定单元32,用于根据广告文本集合,确定第一广告与第二广告之间的语义相似度。

[0138] 第二确定单元33,用于根据用户点击集合,确定第一广告与第二广告之间的点击相似度。

[0139] 第三确定单元34,用于根据语义相似度和点击相似度,确定第一广告与第二广告之间的相似度信息。

[0140] 本实施例的广告的相似度处理装置可执行本发明实施例提供的一种广告的相似度处理方法,其实现原理相类似,此处不再赘述。

[0141] 本实施例通过获取广告文本集合,其中,广告文本集合中包括第一广告的广告文本整体的特征信息、第一广告中的每一个词语的特征信息、第二广告的广告文本整体的特征信息、第二广告中的每一个词语的特征信息、至少一个其他广告的广告文本整体的特征信息、以及至少一个其他广告的每一个其他广告中的每一个词语的特征信息,并获取用户点击集合,其中,用户点击集合中包括第一广告的特征信息、第二广告的特征信息、以及至少一个用户点击过的其他广告的特征信息;根据广告文本集合,确定第一广告与第二广告之间的语义相似度;根据用户点击集合,确定第一广告与第二广告之间的点击相似度;根据语义相似度和点击相似度,确定第一广告与第二广告之间的相似度信息。从而通过对海量的广告中的词语进行提取,根据神经网络模型对海量的广告中的词语进行分析,可以对短文本的广告和长文本的广告都进行分析,便于提取广告中的主题和关键信息;并且,可以从用户点击广告的行为的角度进行分析,去获取到属于同一群体的用户所点击的海量的广告,然后,将属于同一群体的用户所点击过的广告构成一个用户点击集合,去对用户点击集合中的所有广告的特征进行分析,有利于进行广告归类;并且以上过程都是对海量的广告数据进行的分析,可以更准确的确定出广告之间相似性;然后将根据广告文本集合计算得到的语义相似度、以及根据用户点击集合计算得到的点击相似度进行计算,计算得到第一广告与第二广告之间的相似度信息,即第二广告在多大程度上是与第一广告相似的,可以准确确定的确定出广告之间的相似度。进而可以根据以上过程确定出所有广告之间相似度,从而在向用户推送广告的时候,可以将相似的广告推送给用户。

[0142] 图6为本发明实施例提供的另一种广告的相似度处理装置的结构示意图,在图5所示实施例的基础上,如图6所示,本实施例提供的装置中,第一确定单元32,包括:

[0143] 第一建立模块321,用于根据广告文本集合,建立语义相似度目标函数。

[0144] 第一求解模块322,用于对语义相似度目标函数进行求解,以确定在语义相似度目标函数的最优状态下的第一广告的广告文本整体的向量化表示、以及第二广告的广告文本整体的向量化表示。

[0145] 第一确定模块323,用于根据第一广告的广告文本整体的向量化表示、以及第二广告的广告文本整体的向量化表示,确定语义相似度。

[0146] 第一建立模块321,包括:

[0147] 第一建立子模块3211,用于根据广告文本集合,建立广告文本集合中的第 w_t 个特征信息的第一预设函数 $y_{w_t} = b + U * h(w_{t-k}, \dots, w_{t+k}; W)$, 其中, b 表示预设的偏差值, U 表示预设的参数向量, $h(w_{t-k}, \dots, w_{t+k}; W)$ 表示形式化函数, W 表示广告文本集合中的第 w_t 个特征信息, w_{t-k} 表示广告文本集合中第 $t-k$ 个特征信息, w_{t+k} 表示广告文本集合中第 $t+k$ 个特征信息, k 表示待建立的语义相似度目标函数的窗口大小, $t \in [k, T]$, T 表示广告文本集合中的特征信息的个数总和, k, t, T 都是正整数。

