(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 103577452 A
(43)申请公布日 2014.02.12

(21)申请号 201210270201.4

(22)申请日 2012.07.31

(71)申请人 国际商业机器公司
地址 美国纽约

(72)发明人 郭宏蕾 蔡柯柯 包胜华 张硕 吴贤 张俐 苏中

(74)专利代理机构 中国国际贸易促进委员会专利商标事务所 11038
代理人 李镇江

(51)Int.Cl.
G06F 17/30(2006.01)

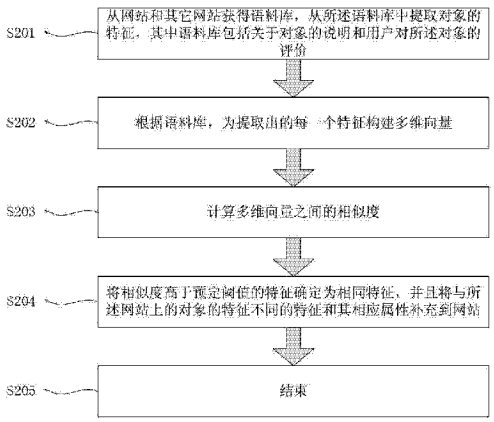
权利要求书5页 说明书14页 附图5页

(54)发明名称

用于丰富网站内容的方法和装置、网站服务器

(57)摘要

本公开涉及一种用于丰富网站内容的装置和方法、网站服务器。本发明的用于丰富网站内容的方法包括：从所述网站和其它网站获得语料库，从所述语料库中提取所述对象的特征，其中所述语料库包括关于对象的说明和用户对所述对象的评价；根据所述语料库，为提取出的特征构建多维向量；针对特定特征，将其多维向量与提取出的其它特征的多维向量进行相似度比较；将相似度高于预定阈值的特征确定为相同特征，并且将与所述网站上的对象的特征不同的特征和其相应属性补充到该网站。本发明提高了网页整合效率。



CN 103577452 A

1. 一种用于丰富网站内容的方法，所述方法包括：

从所述网站和其它网站获得语料库，从所述语料库中提取对象的特征，其中所述语料库包括关于所述对象的说明和用户对所述对象的评价；

根据所述语料库，为提取出的特征构建多维向量；

针对特定特征，将其多维向量与提取出的其它特征的多维向量进行相似度比较；

将相似度高于预定阈值的特征确定为相同特征，并且将与所述网站上的对象的特征不同的特征和其相应属性补充到该网站。

2. 根据权利要求 1 所述的方法，其中从所述网站和其它网站获得语料库包括：

指定所述其它网站；

分析所述网站和其它网站的格式；

按照分析出的所述网站和其它网站的格式，寻找含有与所述对象对应的对象标识的所有的块；

根据块的格式判断寻找到的块是关于对象的说明还是用户对所述对象的评价，将寻找到的关于对象的说明和用户对所述对象的评价作为语料库。

3. 根据权利要求 1 所述的方法，其中从所述语料库中提取所述对象的特征包括：

从所述网站和其它网站中关于对象的说明中提取特征种子，其中按照所述网站和其它网站中关于对象的说明的格式，从相应字段中提取特征种子；

按照提取出的特征种子，从用户对所述对象的评价提取附加特征。

4. 根据权利要求 3 所述的方法，其中从用户对所述对象的评价提取附加特征包括：

从用户对所述对象的评价中提取出所述特征种子附近满足预定条件的名词作为附加特征；

从用户对所述对象的评价中提取出包含所述特征种子的名词词组作为附加特征；

如果提取出的附加特征不在特征种子的列表中，将提取出的附加特征加入到特征种子的列表；

迭代地重复上述步骤，直到不在特征种子的列表中的新提取出的附加特征的数目低于预定阈值为止。

5. 根据权利要求 4 所述的方法，其中满足预定条件的名词是指在特征种子附近的预定范围内的出现频率最高的前 n 名的名词，n 为自然数。

6. 根据权利要求 1 — 5 中任一个所述的方法，其中所构建的多维向量至少包括以下维度中的一个或多个：

特征的情感线索，包括从所述用户对所述对象的评价提取出特定特征的评价词、情感词组成的对或特定特征的评价分类标记、情感词组成的对，其中对于含义类似的评价词给予相同的评价分类标记；

特征的上下文线索，即在从用户对所述对象的评价中特定特征附近满足预定条件的形容词和 / 或名词和 / 或名词短语和 / 或否定词；

特征的可用标签，即所述网站和其它网站赋予特定特征的分组标签信息。

7. 根据权利要求 6 中任一个所述的方法，其中所构建的多维向量还包括如下维度中的至少一个：

特征的名称；

特征的内部线索，其中特征的内部线索包括特定特征的关键词和特定特征的构成词，其中所述关键词和构成词都构成所述多维向量中的维度。

8. 根据权利要求 1-5 任意一项所述的方法，其中针对特定特征将其多维向量与提取出的其它特征的多维向量进行相似度比较包括：

将提取出的所有特征的多维向量的维度进行对准，其中对于特定特征，将提取出的其它特征的多维向量中的维度也视为该特定特征的维度，从而使每个特征的多维向量中的维度数相同；

计算所述语料库中特定特征与其多维向量的每一维度之间的互信息作为每一维度的权重；

根据所述权重计算各特征的多维向量之间的相似度。

9. 根据权利要求 8 所述的方法，其中计算所述语料库中所述特征与其多维向量的每一维度之间的互信息作为每一维度的权重包括：

对于由于将提取出的其它特征的多维向量中的维度视为特定特征的维度导致的增加维度，权重视为 0，否则利用如下公式计算特定特征与其特定维度的互信息作为权重：

$$\mathrm{Weight}(f_j ; p_f) = \log_2 \frac{P(f_j, p_f)}{P(f_j)P(p_f)}$$

其中 p_f 表示所述特定特征，

f_j 表示所述特定特征的第 j 个维度，

$P(f_j;p_f)$ 是在所述语料库中所述特定特征与第 j 个维度在一句话中同时出现的概率，

$P(f_j)$ 是在所述语料库中第 j 个维度在一句话中出现的概率，以及

$P(p_f)$ 是在所述语料库中所述特定特征在一句话中出现的概率。

10. 根据权利要求 8 所述的方法，其中利用欧式距离计算所述各特征的多维向量之间的相似度：

$$dis\tan ce(v_i, v_j) = \sqrt{\sum_{k=1}^{k=n} (w(f_k(v_i)) - w(f_k(v_j)))^2}$$

其中，v_i 表示第 i 个特征的多维向量，v_j 表示第 j 个特征的多维向量；

$w(f_k(v_i))$ 表示多维向量 v_i 的第 k 个维度的权重，

$w(f_k(v_j))$ 表示多维向量 v_j 的第 k 个维度的权重。

11. 根据权利要求 8 所述的方法，其中利用余弦相似度计算所述各特征的多维向量之间的相似度：

$$\sin(v_i, v_j) = \cos\theta = \frac{\sum_{k=1}^{k=n} w(f_k(v_i)) \times w(f_k(v_j))}{\sqrt{(\sum_{k=1}^{k=n} w^2(f_k(v_i)))(\sum_{k=1}^{k=n} w^2(f_k(v_j)))}}$$

其中，v_i 表示第 i 个特征的多维向量，v_j 表示第 j 个特征的多维向量；

$w(f_k(v_i))$ 表示多维向量 v_i 的第 k 个维度的权重，

$w(f_k(v_j))$ 表示多维向量 v_j 的第 k 个维度的权重。

12. 根据权利要求 1 所述的方法，其中将相似度高于预定阈值的特征确定为相同特征并且将与所述网站上的对象的特征不同的特征和其相应属性补充到该网站包括：

将相似度高于预定阈值的特征分组到同一组；

判断所述网站上已有的对象的特征是否属于一个分组成的组，识别出不包含所述网站上已有对象的任何特征的特征组，将该特征组中的特征和其相应属性补充到该网站。

13. 根据权利要求 12 所述的方法，其中将该组的特征和其相应属性补充到该网站包括：统计该特征组的各特征在语料库中的出现次数，将出现次数最高的特征的名称和其相应属性补充到该网站。

