(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 102117289 B
(45)授权公告日 2012.10.10

(21)申请号 200910244539.0

(22)申请日 2009.12.30

(73)专利权人 北京大学
地址 100871 北京市海淀区颐和园路 5 号
专利权人 北大方正集团有限公司
北京方正电子政务信息科技有限公司
北京北大方正电子有限公司

(72)发明人 刘伟 严华梁 万小军 杨建武 肖建国

(74)专利代理机构 北京同达信恒知识产权代理有限公司 11291
代理人 李娟

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

(56)对比文件
CN 101443751 A,2009.05.27, 全文.
CN 101197849 A,2008.06.11, 全文.
US 2008/0243793 A1,2008.10.02, 全文.
CN 101251855 A,2008.08.27, 全文.

审查员 张建

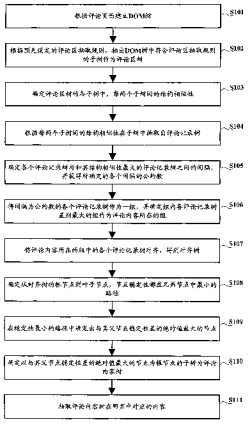
权利要求书 3 页 说明书 8 页 附图 3 页

(54)发明名称

一种从网页中抽取评论内容的方法和装置

(57)摘要

本发明公开了一种从网页中抽取评论内容的方法和装置，涉及信息处理技术，通过建立评论页面的 DOM 树，并选择符合评论区抽取规则的子树抽取出评论区，再利用评论记录间的结构相似性，抽取出评论区中的评论记录，利用包含评论内容的子树的差异性，选择标准差最大的子树作为包含评论内容的子树，最后选取稳定性最小的一条路径中，稳定性差绝对值最大的相邻节点中的孩子节点作为根节点，这个子树就是要抽取的评论内容。由于利用了评论内容的无结构特性来进行抽取，而不是根据网页的模板进行抽取，所以网页的不同不影响抽取的准确性，并且不需要根据网页的模板进行复杂的配置，并通过计算去除了噪声信息，提高了从网页中抽取评论内容的效率和准确性。



CN 102117289 B

1. 一种从网页中抽取评论内容的方法，其特征在于，包括：

根据评论页面建立文档对象模型 DOM 树；

根据预先设定的评论区抽取规则，抽取出 DOM 树中符合所述评论区抽取规则的子树作为评论区树；

在评论区树根节点的各子树中，抽取出评论记录树；

根据每条评论记录中评论内容的差异性，抽取出包括评论内容的评论记录树；

确定出抽取出的包括评论内容的各个评论记录树中的对应节点，并根据每组对应节点所对应的网页内容的差异性，确定出对应的网页内容的差异性满足设定规则的至少一组对应节点；

抽取所述对应的网页内容差异满足设定规则的至少一组对应节点在网页中对应的内容。

2. 如权利要求 1 所述的方法，其特征在于，所述根据预先设定的评论区抽取规则，抽出 DOM 树中符合所述评论区抽取规则的子树作为评论区树，具体包括：

确定出节点所对应的内容所占的面积与页面总面积的比值大于设定的比例的节点；

在节点所对应的内容所占的面积与页面总面积的比值大于设定的比例的节点中，确定所占的面积最小的节点为评论区树的根节点。

3. 如权利要求 2 所述的方法，其特征在于，所述评论区抽取规则还包括：

确定所述评论区树的根点在页面中所占区域的上边界出现在屏幕内；

确定所述评论区树的根节点在页面中所占区域的左边界出现在屏幕内；

确定所述评论区树的根节点在页面中所占区域的长度大于预先设定的长度；

确定所述评论区树的根节点在页面中所占区域的宽度大于预先设定的宽 度。

4. 如权利要求 1 所述的方法，其特征在于，所述在评论区树根节点的各子树中，抽取出评论记录树，具体包括：

确定评论区树的各子树中，每两个子树间的结构相似性；

根据所述每两个子树间的结构相似性在所述子树中抽取出评论记录树。

5. 如权利要求 4 所述的方法，其特征在于，所述确定评论区树的各子树中，每两个子树间的结构相似性，具体包括：

确定评论区树的各子树中，每两个子树中匹配的节点个数为每两个子树间的结构相似性，两个节点匹配具体为两个节点的标签相同和 / 或两个节点对应的文本的字体相同，并且所述两个节点的父节点匹配或者所述两个节点为根节点。