[0148] 第二建立子模块3212,用于根据广告文本集合,建立第一概率分布函数 $p(w_t | w_{t-k}, \dots, w_{t+k}) = (e^{y_{w_t}}) / \left(\sum_i^{t+k} e^{y_{w_i}} \right)$, 其中, $i \in [t-k, t+k]$, i 是正整数; w_t 表示广告文本集合中第 t 个特征信息。

[0149] 第三建立子模块3213,用于根据广告文本集合中的第 w_t 个特征信息的第一预设函数,以及第一概率分布函数,建立语义相似度目标函数 $x = \frac{1}{T} \sum_{t=k}^{T-k} \log p(w_t | w_{t-k}, \dots, w_{t+k})$ 。

[0150] 语义相似度为 $\frac{A \bullet B}{\|A\|_2 \|B\|_2}$, 其中, A 表示第一广告的广告文本整体的向量化表示, B 表示第二广告的广告文本整体的向量化表示。

[0151] 第二确定单元33,包括:

[0152] 第二建立模块331,用于根据用户点击集合,建立点击相似度目标函数。

[0153] 第二求解模块332,用于对点击相似度目标函数进行求解,以确定在点击相似度目标函数的最优状态下的第一广告的向量化表示、以及第二广告的向量化表示。

[0154] 第二确定模块333,用于根据第一广告的向量化表示、以及第二广告的向量化表示,确定点击相似度。

[0155] 第二建立模块331,包括:

[0156] 第四建立子模块3311,用于根据用户点击集合,建立用户点击集合中的第 w'_t 个广告的特征信息的第二预设函数 $y'_{w'_t} = b' + U' * h'(w'_{t'-k'}, \dots, w'_{t'+k'}; W')$, 其中, b' 表示预设的第二偏差值, U' 表示预设的第二参数向量, $h'(w'_{t'-k'}, \dots, w'_{t'+k'}; W')$ 表示形式化函数, W' 表示用户点击集合中的第 w'_t 个广告的特征信息, $w'_{t'-k'}$ 表示用户点击集合中第 $t'-k'$ 个广告的特征信息, $w'_{t'+k'}$ 表示用户点击集合中第 $t'+k'$ 个广告的特征信息, k' 表示待建立的点击相似度目标函数的窗口大小, $t' \in [k', T']$, T' 表示用户点击集合中的广告个数总和, k', T' 都是正整数。

t' 、 T' 都是正整数。

[0157] 第五建立子模块3312,用于根据用户点击集合,建立第二概率分布函数

$$p(w'_t | w'_{t-k'}, \dots, w'_{t+k'}) = \left(e^{y'_{w'_t}} \right) / \left(\sum_i e^{y'_{w'_i}} \right), \text{其中, } i' \in [t'-k', t'+k'], i' \text{ 是正整数; } w'_t \text{ 表示用}$$

户点击集合中第 t' 个广告的特征信息。

[0158] 第六建立子模块3313,用于根据用户点击集合中的第 w'_t 个广告的特征信息的第二预设函数,以及第二概率分布函数,建立点击相似度目标函数

$$x' = \frac{1}{T'} \sum_{t'=k'}^{T'-k'} \log p(w'_t | w'_{t-k'}, \dots, w'_{t+k'})。$$

[0159] 点击相似度为 $\frac{C \cdot D}{\|C\|_2 \|D\|_2}$, 其中,C表示第一广告的向量化表示,D表示第二广告的向量化表示。

[0160] 第三确定单元34,包括:

[0161] 获取模块341,用于获取第二广告的用户点击频次。

[0162] 第三确定模块342,用于根据用户点击频次、语义相似度和点击相似度,确定相似度信息。

[0163] 相似度信息为 $\text{Sim} = (1/\log(\text{TF})) * \text{Sim}_{\text{content}} + \text{Sim}_{\text{session}}$; 其中,TF表示用户点击频次, $\text{Sim}_{\text{content}}$ 表示语义相似度, $\text{Sim}_{\text{session}}$ 表示点击相似度。