14. 一种用于丰富网站内容的装置，所述装置包括：

提取单元，被配置为从所述网站和其它网站获得语料库，从所述语料库中提取对象的特征，其中所述语料库包括关于所述对象的说明和用户对所述对象的评价；

特征向量构建单元，被配置为根据所述语料库，为提取出的特征构建多维向量；

向量比较单元，被配置为针对特定特征，将其多维向量与提取出的其它特征的多维向量进行相似度比较；

补充单元，被配置为将相似度高于预定阈值的特征确定为相同特征，并且将与所述网站上的对象的特征不同的特征和其相应属性补充到该网站。

15. 根据权利要求 14 所述的装置，其中所述提取单元进一步被配置为：

指定所述其它网站；

分析所述网站和其它网站的格式；

按照分析出的所述网站和其它网站的格式，寻找含有与所述对象对应的对象标识的所有的块；

根据块的格式判断寻找到的块是关于对象的说明还是用户对所述对象的评价，将寻找到的关于对象的说明和用户对所述对象的评价作为语料库。

16. 根据权利要求 14 所述的装置，其中所述提取单元进一步被配置为：

从所述网站和其它网站中关于对象的说明中提取特征种子，其中按照所述网站和其它网站中关于对象的说明的格式，从相应字段中提取特征种子；

按照提取出的特征种子，从用户对所述对象的评价提取附加特征。

17. 根据权利要求 16 所述的装置，其中从用户对所述对象的评价提取附加特征包括：

从用户对所述对象的评价中提取出所述特征种子附近满足预定条件的名词作为附加特征；

从用户对所述对象的评价中提取出包含所述特征种子的名词词组作为附加特征；

如果提取出的附加特征不在特征种子的列表中，将提取出的附加特征加入到特征种子的列表；

迭代地重复上述步骤，直到不在特征种子的列表中的新提取出的附加特征的数目低于预定阈值为止。

18. 根据权利要求 17 所述的装置，其中满足预定条件的名词是指在特征种子附近的预定范围内的出现频率最高的前 n 名的名词，n 为自然数。

19. 根据权利要求 14 － 18 中任一个所述的装置，其中所构建的多维向量至少包括以下维度中的一个或多个：

特征的情感线索，包括从所述用户对所述对象的评价提取出特定特征的评价词、情感词组成的对或特定特征的评价分类标记、情感词组成的对，其中对于含义类似的评价词给予相同的评价分类标记；

特征的上下文线索，即在从用户对所述对象的评价中特定特征附近满足预定条件的形容词和 / 或名词和 / 或名词短语和 / 或否定词 ；

特征的可用标签，即所述网站和其它网站赋予特定特征的分组标签信息。

20. 根据权利要求 19 中任一个所述的装置，其中所构建的多维向量还包括如下维度中的至少一个 ：

特征的名称 ；

特征的内部线索，其中特征的内部线索包括特定特征的关键词和特定特征的构成词，其中所述关键词和构成词都构成所述多维向量中的维度。

21. 根据权利要求 14 — 18 任意一项所述的装置，其中向量比较单元进一步被配置为 ：

将提取出的所有特征的多维向量的维度进行对准，其中对于特定特征，将提取出的其它特征的多维向量中的维度也视为该特定特征的维度，从而使每个特征的多维向量中的维度数相同 ；

计算所述语料库中特定特征与其多维向量的每一维度之间的互信息作为每一维度的权重 ；

根据所述权重计算各特征的多维向量之间的相似度。

22. 根据权利要求 21 所述的装置，其中计算所述语料库中所述特征与其多维向量的每一维度之间的互信息作为每一维度的权重包括 ：

对于由于将提取出的其它特征的多维向量中的维度视为特定特征的维度导致的增加维度，权重视为 0，否则利用如下公式计算特定特征与其特定维度的互信息作为权重 ：

$$\mathrm{Weight}(\mathrm{f_j}\,;\mathrm{p_f}) = \log_2 \frac{P(\mathrm{f_j}, p_f)}{P(f_j)P(p_f)}$$

其中 p_f 表示所述特定特征，

f_j 表示所述特定特征的第 j 个维度，

$P(f_j;p_f)$ 是在所述语料库中所述特定特征与第 j 个维度在一句话中同时出现的概率，

$P(f_j)$ 是在所述语料库中第 j 个维度在一句话中出现的概率，以及

$P(p_f)$ 是在所述语料库中所述特定特征在一句话中出现的概率。

23. 根据权利要求 21 所述的装置，其中利用欧式距离计算所述各特征的多维向量之间的相似度 ：

$$dis\tan ce(v_i, v_j) = \sqrt{\sum_{k=1}^{k=n} (w(f_k(v_i)) - w(f_k(v_j)))^2}$$

其中，v_i 表示第 i 个特征的多维向量，v_j 表示第 j 个特征的多维向量 ；

$w(f_k(v_i))$ 表示多维向量 v_i 的第 k 个维度的权重，

$w(f_k(v_j))$ 表示多维向量 v_j 的第 k 个维度的权重。

24. 根据权利要求 21 所述的装置，其中利用余弦相似度计算所述各特征的多维向量之间的相似度 ：

$$\sin(v_i, v_j) = \cos\theta = \frac{\sum_{k=1}^{k=n} w(f_k(v_i)) \times w(f_k(v_j))}{\sqrt{(\sum_{k=1}^{k=n} w^2(f_k(v_i)))(\sum_{k=1}^{k=n} w^2(f_k(v_j)))}}$$

其中，v_i 表示第 i 个特征的多维向量，v_j 表示第 j 个特征的多维向量 ；

$w(f_k(v_i))$ 表示多维向量 v_i 的第 k 个维度的权重，

$w(f_k(v_j))$ 表示多维向量 v_j 的第 k 个维度的权重。

25. 根据权利要求 14 所述的装置，其中补充单元进一步被配置为：

将相似度高于预定阈值的特征分组到同一组；

判断所述网站上已有的对象的特征是否属于一个分组成的组，识别出不包含所述网站上已有对象的任何特征的特征组，将该特征组中的特征和其相应属性补充到该网站。

26. 根据权利要求 25 所述的装置，其中将该组的特征和其相应属性补充到该网站包括：统计该特征组的各特征在语料库中的出现次数，将出现次数最高的特征的名称和其相应属性补充到该网站。

27. 一种网站服务器，包括根据权利要求 14 － 26 中的任一个的用于丰富网站内容的装置。

用于丰富网站内容的方法和装置、网站服务器

技术领域

[0001] 本公开总体来说涉及一种丰富网站内容的方法和装置，更具体地，本公开涉及利用语义分析、计算方法来丰富网站中关于对象的说明。

背景技术

[0002] 如今，各种各样的网站提供各种各样的网络内容。网页上经常保护对一个对象的说明，例如对一个事件、一个产品、一个人物的说明等。用户看到该网页上，往往会产生一种需求，即想看到关于这个事件、这个产品、这个人物其它方面的说明，并希望能够看到一个网页，在该网页上将在该网站和其它网站上找到的关于这个事件、这个产品、这个人物的说明整合在一起，便于用户阅读。

[0003] 作为一个例子，用户在网页上看到一个人物的说明。用户非常想知道该人物更多的方面，但本网页上只有该人物的年龄、身高、性别。如果用户想知道关于该人物的其它方面，该用户必须查询其它的网页。用户希望看到一个整合的网页，该网页上将在本网站和其它网站上找到的关于这个人物的说明例如按照年龄、身高、性别、兴趣、职业、血型、星座等方面整合在一起，便于用户阅读。

[0004] 作为另一个例子，用户在网页上看到一个产品的说明。用户非常想知道该产品更多的方面，但本网页上只有该产品的型号、颜色、价格。如果用户想知道关于该产品的其它方面，该用户必须查询其它的网页。用户希望看到一个整合的网页，该网页上将在本网站和其它网站上找到的关于这个产生的说明例如按照年型号、颜色、价格、尺寸、芯片、内存、重量等方面整合在一起，便于用户阅读。

[0005] 一般来说，在具有关于人物的说明的网站上，还有其它用户看了该关于人物的说明后的一些评论或感想。在具有关于产品的说明的网站上，还有其它用户看了该关于产品的说明后的一些评论或感想。这些评论或感想与关于人物或产品的说明位于网页格式的不同的块中。

[0006] 在各个网站上关于对象的说明往往采用了不同的词语。如果简单地收集网站所提供的说明和数据，很可能提供了重复的信息。例如不同的网站上对于同一对象的说明中可能分别出现了屏幕、显示器、手机屏、显示屏等等，但实际上它们的含义是基本相同的。如果把关于它们的信息都整合进网页中，提供了重复的信息且页面可读性差。

[0007] 另一方面，现有技术中仅仅关注了对网站说明的收集和提取。但实际上，用户的评论或感想中也存在着大量有用的信息。现有技术没有实现网页信息利用最大化。

发明内容

[0008] 本发明解决的一个技术问题是提供一种丰富网络内容的方法、装置及网络服务器，其能够用其它网站上关于对象的说明来丰富网站上关于对象的说明而不引入重复信息，提高网页整合效率。

[0009] 根据本发明的一方面，提供了一种用于丰富网站内容的方法，所述方法包括：从所

述网站和其它网站获得语料库，从所述语料库中提取对象的特征，其中所述语料库包括关于所述对象的说明和用户对所述对象的评价；根据所述语料库，为提取出的特征构建多维向量；针对特定特征，将其多维向量与提取出的其它特征的多维向量进行相似度比较；将相似度高于预定阈值的特征确定为相同特征，并且将与所述网站上的对象的特征不同的特征和其相应属性补充到该网站。

[0010] 可选地，从所述网站和其它网站获得语料库包括：指定所述其它网站；分析所述网站和其它网站的格式；按照分析出的所述网站和其它网站的格式，寻找含有与所述对象对应的对象标识的所有的块；根据块的格式判断寻找到的块是关于对象的说明还是用户对所述对象的评价，将寻找到的关于对象的说明和用户对所述对象的评价作为语料库。

[0011] 可选地，从所述语料库中提取所述对象的特征包括：从所述网站和其它网站中关于对象的说明中提取特征种子，其中按照所述网站和其它网站中关于对象的说明的格式，从相应字段中提取特征种子；按照提取出的特征种子，从用户对所述对象的评价提取附加特征。