6. 如权利要求 4 所述的方法，其特征在于，所述根据所述每两个子树间的结构相似性在所述子树中抽取出评论记录树，具体包括：

确定每个子树的全局相似性为该子树与其它子树的结构相似性中的最大值；

确定每个子树与其左相邻子树的全局相似性的差值；

确定所述差值最大和最小的两棵子树，并抽取所述差值最小的子树以及所述差值最大和最小的两棵子树之间的各子树作为评论记录树。

7. 如权利要求 4 所述的方法，其特征在于，所述根据每条评论记录中评论内容的差异性，抽取出包括评论内容的评论记录树，具体包括：

确定各个评论记录树与和其结构相似性最大的评论记录树，并根据表征各评论记录相

应区域的评论记录树结构相似性最大的规则，确定构成一条完整评论记录的评论记录树的数量；

根据表征各评论记录相同区域的各评论记录树之间的差异，确定差异最大的一组评论记录树是表征评论记录中评论内容区域的评论记录树；

抽取出各个表征评论记录中评论内容区域的评论记录树作为包括评论内 容的评论记录树。

8. 如权利要求 7 所述的方法，其特征在于，所述根据表征各评论记录相同区域的各评论记录树之间的差异，确定差异最大的一组评论记录树是表征评论记录中评论内容区域的评论记录树，具体包括：

将评论记录树中表征各评论记录相同区域的各评论记录树作为一组；

计算出每组中各个评论记录树的节点数量的标准差，或者计算出每组中各个评论记录树对应的文本长度的标准差；

确定所述标准差最大的组中，各个评论记录树是表征评论记录中评论内容区域的评论记录树。

9. 如权利要求 1 所述的方法，其特征在于，所述确定出抽取出的包括评论内容的各个评论记录树中的对应节点，并根据每组对应节点所对应的网页内容的差异性，确定出对应的网页内容的差异性满足设定规则的至少一组对应节点，具体包括：

将所述抽取出的包括评论内容的各个评论记录树对齐，得到对齐树；

确定从所述对齐树的根节点到叶子节点，节点稳定性都在兄弟节点中最小的路径，所述节点稳定性根据所述对齐树中各个节点的重复次数以及该节点在所述评论内容所在的组中的各个评论记录树中的文本长度确定；

在所述稳定性最小的路径中确定出与其父节点稳定性差的绝对值最大的节点；

确定以所述与其父节点稳定性差的绝对值最大的节点为根节点的子树为评论内容树；

确定所述评论内容树中的各个节点在抽取出的包括评论内容的各个评论记录树中对应的节点，为对应的网页内容的差异性满足设定规则的至少一组对应节点。

10. 如权利要求 9 所述的方法，其特征在于，所述节点稳定性根据所述对齐树中各个节点的重复次数以及该节点在所述评论内容所在的组中的各个评 论记录树中的文本长度确定，具体为：

确定所述对齐树中节点 a 的稳定性为 $S(a)=\sum_{i=1}^{m}-\frac{|Wi|}{|W|}\ln\frac{|Wi|}{|W|}$，其中，m 是节点 a 在所述对齐树中出现的次数，W 是所述节点 a 在构成所述对齐树的所有评论记录树中对应的文本长度的总和，Wi 是所述节点 a 在构成所述对齐树的评论记录树 i 中对应的文本长度。

11. 一种从网页中抽取评论内容的装置，其特征在于，包括：

用于根据评论页面建立文档对象模型 DOM 树的单元；

用于根据预先设定的评论区抽取规则，抽取出 DOM 树中符合所述评论区抽取规则的子树作为评论区树的单元；

用于在评论区树根节点的各子树中，抽取出评论记录树的单元；

用于根据每条评论记录中评论内容的差异性，抽取出包括评论内容的评论记录树的单

元；

用于确定出抽取出的包括评论内容的各个评论记录树中的对应节点，并根据每组对应节点所对应的网页内容的差异性，确定出对应的网页内容的差异性满足设定规则的至少一组对应节点的单元；

用于抽取所述对应的网页内容差异最大的一组对应节点在网页中对应的内容的单元。

12. 如权利要求 11 所述的装置，其特征在于，所述用于根据预先设定的评论区抽取规则，抽出 DOM 树中符合所述评论区抽取规则的子树作为评论区树的单元，具体包括：