[0164] 本实施例的广告的相似度处理装置可执行本发明实施例提供的另一种广告的相似度处理方法,其实现原理相类似,此处不再赘述。

[0165] 本实施例通过获取广告文本集合,其中,广告文本集合中包括第一广告的广告文本整体的特征信息、第一广告中的每一个词语的特征信息、第二广告的广告文本整体的特征信息、第二广告中的每一个词语的特征信息、至少一个其他广告的广告文本整体的特征信息、以及至少一个其他广告的每一个其他广告中的每一个词语的特征信息,并获取用户点击集合,其中,用户点击集合中包括第一广告的特征信息、第二广告的特征信息、以及至少一个用户点击过的其他广告的特征信息;根据广告文本集合,确定第一广告与第二广告之间的语义相似度;根据用户点击集合,确定第一广告与第二广告之间的点击相似度;根据语义相似度和点击相似度,确定第一广告与第二广告之间的相似度信息。从而通过对海量的广告中的词语进行提取,根据神经网络模型对海量的广告中的词语进行分析,可以对短文本的广告和长文本的广告都进行分析,便于提取广告中的主题和关键信息;并且,可以从用户点击广告的行为的角度进行分析,去获取到属于同一群体的用户所点击的海量的广告,然后,将属于同一群体的用户所点击过的广告构成一个用户点击集合,去对用户点击集合中的所有广告的特征进行分析,有利于进行广告归类;并且以上过程都是对海量的广告数据进行的分析,可以更准确的确定出广告之间相似性;然后将根据广告文本集合计算得到的语义相似度、以及根据用户点击集合计算得到的点击相似度进行计算,计算得到第一广告与第二广告之间的相似度信息,即第二广告在多大程度上是与第一广告相似的,可以准确确定的确定出广告之间的相似度。进而可以根据以上过程确定出所有广告之间相似度,从而在向用户推送广告的时候,可以将相似的广告推送给用户。

[0166] 在本发明所提供的几个实施例中,应该理解到,所揭露的装置和方法,可以通过其它的方式实现。例如,以上所描述的装置实施例仅仅是示意性的,例如,单元的划分,仅仅为一种逻辑功能划分,实际实现时可以有另外的划分方式,例如多个单元或组件可以结合或者可以集成到另一个系统,或一些特征可以忽略,或不执行。另一点,所显示或讨论的相互之间的耦合或直接耦合或通信连接可以是通过一些接口,装置或单元的间接耦合或通信连接,可以是电性,机械或其它的形式。

[0167] 作为分离部件说明的单元可以是或者也可以不是物理上分开的,作为单元显示的部件可以是或者也可以不是物理单元,即可以位于一个地方,或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部单元来实现本实施例方案的目的。

[0168] 另外,在本发明各个实施例中的各功能单元可以集成在一个处理单元中,也可以是各个单元单独物理存在,也可以两个或两个以上单元集成在一个单元中。上述集成的单元既可以采用硬件的形式实现,也可以采用硬件加软件功能单元的形式实现。

[0169] 上述以软件功能单元的形式实现的集成的单元,可以存储在一个计算机可读取存储介质中。上述软件功能单元存储在一个存储介质中,包括若干指令用以使得一台计算机设备(可以是个人计算机,服务器,或者网络设备等)或处理器(processor)执行本发明各个实施例的方法的部分步骤。而前述的存储介质包括:U盘、移动硬盘、只读存储器(Read-Only Memory,ROM)、随机存取存储器(Random Access Memory,RAM)、磁碟或者光盘等各种可以存储程序代码的介质。

[0170] 本领域技术人员在考虑说明书及实践这里公开的发明后,将容易想到本公开的其它实施方案。本发明旨在涵盖本公开的任何变型、用途或者适应性变化,这些变型、用途或者适应性变化遵循本公开的一般性原理并包括本公开未公开的本技术领域中的公知常识或惯用技术手段。说明书和实施例仅被视为示例性的,本公开的真正范围和精神由下面的权利要求书指出。