[0012] 可选地，从用户对所述对象的评价提取附加特征包括：从用户对所述对象的评价中提取出所述特征种子附近满足预定条件的名词作为附加特征；从用户对所述对象的评价中提取出包含所述特征种子的名词词组作为附加特征；如果提取出的附加特征不在特征种子的列表中，将提取出的附加特征加入到特征种子的列表；迭代地重复上述步骤，直到不在特征种子的列表中的新提取出的附加特征的数目低于预定阈值为止。

[0013] 可选地，满足预定条件的名词是指在特征种子附近的预定范围内的出现频率最高的前 n 名的名词，n 为自然数。

[0014] 可选地，所构建的多维向量至少包括以下维度中的一个或多个：特征的情感线索，包括从所述用户对所述对象的评价提取出特定特征的评价词、情感词组成的对或特定特征的评价分类标记、情感词组成的对，其中对于含义类似的评价词给予相同的评价分类标记；特征的上下文线索，即在从用户对所述对象的评价中特定特征附近满足预定条件的形容词和 / 或名词和 / 或名词短语和 / 或否定词；特征的可用标签，即所述网站和其它网站赋予特定特征的分组标签信息。

[0015] 可选地，所构建的多维向量还包括如下维度中的至少一个：特征的名称；特征的内部线索，其中特征的内部线索包括特定特征的关键词和特定特征的构成词，其中所述关键词和构成词都构成所述多维向量中的维度。

[0016] 可选地，针对特定特征将其多维向量与提取出的其它特征的多维向量进行相似度比较包括：将提取出的所有特征的多维向量的维度进行对准，其中对于特定特征，将提取出的其它特征的多维向量中的维度也视为该特定特征的维度，从而使每个特征的多维向量中的维度数相同；计算所述语料库中特定特征与其多维向量的每一维度之间的互信息作为每一维度的权重；根据所述权重计算各特征的多维向量之间的相似度。

[0017] 可选地，计算所述语料库中所述特征与其多维向量的每一维度之间的互信息作为每一维度的权重包括：

[0018] 对于由于将提取出的其它特征的多维向量中的维度视为特定特征的维度导致的增加维度，权重视为 0，否则利用如下公式计算特定特征与其特定维度的互信息作为权重：

[0019] $$\mathrm{Weight}(f_j ; p_f) = \log_2 \frac{P(f_j, p_f)}{P(f_j)P(p_f)}$$

[0020] 其中 p_f 表示所述特定特征，

[0021] f_j 表示所述特定特征的第 j 个维度，

[0022] $P(f_j;p_f)$ 是在所述语料库中所述特定特征与第 j 个维度在一句话中同时出现的概率，

[0023] $P(f_j)$ 是在所述语料库中第 j 个维度在一句话中出现的概率，以及

[0024] $P(p_f)$ 是在所述语料库中所述特定特征在一句话中出现的概率。

[0025] 可选地，利用欧式距离计算所述各特征的多维向量之间的相似度：

[0026] $$dis\tan ce(v_i, v_j) = \sqrt{\sum_{k=1}^{k=n}(w(f_k(v_i)) - w(f_k(v_j)))^2}$$

[0027] 其中，v_i 表示第 i 个特征的多维向量，v_j 表示第 j 个特征的多维向量；

[0028] $w(f_k(v_i))$ 表示多维向量 v_i 的第 k 个维度的权重，

[0029] $w(f_k(v_j))$ 表示多维向量 v_j 的第 k 个维度的权重。

[0030] 可选地，利用余弦相似度计算所述各特征的多维向量之间的相似度：

[0031] $$\sin(v_i, v_j) = \cos\theta = \frac{\sum_{k=1}^{k=n} w(f_k(v_i)) \times w(f_k(v_j))}{\sqrt{(\sum_{k=1}^{k=n} w^2(f_k(v_i)))(\sum_{k=1}^{k=n} w^2(f_k(v_j)))}}$$

[0032] 其中，v_i 表示第 i 个特征的多维向量，v_j 表示第 j 个特征的多维向量；

[0033] $w(f_k(v_i))$ 表示多维向量 v_i 的第 k 个维度的权重，

[0034] $w(f_k(v_j))$ 表示多维向量 v_j 的第 k 个维度的权重。

[0035] 可选地，将相似度高于预定阈值的特征确定为相同特征并且将与所述网站上的对象的特征不同的特征和其相应属性补充到该网站包括：将相似度高于预定阈值的特征分组到同一组；判断所述网站上已有的对象的特征是否属于一个分组成的组，识别出不包含所述网站上已有对象的任何特征的特征组，将该特征组中的特征和其相应属性补充到该网站。

[0036] 可选地，将该组的特征和其相应属性补充到该网站包括：统计该特征组的各特征在语料库中的出现次数，将出现次数最高的特征的名称和其相应属性补充到该网站。

[0037] 根据本发明的一方面，提供了一种用于丰富网站内容的装置，所述装置包括：提取单元，被配置为从所述网站和其它网站获得语料库，从所述语料库中提取对象的特征，其中所述语料库包括关于所述对象的说明和用户对所述对象的评价；特征向量构建单元，被配置为根据所述语料库，为提取出的特征构建多维向量；向量比较单元，被配置为针对特定特征，将其多维向量与提取出的其它特征的多维向量进行相似度比较；补充单元，被配置为将相似度高于预定阈值的特征确定为相同特征，并且将与所述网站上的对象的特征不同的特征和其相应属性补充到该网站。

[0038] 可选地，所述提取单元进一步被配置为：指定所述其它网站；分析所述网站和其它网站的格式；按照分析出的所述网站和其它网站的格式，寻找含有与所述对象对应的对象标识的所有的块；根据块的格式判断寻找到的块是关于对象的说明还是用户对所述对象的评价，将寻找到的关于对象的说明和用户对所述对象的评价作为语料库。

[0039] 可选地，所述提取单元进一步被配置为：从所述网站和其它网站中关于对象的说明中提取特征种子，其中按照所述网站和其它网站中关于对象的说明的格式，从相应字段中提取特征种子；按照提取出的特征种子，从用户对所述对象的评价提取附加特征。

[0040] 可选地，从用户对所述对象的评价提取附加特征包括：从用户对所述对象的评价中提取出所述特征种子附近满足预定条件的名词作为附加特征；从用户对所述对象的评价中提取出包含所述特征种子的名词词组作为附加特征；如果提取出的附加特征不在特征种子的列表中，将提取出的附加特征加入到特征种子的列表；迭代地重复上述步骤，直到不在特征种子的列表中的新提取出的附加特征的数目低于预定阈值为止。

[0041] 可选地，满足预定条件的名词是指在特征种子附近的预定范围内的出现频率最高的前 n 名的名词，n 为自然数。

[0042] 可选地，所构建的多维向量至少包括以下维度中的一个或多个：特征的情感线索，包括从所述用户对所述对象的评价提取出特定特征的评价词、情感词组成的对或特定特征的评价分类标记、情感词组成的对，其中对于含义类似的评价词给予相同的评价分类标记；特征的上下文线索，即在从用户对所述对象的评价中特定特征附近满足预定条件的形容词和 / 或名词和 / 或名词短语和 / 或否定词；特征的可用标签，即所述网站和其它网站赋予特定特征的分组标签信息。

[0043] 可选地，所构建的多维向量还包括如下维度中的至少一个：特征的名称；特征的内部线索，其中特征的内部线索包括特定特征的关键词和特定特征的构成词，其中所述关键词和构成词都构成所述多维向量中的维度。

[0044] 可选地，向量比较单元进一步被配置为：将提取出的所有特征的多维向量的维度进行对准，其中对于特定特征，将提取出的其它特征的多维向量中的维度也视为该特定特征的维度，从而使每个特征的多维向量中的维度数相同；计算所述语料库中特定特征与其多维向量的每一维度之间的互信息作为每一维度的权重；根据所述权重计算各特征的多维向量之间的相似度。

[0045] 可选地，计算所述语料库中所述特征与其多维向量的每一维度之间的互信息作为每一维度的权重包括：

[0046] 对于由于将提取出的其它特征的多维向量中的维度视为特定特征的维度导致的增加维度，权重视为 0，否则利用如下公式计算特定特征与其特定维度的互信息作为权重：

[0047] $$\mathrm{Weight}(f_j ; p_f) = \log_2 \frac{P(f_j, p_f)}{P(f_j)P(p_f)}$$

[0048] 其中 p_f 表示所述特定特征，

[0049] f_j 表示所述特定特征的第 j 个维度，

[0050] $P(f_j;p_f)$ 是在所述语料库中所述特定特征与第 j 个维度在一句话中同时出现的概率，

[0051] $P(f_j)$ 是在所述语料库中第 j 个维度在一句话中出现的概率，以及

[0052] $P(p_f)$ 是在所述语料库中所述特定特征在一句话中出现的概率。

[0053] 可选地，利用欧式距离计算所述各特征的多维向量之间的相似度：

[0054] $$dis\tan ce(v_i, v_j) = \sqrt{\sum_{k=1}^{k=n} (w(f_k(v_i)) - w(f_k(v_j)))^2}$$