用于确定出节点所对应的内容所占的面积与页面总面积的比值大于设定的比例的节点的子单元；

用于在节点所对应的内容所占的面积与页面总面积的比值大于设定的比例的节点中，确定所占的面积最小的节点为评论区树的根节点的子单元。

13. 如权利要求 11 所述的装置，其特征在于，所述用于在评论区树根节 点的各子树中，抽取出评论记录树的单元，具体包括：

用于确定评论区树的各子树中，每两个子树间的结构相似性的子单元；

用于根据所述每两个子树间的结构相似性在所述子树中抽取出评论记录树子单元。

14. 如权利要求 13 所述的装置，其特征在于，所述用于根据每条评论记录中评论内容的差异性，抽取出包括评论内容的评论记录树的单元，具体包括：

用于确定各个评论记录树与和其结构相似性最大的评论记录树，并根据表征各评论记录相应区域的评论记录树结构相似性最大的规则，确定构成一条完整评论记录的评论记录树的数量的子单元；

用于根据表征各评论记录相同区域的各评论记录树之间的差异，确定差异最大的一组评论记录树是表征评论记录中评论内容区域的评论记录树的子单元；

用于抽取出各个表征评论记录中评论内容区域的评论记录树作为包括评论内容的评论记录树的子单元。

15. 如权利要求 11 所述的装置，其特征在于，所述用于确定出抽取出的包括评论内容的各个评论记录树中的对应节点，并根据每组对应节点所对应的网页内容的差异性，确定出对应的网页内容的差异性满足设定规则的至少一组对应节点的单元，具体包括：

用于将所述抽取出的包括评论内容的各个评论记录树对齐，得到对齐树的子单元；

用于确定从所述对齐树的根节点到叶子节点，节点稳定性都在兄弟节点中最小的路径的子单元，所述节点稳定性根据所述对齐树中各个节点的重复次数以及该节点在所述评论内容所在的组中的各个评论记录树中的文本长度确定；

用于在所述稳定性最小的路径中确定出与其父节点稳定性差的绝对值最大的节点的子单元；

用于确定以所述与其父节点稳定性差的绝对值最大的节点为根节点的子 树为评论内容树的子单元；

用于确定所述评论内容树中的各个节点在抽取出的包括评论内容的各个评论记录树中对应的节点，为对应的网页内容的差异性满足设定规则的至少一组对应节点的子单元。

一种从网页中抽取评论内容的方法和装置

技术领域

[0001] 本发明涉及信息处理技术，尤其涉及一种从网页中抽取评论内容的方法和装置。

背景技术

[0002] Web 自上世纪 90 年代初诞生以来便以惊人的速度发展，到目前 Web 已经成为了世界上最大的信息仓库，覆盖了生活中的各个领域，成为了人类工作生活获取信息主要途径之一。在 Web 中，主要是以网页的形式发布信息。然而，Web 中网页的数量十分庞大，目前，Web 中网页的数量已经超过了 5500 亿，显然，在如此庞大的数目下，手工方式的访问已经很难满足人们信息获取的需要，为了让人们更有效地访问和利用 Web 中的信息，自上世纪 90 年代中期开始，研究者们便开始了 Web 信息搜索和集成领域的研究，同时也出现了各种 Web 信息搜索和集成相关的应用，比如垂直搜索引擎、舆情分析等。这些应用实现的一个必要步骤就是将所需的信息从结构化程度很差的网页中准确地抽取出来。

[0003] Web 中的评论是指浏览者在具有可以发布评论的网站中，针对网页的主题所发布的评论，是目前人们在互联网上非常重要的信息获取来源。评论内容在 Web 信息中占有很大的比例。基于评论内容产生了许多重要的应用和研究课题，主要包括以下两个方面 ：

[0004] 评论搜索引擎 ：面向评论的垂直搜索引擎，从网站中获取并集成评论，可以为人们提供即时全面的评论搜索。为了保证评论信息的及时性和全面性，必然要能够对大量的评论页面的及时处理。

[0005] 舆情分析 ：是近十年自然语言处理和信息检索领域的热点研究课题。其目标是从连续的记录中识别出系统未知的话题以及与该话题相关的报道。其主要的信息来源之一就是 Web 中发布的评论信息。

[0006] 由上面对两类应用的介绍可以看出，评论内容是它们非常重要的数据来源。但由于 Web 中网站数量众多，而且评论内容所在的网页通常会包含大量无用的信息，即噪音信息，这些噪音信息必然会严重影响对信息处理的效率和检索的质量。因此，对评论内容的自动抽取是许多重要应用迫切需要解决的关键技术问题之一。