[0171] 应当理解的是,本公开并不局限于上面已经描述并在附图中示出的精确结构,并且可以在不脱离其范围进行各种修改和改变。本公开的范围仅由所附的权利要求书来限制。

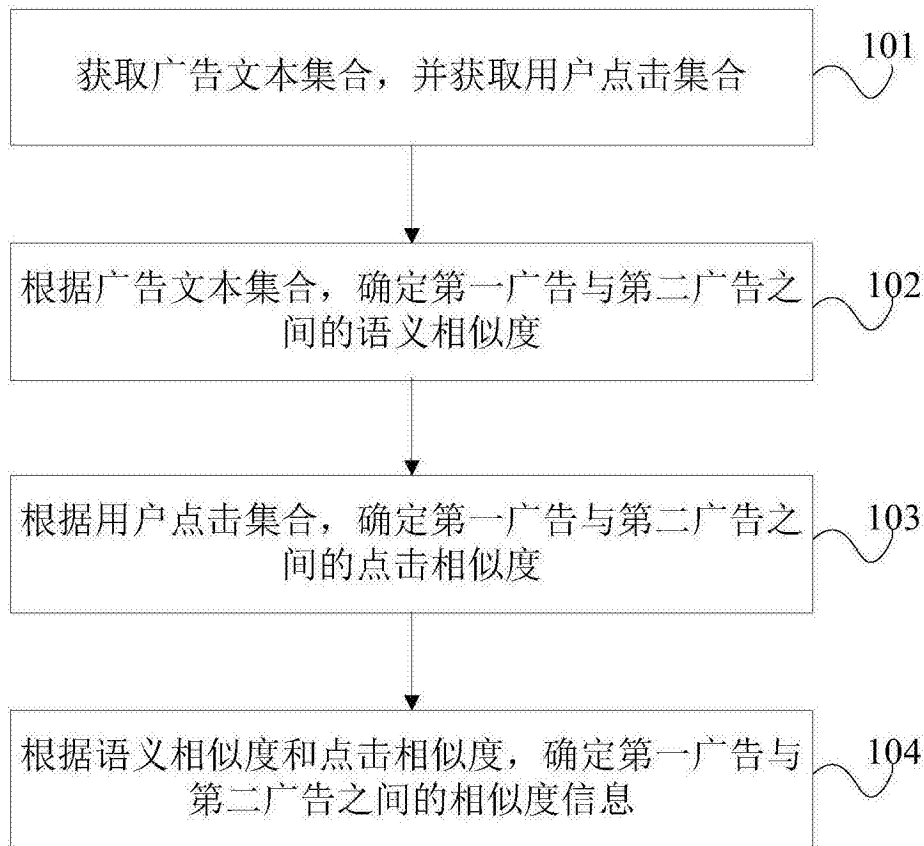