[0055] 其中，v_i 表示第 i 个特征的多维向量，v_j 表示第 j 个特征的多维向量；

[0056] $w(f_k(v_i))$ 表示多维向量 v_i 的第 k 个维度的权重，

[0057] $w(f_k(v_j))$ 表示多维向量 v_j 的第 k 个维度的权重。

[0058] 可选地，利用余弦相似度计算所述各特征的多维向量之间的相似度：

[0059] $$\sin(v_i, v_j) = \cos\theta = \frac{\sum_{k=1}^{k=n} w(f_k(v_i)) \times w(f_k(v_j))}{\sqrt{(\sum_{k=1}^{k=n} w^2(f_k(v_i)))(\sum_{k=1}^{k=n} w^2(f_k(v_j)))}}$$

[0060] 其中，v_i 表示第 i 个特征的多维向量，v_j 表示第 j 个特征的多维向量；

[0061] $w(f_k(v_i))$ 表示多维向量 v_i 的第 k 个维度的权重，

[0062] $w(f_k(v_j))$ 表示多维向量 v_j 的第 k 个维度的权重。

[0063] 可选地，补充单元进一步被配置为：将相似度高于预定阈值的特征分组到同一组；判断所述网站上已有的对象的特征是否属于一个分组成的组，识别出不包含所述网站上已有对象的任何特征的特征组，将该特征组中的特征和其相应属性补充到该网站。

[0064] 可选地，将该组的特征和其相应属性补充到该网站包括：统计该特征组的各特征在语料库中的出现次数，将出现次数最高的特征的名称和其相应属性补充到该网站。

[0065] 根据本发明的一方面，提供了一种包括如上所述的用于丰富网站内容的装置的网站服务器，

[0066] 本发明实现的一个有益效果是用其它网站上关于对象的说明来丰富网站上关于对象的说明而不引入重复信息，提高网页整合效率。

[0067] 本发明还实现了提高在网页整合时的网页信息利用率的有益效果。

附图说明

[0068] 通过结合附图对本公开示例性实施方式进行更详细的描述，本公开的上述以及其它目的、特征和优势将变得更加明显，其中，在本公开示例性实施方式中，相同的参考标号通常代表相同部件。

[0069] 图 1 示出了适于用来实现本发明实施方式的示例性计算系统 100 的框图。

[0070] 图 2 示例性地示出了根据本公开实施例的方法的流程图。

[0071] 图 3 示例性地示出了根据本公开实施例的提取特征的处理的流程图。

[0072] 图 4 示例性地示出了一个示例性网站上对于对象的说明。

[0073] 图 5 示例性地示出了根据本公开一个实施例的装置的框图。

[0074] 图 6 示出了根据本公开一个实施例的补充内容后的网页的示意图。

具体实施方式

[0075] 下面将参照附图更详细地描述本公开的优选实施方式。虽然附图中显示了本公开的优选实施方式，然而应该理解，可以以各种形式实现本公开而不应被这里阐述的实施方式所限制。相反，提供这些实施方式是为了使本公开更加透彻和完整，并且能够将本公开的范围完整的传达给本领域的技术人员。

[0076] 图 1 示出了适于用来实现本发明实施方式的示例性计算系统 100 的框图。如图 1 所示，计算机系统 100 可以包括：CPU（中央处理单元）101、RAM（随机存取存储器）102、ROM

(只读存储器)103、系统总线 104、硬盘控制器 105、键盘控制器 106、串行接口控制器 107、并行接口控制器 108、显示控制器 109、硬盘 110、键盘 111、串行外部设备 112、并行外部设备 113 和显示器 114。在这些设备中,与系统总线 104 耦合的有 CPU 101、RAM 102、ROM 103、硬盘控制器 105、键盘控制器 106、串行控制器 107、并行控制器 108 和显示控制器 109。硬盘 110 与硬盘控制器 105 耦合,键盘 111 与键盘控制器 106 耦合,串行外部设备 112 与串行接口控制器 107 耦合,并行外部设备 113 与并行接口控制器 108 耦合,以及显示器 114 与显示控制器 109 耦合。应当理解,图 1 所述的结构框图仅仅是为了示例的目的,而不是对本发明范围的限制。在某些情况下,可以根据具体情况增加或减少某些设备。

[0077] 所属技术领域的技术人员知道,本发明可以实现为系统、方法或计算机程序产品。因此,本公开可以具体实现为以下形式,即 :可以是完全的硬件、也可以是完全的软件(包括固件、驻留软件、微代码等),还可以是硬件和软件结合的形式,本文一般称为“电路”、“模块”或“系统”。此外,在一些实施例中,本发明还可以实现为在一个或多个计算机可读介质中的计算机程序产品的形式,该计算机可读介质中包含计算机可读的程序代码。

[0078] 可以采用一个或多个计算机可读的介质的任意组合。计算机可读介质可以是计算机可读信号介质或者计算机可读存储介质。计算机可读存储介质例如可以是——但不限于——电、磁、光、电磁、红外线、或半导体的系统、装置或器件,或者任意以上的组合。计算机可读存储介质的更具体的例子(非穷举的列表)包括 :具有一个或多个导线的电连接、便携式计算机磁盘、硬盘、随机存取存储器(RAM)、只读存储器 (ROM)、可擦式可编程只读存储器 (EPROM 或闪存)、光纤、便携式紧凑磁盘只读存储器 (CD-ROM)、光存储器件、磁存储器件、或者上述的任意合适的组合。在本文件中,计算机可读存储介质可以是任何包含或存储程序的有形介质,该程序可以被指令执行系统、装置或者器件使用或者与其结合使用。

[0079] 计算机可读的信号介质可以包括在基带中或者作为载波一部分传播的数据信号,其中承载了计算机可读的程序代码。这种传播的数据信号可以采用多种形式,包括——但不限于——电磁信号、光信号或上述的任意合适的组合。计算机可读的信号介质还可以是计算机可读存储介质以外的任何计算机可读介质,该计算机可读介质可以发送、传播或者传输用于由指令执行系统、装置或者器件使用或者与其结合使用的程序。

[0080] 计算机可读介质上包含的程序代码可以用任何适当的介质传输,包括——但不限于——无线、电线、光缆、RF 等等,或者上述的任意合适的组合。

[0081] 可以以一种或多种程序设计语言或其组合来编写用于执行本发明操作的计算机程序代码,所述程序设计语言包括面向对象的程序设计语言—诸如 Java、Smalltalk、C++,还包括常规的过程式程序设计语言—诸如”C”语言或类似的程序设计语言。程序代码可以完全地在用户计算机上执行、部分地在用户计算机上执行、作为一个独立的软件包执行、部分在用户计算机上部分在远程计算机上执行、或者完全在远程计算机或服务器上执行。在涉及远程计算机的情形中,远程计算机可以通过任意种类的网络——包括局域网 (LAN) 或广域网 (WAN)-- 连接到用户计算机,或者,可以连接到外部计算机(例如利用因特网服务提供商来通过因特网连接)。

[0082] 下面将参照本发明实施例的方法、装置(系统)和计算机程序产品的流程图和 / 或框图描述本发明。应当理解,流程图和 / 或框图的每个方框以及流程图和 / 或框图中各方框的组合,都可以由计算机程序指令实现。这些计算机程序指令可以提供给通用计算机、专

用计算机或其它可编程数据处理装置的处理器，从而生产出一种机器，这些计算机程序指令通过计算机或其它可编程数据处理装置执行，产生了实现流程图和 / 或框图中的方框中规定的功能 / 操作的装置。

[0083] 也可以把这些计算机程序指令存储在能使得计算机或其它可编程数据处理装置以特定方式工作的计算机可读介质中，这样，存储在计算机可读介质中的指令就产生出一个包括实现流程图和 / 或框图中的方框中规定的功能 / 操作的指令装置 (instruction means) 的制造品(manufacture)。

[0084] 也可以把计算机程序指令加载到计算机、其它可编程数据处理装置、或其它设备上，使得在计算机、其它可编程数据处理装置或其它设备上执行一系列操作步骤，以产生计算机实现的过程，从而使得在计算机或其它可编程装置上执行的指令能够提供实现流程图和 / 或框图中的方框中规定的功能 / 操作的过程。

[0085] 根据本公开，在一个实施例中，从各个网站抽取关于对象的说明以及用户对于对象的评价，上述说明和评价构成了语料库。从语料库提取该对象的特征。根据语料库中的内容，为每一个特征构建多维向量。并且针对特定特征，将其多维向量与提取出的其它特征的多维向量进行相似度比较，将相似度高于预定阈值的特征分组到同一个组中。将分到同一组中的特征认为是语义相同的。将分组的结果与某一网站上的对象的特征进行比较，将与所述网站上的对象的特征不同的特征和其相应属性补充到网站。

[0086] 以下将参考附图讲述根据本公开的实施例的细节。如图 2 所示，根据本公开的方法始于步骤 S201。在 S201，从网站和其它网站获得语料库，从所述语料库中提取对象的特征，其中语料库包括关于对象的说明和用户对所述对象的评价。关于对象的说明例如网络、存储、显示等多类参数，如图 4 所示。关于用户对对象的评价是一些用户发表的文本片段 (图 4 未示出)，比如“该手机比较注重硬件配置，摄像头的表现更是十分引人关注的部分”，“我喜欢这款手机具有大屏幕、大内存”。如下文将要详细描述的，语料库是判断对象的特征是否相似的资料库。

[0087] 可选地，从所述网站和其它网站获得语料库包括：指定所述其它网站，这种指定是人为的，但也可以通过由计算机查询存储有各种对象可能对应的网站的数据库实现；分析所述网站和其它网站的格式；按照分析出的所述网站和其它网站的格式，寻找含有与所述对象对应的对象标识的所有的块，这可以通过例如使用本领域技术人员所公知的爬虫技术实现，对象标识例如产品的名称、型号、图像数据等中的至少一个；提取所述寻找到的块中的内容作为语料库，其中提取与关于对象的说明对应的格式的块中的内容作为关于对象的说明，提取与用户对所述对象的评价对应的格式的块中的内容作为用户对所述对象的评价。

[0088] 在一个实施例中，获得了语料库之后，提取该对象的特征。对象的特征可以包括如图 4 所示的网络制式、机身内存、屏幕尺寸等内容。参考附图 3 描述特征提取的细节。首先，在步骤 S301，从网站和其它网站中关于对象的说明中提取特征种子。特征种子是处于网站的网页的格式的预定字段中的特征。因此，按照所述网站和其它网站中关于对象的说明的格式，从预定字段提取特征种子。如图 4 所示，从图 4 的网页上提取左边一列的字段中的内容，即网络制式、网络频率、数据业务、浏览器、机身内存、可用空间、储存卡类型、最大存储扩展等，作为特征种子。接着，在步骤 S302 — S305，按照提取出的特征种子，从用户对对象

的评价中提取附加特征。首先，在步骤S302，从用户对对象的评价中提取出位于特征种子附近的、满足预定条件的名词作为附加特征。例如，可以在所有的用户对对象的评价中提取在特征种子附近（例如，在特征种子所处的一个完整语句或固定长度的周围文本片段中）出现频率最高的名词的前 n 名（n 为自然数）提取作为附加种子。关于如何去确定特征种子所处的一个完整语句的起始位置或固定长度的周围文本片段的起始位置，以及如何在一个范围内识别出名词，目前是成熟技术，比如词性分析技术（见 http://en.wikipedia.org/wiki/Part-of-speech_tagging）。

[0089] 接下来，在步骤S303，通过本领域技术人员所公知的语义分析技术，在用户对对象的评价中提取出与种子特征紧邻的名词与所述特征一起构成的词组作为附加特征。例如，通过语义分析发现，特征“屏幕”和与其紧邻的名词“LCD”同时出现，因此将“LCD 屏幕”提取作为附加特征。关于如何确定名词、如何在句中分词，目前是成熟技术。

[0090] 在步骤S304，在提取出附加特征之后，将在 S302 和 S303 中提取的附加特征与当前特征种子列表中的特征种子进行比较，将不与当前特征种子列表中的特征重叠的附加特征加入到该特征种子列表。在步骤S305，判断在 S302 和 S303 中提取的附加特征的数目是否低于预定阈值。如果在 S302 和 S303 中提取的附加特征的数目大于或等于预定阈值，则迭代地重复步骤 S302 至 S304。如果在 S302 和 S303 中提取的附加特征的数目小于预定阈值，则提取特征的处理在步骤 S306 处结束。

[0091] 预定阈值取值越低，则从语料库中提取出的附加特征越多，则可以实现相对更全面的补充，但计算量相对较大。预定阈值取值越高，则从语料库中提取出的附加特征越少，则可能最后补充的数据较少，但计算量相对较小。

[0092] 回到图 2，在提取了语料库和对象的特征之后，方法进行到步骤 S202，即根据语料库，为提取出的每一个特征构建多维向量。在一个实施例中，一个特征的多维向量可以用下式表示：

[0093] $v(feature)=(f_1,w_1;f_2,w_2;\cdots f_n,w_n)$ （1）

[0094] 其中 v（feature）表示某一特征的多维向量，f_1-f_n 表示该特征的 n 个维度，w_1-w_n 表示 n 个维度对应的权重。

[0095] 在一个实施例中，所构建的多维向量至少包括以下中的一个或多个：特征的情感线索；特征的上下文线索；以及特征的可用标签。所构建的多维向量还可选地包括以下中的至少一个：特征的名称；特征的内部线索。

[0096] 在一个实施例中，用某一特征与所述特征的多维向量的每一维度之间的互信息作为每一维度的权重，互信息可以用如下公式计算：

[0097] $$\mathrm{Weight}(f_j;p_f)=\log_2\frac{P(f_j,p_f)}{P(f_j)P(p_f)} \qquad (2)$$

[0098] p_f 表示所述特定特征。

[0099] f_j 表示所述特征的第 j 个维度。

[0100] $P(f_j;p_f)$ 是在所述语料库中所述特定特征与第 j 个维度在一句话中同时出现的概率。这里的语料库是指步骤 S201 中收集上来的所有语料库。例如所有语料库中有 1000 句话，其中 3 句话中特征“显示屏”与维度“大”同时出现，则该概率为 0.3%。

[0101] $P(f_j)$ 是在所述语料库中第 j 个维度在一句话中出现的概率。例如语料库中有

1000 句话，其中 30 句出现了“大”，则该概率为 3%。

[0102] $P(p_f)$ 是在所述语料库中所述特征在一句话中出现的概率。例如语料库中有 1000 句话，其中 100 句出现了“显示屏”，则该概率为 10%。

[0103] 为了清楚起见，以下以“LCD 屏幕”作为对象的一个特征，示例性地说明如何为该特征构建多维向量 v（LCD 屏幕）。

[0104] 1）特征名称

[0105] 以特征“LCD 屏幕”为例，其特征名称即 LCD 屏幕。

[0106] 2）特征的内部线索

[0107] 特征的内部线索包括特征的关键词和特征的构成词，其中所述关键词和构成词都构成所述多维向量中的维度。仍然以“LCD 屏幕”为例，利用本领域技术人员公知的自然语言处理技术，比如句子成分解析技术可以识别并分析出，其中“屏幕”为该特征的关键词，而“LCD”是该特征的构成词。

[0108] 3）特征的情感线索

[0109] 根据一个实施例，为每个特征使用情感线索以提升语义识别的准确度。更具体地，特征的情感线索包括从用户对对象的评价提取出的评价词和情感词，其中评价词和情感词一起构成所述多维向量中的一个维度。

[0110] 通过语义分析工具，可以从用户对对象的评价中提取对特征的评价词和情感词。举例来说，用户对于某款手机的 LCD 屏幕的评价可能是这样的句子：

[0111] “我的新的 iPad 非常好，因为它的 LCD 屏幕非常大且清晰”。

[0112] 一般来说，通过语义分析工具，寻找“LCD 屏幕”前面的定语和后面的表语，即可获得与特征“LCD 屏幕”有关的评价词——“非常大”、“清晰”。关于情感词提取的有关内容，可以参考 In Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, 2005, 339-346. 中的“Opinion Mining:Extracting Product Feature Assessments from Reviews”，作者为 Ana-Maria Popescu 和 Oren Etzioni，其全文通过引用的方式并入本文中。进一步地，可以通过分析语义，判断此处的评价词“非常大”、“清晰”的情感，也即，是褒义的还是贬义的。例如，此处判断出清晰用于评价 LCD 屏幕时是褒义的，因此给予其情感词“positive”。同样是“非常大”，对于屏幕来说就是褒义的，因此给予其情感词“positive”；对于手机尺寸来说就是贬义的，因此给予其情感词“negative”。

[0113] 因此，找到的一个维度为评价词为“非常大”且情感词为“positive”的整体，另一个维度为评价词为“清晰”且情感词为“positive”二者。

[0114] 在一个实施例中，可以进一步使用评价分类标记减少特征的维度，从而降低计算量。例如，在用户的评价中，当评价屏幕时，使用“清晰”和“清楚”实际上表达了相似的含义。此时，当执行语义分析之后，为“清晰”和“清楚”分配相同的评价分类标记。由此，在计算时，具有褒义情感时的清晰和清楚将被视为同一个维度。由此可见，使用评价分类标记可以减少多维向量的维度，进而降低计算的复杂性。这种情况下，评价分类标记和情感词就一起构成所述多维向量中的一个维度。

[0115] 评价分类标记的分配可以通过查表来进行。例如，维护一个特征、评价词和评价分类标记的对应关系表，其中其中为同义词分配同样的评价分类标记。在使用时，通过按照特

征、评价词，查找对应的评价分类标记。

[0116] 4）特征的上下文线索

[0117] 特征的上下文线索包括所述特征附近满足预定条件的形容词和 / 或名词和 / 或名词短语和 / 或否定词，其中满足预定条件的形容词和 / 或名词和 / 或名词短语和 / 或否定词中的每一个均构成所述多维向量的一个维度。预定条件例如是左右最近的 n 个。

[0118] 以如下示例性的用户评价为例，设 n=3，即寻找左边和右边离特征最近的 3 个形容词和 / 或名词和 / 或名词短语和 / 或否定词：

[0119] “我的新的 iPad 非常好，因为它的 LCD 屏幕非常大且清晰。”