图 1

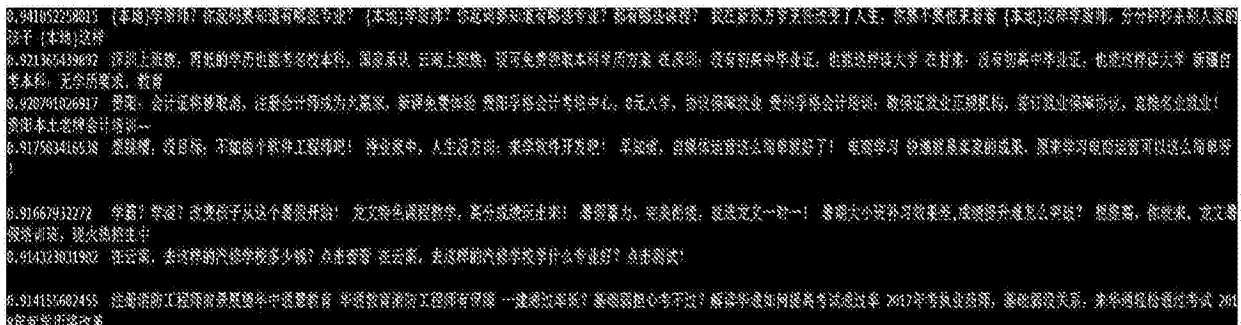


图2

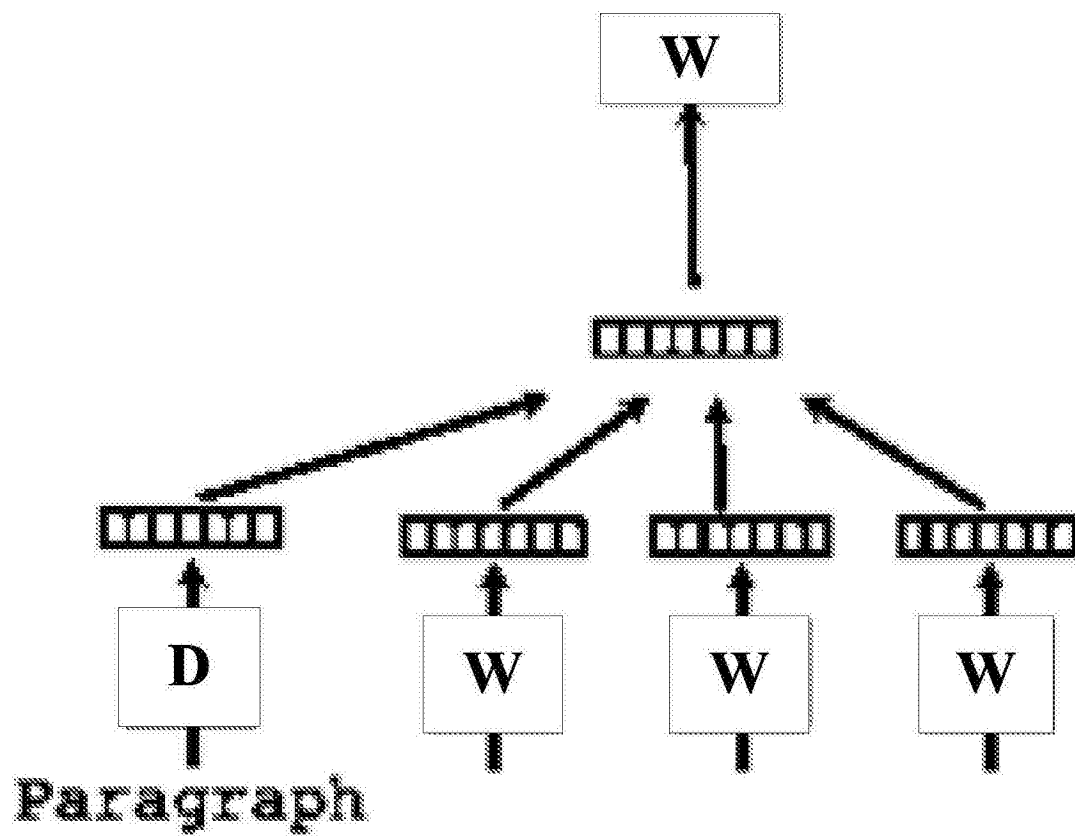