[0120] 以特征“LCD 屏幕”为分界，分别提取“LCD 屏幕”左侧和右侧的 3 个形容词、3 个名词、3 个名词短语和 3 个否定词。参考上述例子，可以得出：

[0121] “LCD 屏幕”左侧的形容词为“好”、“新的”；

[0122] “LCD 屏幕”右侧的形容词为“大”、“清晰”；

[0123] “LCD 屏幕”左侧的名词为“iPad”；

[0124] “LCD 屏幕”右侧的名词为空；

[0125] “LCD 屏幕”左侧的否定词为空；

[0126] “LCD 屏幕”右侧的否定词为空。

[0127] 在该例中，左边和右边的形容词、名词、否定词的个数均低于 3 个。

[0128] 由此，有维度 OL（好）、OL（新的）、OR（大）、OR（清楚）、NL（ipad），将这些内容作为 5 个维度增加到每个特征的多维向量中。

[0129] 5）特征的可用标签

[0130] 特征的可用标签包括当前网站和其它网站赋予特征的分组标签信息，该可用标签构成所述多维向量中的一个维度。如图 4 所示，某些网站已经将特征进行了分组，并且给予了每个分组一个标签，例如图 4 中的“网络”、“存储”等。在这种情况下，该标签实际上可能代表某个特征的共性。因此，本公开考虑使用该标签作为分组的判断依据之一。

[0131] 上面的过程产生一个问题，即由于为每个特征找到的特征的情感线索中包含的评价词、情感词对的数目不同，为每个特征找到的特征的上下文线索中包含的形容词和 / 或名词和 / 或名词短语和 / 或否定词的数目不同等，最后为每个特征确定的多维向量的维度数也不同。为了计算各特征向量之间的相似度，首先要统一各特征向量的维度数。一种方案是，将提取出的所有特征的多维向量的维度进行对准，其中对于特定特征，将提取出的其它特征的多维向量中的维度也视为该特定特征的维度，从而使每个特征的多维向量中的维度数相同。但在计算互信息作为权重时，对于由于将提取出的其它特征的多维向量中的维度视为特定特征的维度导致的增加维度，权重视为 0。

[0132] 例如，从语料库中只提取出所述对象的两个特征“LCD 屏幕”、“显示器”。对于“LCD 屏幕”，特征名称为“LCD 屏幕”，特征的内部线索为“屏幕”、“LCD”，特征的情感线索为“非常大、positive”和“清晰、positive”，特征的上下文线索为“OL（好）”、“OL（新的）”、“OR（大）”、“OR（清楚）”、“NL（ipad）”，特征的可用标签为“显示”。对于“显示器”，特征名称为“显示器”，特征的内部线索为“显示器”，特征的情感线索为“清晰、positive”，特征的上下文线索为“OL（平的）”、“OL（新的）”、“OR（大）”、“OR（清楚）”、“NL（ipad）”，特征的可用标签为“显示”。因此，为“LCD 屏幕”构建的多维向量中的维度不但有“LCD 屏幕”、“屏幕”、

“LCD”、“非常大、positive”、“清晰、positive”、“OL（好）”、“OL（新的）”、“OR（大）”、“OR（清楚）”、“NL（ipad）”、“显示”，还有“显示器”、“OL（平的）”，但对于“显示器”、“OL（平的）”来说，在计算特征向量的相似度时，权重视为 0。

[0133] 以下将讲述多维向量对准之后的相似度计算。如上文所述，在为每个特征构建了多维向量之后，继续进行到步骤 S203。当为每个特征构建了多维向量之后，可以利用多种方法计算两个多维向量之间的相似度。例如，可以根据欧式距离或者余弦相似度来计算两个多维向量之间的相似度。

[0134] 欧式距离的公式为：

[0135] $$dis\tan ce(v_i,v_j)=\sqrt{\sum_{k=1}^{k=n}(w(f_k(v_i))-w(f_k(v_j)))^2} \qquad (3)$$

[0136] 余弦相似度的公式为：

[0137] $$\sin(v_i,v_j)=\cos\theta=\frac{\sum_{k=1}^{k=n}w(f_k(v_i))\times w(f_k(v_j))}{\sqrt{(\sum_{k=1}^{k=n}w^2(f_k(v_i)))(\sum_{k=1}^{k=n}w^2(f_k(v_j)))}} \qquad (4)$$

[0138] 其中 v_i 表示第 i 个多维向量，v_j 表示第 j 个多维向量；

[0139] $w(f_k(v_i))$ 表示多维向量 v_i 的第 k 个维度的权重；类似地，$w(f_k(v_j))$ 表示多维向量 v_j 的第 k 个维度的权重。

[0140] 欧式距离表示两个数据点之间的距离，该距离越小表明两个数据点之间的距离越近；而余弦相似度则表示两个向量之间的夹角，两个向量之间的夹角越小（其取值越接近 1），则两个向量之间的相似度越高。

[0141] 如上文所述，由于维度对准后各特征的维度向量中维度的数目以及顺序是相同的，因此，利用两个特征的各维度向量中的各维度的权重计算两个特征向量之间的相似度。值得注意的是，在多个特征中两两计算相似度。

[0142] 在步骤 S204，将相似度高于预定阈值的特征确定为相同特征，并且将与所述网站上的对象的特征不同的特征和其相应属性补充到网站。具体地，将相似度高于预定阈值的特征分组到同一组；判断所述网站上已有的对象的特征是否属于一个分组成的组，识别出不包含所述网站上已有对象的任何特征的特征组，将该特征组中的特征和其相应属性补充到该网站。

[0143] 例如，总共具有 n 个特征，则首先计算第一个特征的特征向量与其他 n-1 个特征的特征向量的相似度，将相似度高于预定阈值的那些特征分到一个组。例如，相似度高于预定阈值的，共有 m 个特征。然后这 m 个特征就不再参加比较。对于剩下的 n-m 个特征，再取出第一个特征，将其与 n-m 个特征中余下的 n-m-1 个特征进行特征向量的相似度比较。

[0144] 在一个实施例中，可以同时利用公式（3）和（4）计算多维向量之间的相似度，并且综合二者计算的结果来判断相似度。

[0145] 经过分组，例如“LCD 屏幕”、“LCD 显示器”和“显示屏”的特征将会被归类到相同的组中，表示这些特征实际上指代了相同的内容。

[0146] 此时，判断待补充内容的当前网站上的各特征属于归类的组中的哪一组。由于在提取语料库时已经提取了当前网站上的特征，因此，这些特征必然会经历上述分组处理并且被包括在某一组中。通过查找可以容易地定位这些特征归属于哪一组。对于包括当前网站上的特征的特征组将被排除在补充内容之外。换言之，对于不包含当前网站上已有的对

象的任何特征的新组，仅将这些组的特征和其相应属性补充到该网站。

[0147] 在补充时，在某一个分组中的特征名称可能具有较大的差异。举例来说，某个分组中可能包括特征“LCD 屏幕”、“LCD 显示器”和“显示屏”，它们都指代相同的内容。此时，在语料库中进行统计，统计特征组的各特征在语料库中的出现次数，将出现次数最高的特征的名称和其相应属性补充到该网站。

[0148] 例如，在上述特征组中，经过统计发现“显示器”在语料库中出现的次数最高，因此在补充到当前网站时使用“显示器”作为该特征的名称。

[0149] 另一方面，当确定了补充特征的名称时，查找该特征是从哪个网站获得的，并且从该网站获得该特征的属性。更具体地，属性可以包括属性词和属性值。例如，对于图 6 的“其它使用时间”来说，其在网站上表示的属性词是“音乐播放时间”、“视频播放时间”、“视频录制时间”、“视频通话时间”，对应的属性值为“54 小时”、“6.5 小时”、“3.8 小时”、“2.6 小时”。将属性词和属性值与特征“其它使用时间”对应地补充到网站。

[0150] 方法随后结束于 S205。

[0151] 值得说明的是，本公开与语言种类无关。尽管参考中文描述的情况描述了本公开，然而，其他语言例如英文也是可以实现的。如上所述，词语提取、情感词分析等都是可以使用其他语言完成的。

[0152] 以下参考附图 5 描述根据本公开的用于丰富网站内容的装置 500，其包括：提取单元 501，被配置为从所述网站和其它网站获得语料库，从所述语料库中提取对象的特征，其中所述语料库包括关于所述对象的说明和用户对所述对象的评价；特征向量构建单元 502，被配置为根据所述语料库，为提取出的特征构建多维向量；向量比较单元 503，被配置为针对特定特征，将其多维向量与提取出的其它特征的多维向量进行相似度比较；补充单元 504，被配置为将相似度高于预定阈值的特征确定为相同特征，并且将与所述网站上的对象的特征不同的特征和其相应属性补充到该网站。

[0153] 可选地，提取单元 501 进一步被配置为：指定所述其它网站；分析所述网站和其它网站的格式；按照分析出的所述网站和其它网站的格式，寻找含有与所述对象对应的对象标识的所有的块；根据块的格式判断寻找到的块是关于对象的说明还是用户对所述对象的评价，将寻找到的关于对象的说明和用户对所述对象的评价作为语料库。

[0154] 可选地，提取单元 501 进一步被配置为：从所述网站和其它网站中关于对象的说明中提取特征种子，其中按照所述网站和其它网站中关于对象的说明的格式，从相应字段中提取特征种子；按照提取出的特征种子，从用户对所述对象的评价提取附加特征。