图3



图4

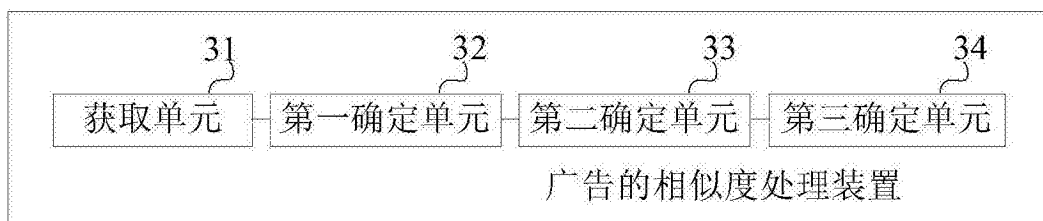


图5

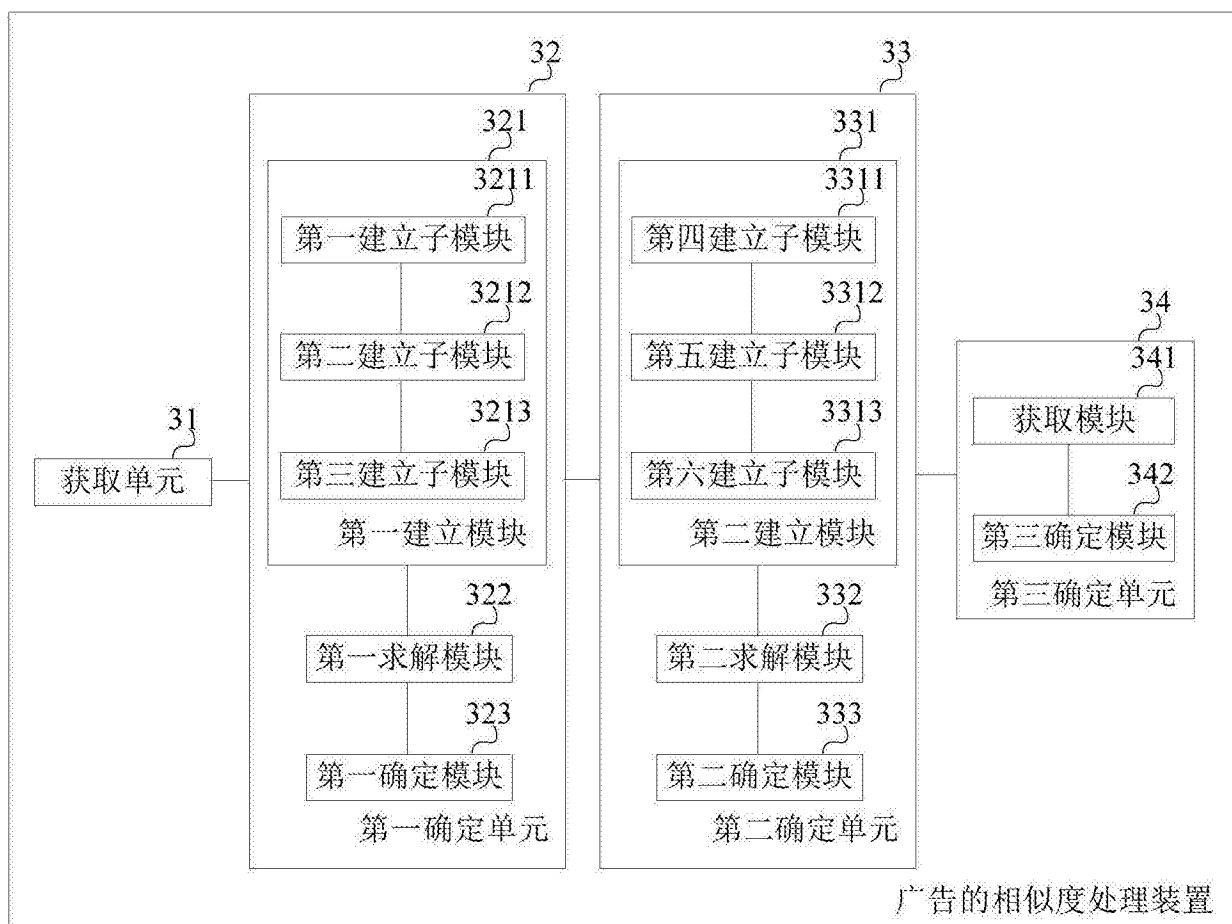


图6