[0155] 可选地，从用户对所述对象的评价提取附加特征包括：从用户对所述对象的评价中提取出所述特征种子附近满足预定条件的名词作为附加特征；从用户对所述对象的评价中提取出包含所述特征种子的名词词组作为附加特征；如果提取出的附加特征不在特征种子的列表中，将提取出的附加特征加入到特征种子的列表；迭代地重复上述步骤，直到不在特征种子的列表中的新提取出的附加特征的数目低于预定阈值为止。

[0156] 可选地，满足预定条件的名词是指在特征种子附近的预定范围内的出现频率最高的前 n 名的名词，n 为自然数。

[0157] 可选地，所构建的多维向量至少包括以下维度中的一个或多个：特征的情感线索，包括从所述用户对所述对象的评价提取出特定特征的评价词、情感词组成的对或特定特征

的评价分类标记、情感词组成的对，其中对于含义类似的评价词给予相同的评价分类标记；特征的上下文线索，即在从用户对所述对象的评价中特定特征附近满足预定条件的形容词和/或名词和/或名词短语和/或否定词；特征的可用标签，即所述网站和其它网站赋予特定特征的分组标签信息。

[0158] 可选地，所构建的多维向量还包括如下维度中的至少一个：特征的名称；特征的内部线索，其中特征的内部线索包括特定特征的关键词和特定特征的构成词，其中所述关键词和构成词都构成所述多维向量中的维度。

[0159] 可选地，向量比较单元进一步被配置为：将提取出的所有特征的多维向量的维度进行对准，其中对于特定特征，将提取出的其它特征的多维向量中的维度也视为该特定特征的维度，从而使每个特征的多维向量中的维度数相同；计算所述语料库中特定特征与其多维向量的每一维度之间的互信息作为每一维度的权重；根据所述权重计算各特征的多维向量之间的相似度。

[0160] 可选地，计算所述语料库中所述特征与其多维向量的每一维度之间的互信息作为每一维度的权重包括：

[0161] 对于由于将提取出的其它特征的多维向量中的维度视为特定特征的维度导致的增加维度，权重视为0，否则利用如下公式计算特定特征与其特定维度的互信息作为权重：

[0162] $$\mathrm{Weight}(f_j ; p_f) = \log_2 \frac{P(f_j, p_f)}{P(f_j)P(p_f)}$$

[0163] 其中 p_f 表示所述特定特征，

[0164] f_j 表示所述特定特征的第 j 个维度，

[0165] $P(f_j;p_f)$ 是在所述语料库中所述特定特征与第 j 个维度在一句话中同时出现的概率，

[0166] $P(f_j)$ 是在所述语料库中第 j 个维度在一句话中出现的概率，以及

[0167] $P(p_f)$ 是在所述语料库中所述特定特征在一句话中出现的概率。

[0168] 可选地，利用欧式距离计算所述各特征的多维向量之间的相似度：

[0169] $$dis\tan ce(v_i, v_j) = \sqrt{\sum_{k=1}^{k=n} (w(f_k(v_i)) - w(f_k(v_j)))^2}$$

[0170] 其中，v_i 表示第 i 个特征的多维向量，v_j 表示第 j 个特征的多维向量；

[0171] $w(f_k(v_i))$ 表示多维向量 v_i 的第 k 个维度的权重，

[0172] $w(f_k(v_j))$ 表示多维向量 v_j 的第 k 个维度的权重。

[0173] 可选地，利用余弦相似度计算所述各特征的多维向量之间的相似度：

[0174] $$\sin(v_i, v_j) = \cos\theta = \frac{\sum_{k=1}^{k=n} w(f_k(v_i)) \times w(f_k(v_j))}{\sqrt{(\sum_{k=1}^{k=n} w^2(f_k(v_i)))(\sum_{k=1}^{k=n} w^2(f_k(v_j)))}}$$

[0175] 其中，v_i 表示第 i 个特征的多维向量，v_j 表示第 j 个特征的多维向量；

[0176] $w(f_k(v_i))$ 表示多维向量 v_i 的第 k 个维度的权重，

[0177] $w(f_k(v_j))$ 表示多维向量 v_j 的第 k 个维度的权重。

[0178] 可选地，补充单元进一步被配置为：将相似度高于预定阈值的特征分组到同一组；判断所述网站上已有的对象的特征是否属于一个分组成的组，识别出不包含所述网站上已

有对象的任何特征的特征组，将该特征组中的特征和其相应属性补充到该网站。

[0179] 可选地，将该组的特征和其相应属性补充到该网站包括：统计该特征组的各特征在语料库中的出现次数，将出现次数最高的特征的名称和其相应属性补充到该网站。

[0180] 值得说明的是，由于计算机软件和硬件之间是可以相互转换的，例如，一段软件代码可以通过硬件描述语言(例如 Verlog 等)转化为相应的硬件，例如现场可编程门阵列(FPGA)或者被转化为相应的专用芯片。因此，本公开的实施例可以使用软件实现，可以使用硬件实现，也可以固件的形式实现。本公开已经充分地公开了能够实现本公开目的的装置的组成部件以及通过信号的传递公开了各组成部件之间的连接关系，因此，本领域技术人员完全能够理解，此处公开的技术可以以硬件或固件的方式实现。此外，为了简明起见，本文仅仅描述了与实现本公开密切相关的那些步骤、模块，而省略了其他的组成部件。然而，本领域技术人员应当理解，本公开的方法或装置还可以包括除了上述之外的步骤和组成模块。

[0181] 附图 6 示出了根据本公开的实施例的效果。可以看出，经过了补充之后，已经列出了当前网站缺少的若干特征，并且补充了从其他网站提取的许多相当实用的特征，使得用户可以不必搜索其他网站即可获得全面的信息。

[0182] 附图中的流程图和框图显示了根据本发明的多个实施例的系统、方法和计算机程序产品的可能实现的体系架构、功能和操作。在这点上，流程图或框图中的每个方框可以代表一个模块、程序段或代码的一部分，所述模块、程序段或代码的一部分包含一个或多个用于实现规定的逻辑功能的可执行指令。也应当注意，在有些作为替换的实现中，方框中所标注的功能也可以以不同于附图中所标注的顺序发生。例如，两个连续的方框实际上可以基本并行地执行，它们有时也可以按相反的顺序执行，这依所涉及的功能而定。也要注意的是，框图和 / 或流程图中的每个方框、以及框图和 / 或流程图中的方框的组合，可以用执行规定的功能或操作的专用的基于硬件的系统来实现，或者可以用专用硬件与计算机指令的组合来实现。

[0183] 以上已经描述了本发明的各实施例，上述说明是示例性的，并非穷尽性的，并且也不限于所披露的各实施例。在不偏离所说明的各实施例的范围和精神的情况下，对于本技术领域的普通技术人员来说许多修改和变更都是显而易见的。本文中所用术语的选择，旨在最好地解释各实施例的原理、实际应用或对市场中的技术的技术改进，或者使本技术领域的其它普通技术人员能理解本文披露的各实施例。

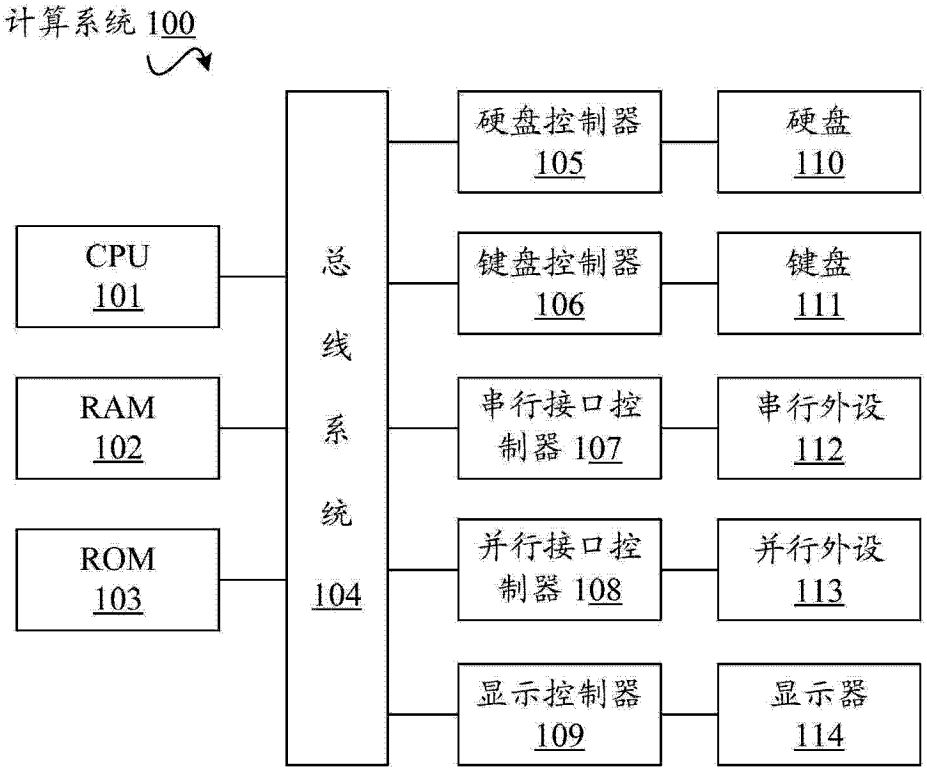

计算系统 100
CPU 101
RAM 102
ROM 103
总线系统 104
硬盘控制器 105
键盘控制器 106
串行接口控制器 107
并行接口控制器 108
显示控制器 109
硬盘 110
键盘 111
串行外设 112
并行外设 113
显示器 114

图 1

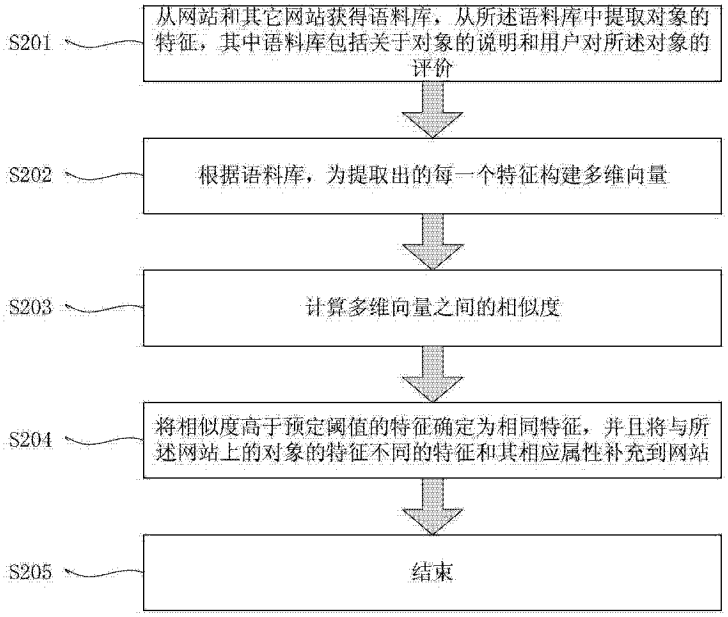

S201
从网站和其它网站获得语料库，从所述语料库中提取对象的特征，其中语料库包括关于对象的说明和用户对所述对象的评价
S202
根据语料库，为提取出的每一个特征构建多维向量
S203
计算多维向量之间的相似度
S204
将相似度高于预定阈值的特征确定为相同特征，并且将与所述网站上的对象的特征不同的特征和其相应属性补充到网站
S205
结束

图 2

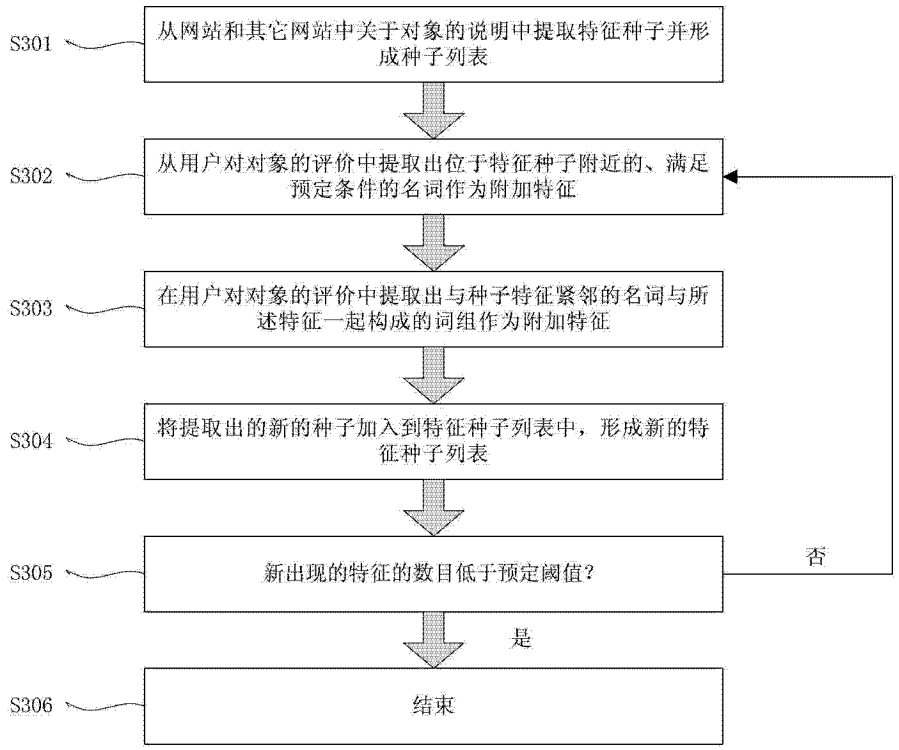

S301
从网站和其它网站中关于对象的说明中提取特征种子并形成种子列表
S302
从用户对对象的评价中提取出位于特征种子附近的、满足预定条件的名词作为附加特征
S303
在用户对对象的评价中提取出与种子特征紧邻的名词与所述特征一起构成的词组作为附加特征
S304
将提取出的新的种子加入到特征种子列表中，形成新的特征种子列表
S305
新出现的特征的数目低于预定阈值？
否
是
S306
结束

图 3

	网络
网络制式	WCDMA/GSM/CDMA2000/CDMA
网络频率	2G：GSM 900/1800/1900 3G：CDMA EVDO 800/1900 3G：WCDMA 900/2100MHz
数据业务	GPRS，CDMA 1X，EVDO rev.A，EDGE，HSDPA
浏览器	支持
	存储
机身内存	4GB ROM+512MB RAM
可用空间	1 GB
储存卡类型	MicroSD卡，支持App2SD功能
最大存储扩展	32GB
	显示
屏幕尺寸	3.8英寸
屏幕色彩	1600万色
屏幕材质	SLCD
分辨率	480*800像素
触摸屏	电容屏
重力感应	支持
光线传感器	支持
距离感应	支持
	娱乐功能
音乐播放	MP3/AAC/AMR/WAV/MID等格式
视频播放	MP4/3GP/AVC/AVI/MPEG-4等格式
电子书	支持

图 4

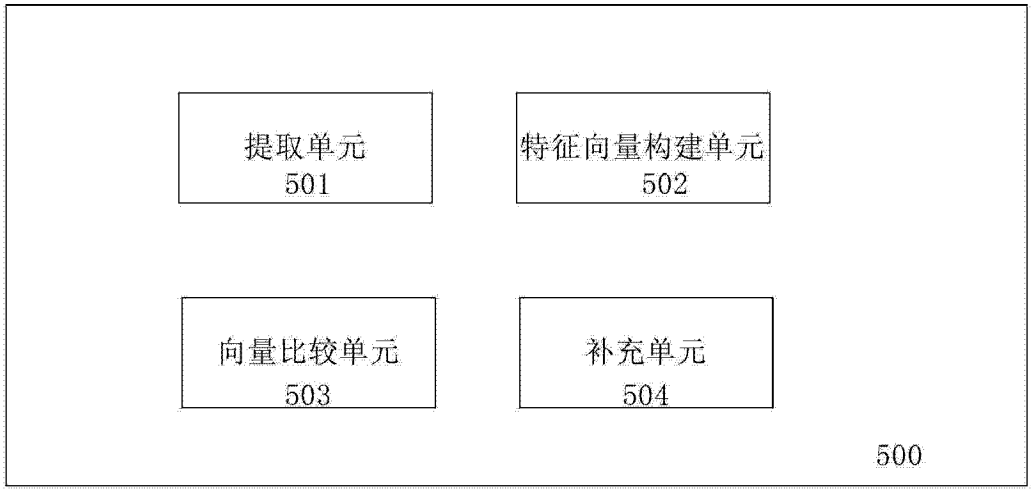


图 5

原始说明

基本参数	
品牌	Nokia/诺基亚
手机价格区间	1001-2000元
上市时间	2010年
网络类型	联通3G GSM/WCDMA
外观样式	直板
主屏尺寸	3.5英寸
摄像头	800万
是否智能手机	智能手机
操作系统	Symbian/塞班
高级功能	WIFI上网, GPS导航,
触摸屏	电容式触摸屏
手机CPU	680M
运行内存RAM	256M
机身内存ROM	8g
键盘类型	虚拟触屏键盘
厚度	普通(大于1cm)
主屏分辨率	640x360像素

补充后的说明

机身尺寸	117.3 x 56.8 x 10.5 毫米
机身重量	130克
网络频率	WCDMA 850/900/1700/1900/2100 GSM 850/900,
屏幕色彩	1600万色
屏幕材质	AMOLED
重力感应	支持
副摄像头	30万像素
传感器类型	CMOS
闪光灯	LED闪光灯
视频拍摄	720p(1280×720，25帧/秒)视频录制
连拍功能	720p(1280×720，25帧/秒)视频录制 720p(1280×720，30帧/秒)视频播放 视频通话的辅助照相/摄像机480p(640×480
电池型号	BL-5K
电池类型	锂电池
电池容量	1200毫安
理论通话时间	576分钟(2G)，318分钟(3G)
理论待机时间	555小时(2G)，656小时(3G)
其它使用时间	音乐播放时间：54小时 视频播放时间：6.5小时 视频录制时间：3.8小时 视频通话时间：2.6小时

图 6