[0007] 由于网页绝大部分都是以 HTML(Hypertext Markup Language 超文本链接标示语言）语言编写，文档结构化程度很低，而且评论本身缺乏语义的连续性，因此使用传统的数据库技术和文本处理技术很难从网页中直接识别所需的信息。Web 中评论内容的抽取一直是 Web 搜索与集成研究领域的热点问题，虽然针对不同的应用场景已经开展了大量的研究工作，但主要是对在网页中以结构化形式展现的数据的抽取，比如 Deep Web 数据的抽取，对于非结构化格式数据抽取的研究至今还没有得到解决，尤其是对无结构化格式的评论内容的抽取。它们的无结构主要表现在三个方面 ：

[0008] 1、评论是由评论者自由撰写的，评论内容在文本长度、信息类型（文本、图片、表格等）往往没有严格的结构和格式要求。

[0009] 2、评论内容信息的表现格式不一致，即表示相同类型语义的信息格式、表现形式多样性化，没有统一的标准，比如评论内容在网页中使用的字体以及在页面中位置繁杂不

一。

[0010] 3、缺乏统一的布局标准，即没有一种对同一类体裁的信息统一的布局标准。

[0011] 由于评论内容的无结构，导致很难定义一种严格的抽取模型实现准确的抽取，给评论内容的抽取带来了极大的挑战性。

[0012] 目前，对评论内容的抽取工作主要存在三个方面的不足：

[0013] 抽取的评论内容不完整：当评论内容较长并且包含多种信息类型时，抽取结果常不能包含所有评论内容；

[0014] 抽取的评论内容中混杂有噪音信息：比如对评论内容的抽取目前的方法主要是在网页层次的抽取，由于评论网页通常包含大量噪音信息，严重影响信息处理的质量；

[0015] 抽取准确性不高且不稳定：由于目前的抽取方法依赖于评论网页的模板，而不同网站之间的评论网页得模板存在着较大的差异，所以抽取的准确性较低，根据网页的不同，抽取的准确性波动也较大，较好的情况也仅在 80％左右。

[0016] 怎样用快速的自动方法，准确抽取出网页中的评论内容是很多重要应用所需要的，但目前尚未发现此类的方法和系统。

发明内容

[0017] 本发明实施例提供一种从网页中抽取评论内容的方法和装置，以提高从网页中抽取评论内容的效率和准确性。

[0018] 一种从网页中抽取评论内容的方法，包括：

[0019] 根据评论页面建立文档对象模型 DOM 树；

[0020] 根据预先设定的评论区抽取规则，抽取出 DOM 树中符合所述评论区抽取规则的子树作为评论区树；

[0021] 在评论区树根节点的各子树中，抽取出评论记录树；

[0022] 根据每条评论记录中评论内容的差异性，抽取出包括评论内容的评论记录树；

[0023] 确定出抽取出的包括评论内容的各个评论记录树中的对应节点，并根据每组对应节点所对应的网页内容的差异性，确定出对应的网页内容的差异性满足设定规则的至少一组对应节点；

[0024] 抽取所述对应的网页内容差异满足设定规则的至少一组对应节点在网页中对应的内容。

[0025] 进一步，所述根据预先设定的评论区抽取规则，抽出 DOM 树中符合所述评论区抽取规则的子树作为评论区树，具体包括：

[0026] 确定出节点所对应的内容所占的面积与页面总面积的比值大于设定的比例的节点；

[0027] 在节点所对应的内容所占的面积与页面总面积的比值大于设定的比例的节点中，确定所占的面积最小的节点为评论区树的根节点。

[0028] 进一步，所述在评论区树根节点的各子树中，抽取出评论记录树，具体包括：

[0029] 确定评论区树的各子树中，每两个子树间的结构相似性；

[0030] 根据所述每两个子树间的结构相似性在所述子树中抽取出评论记录树。

[0031] 进一步，所述根据每条评论记录中评论内容的差异性，抽取出包括评论内容的评

论记录树，具体包括：

[0032] 确定各个评论记录树与和其结构相似性最大的评论记录树，并根据表征各评论记录相应区域的评论记录树结构相似性最大的规则，确定构成一条完整评论记录的评论记录树的数量；

[0033] 根据表征各评论记录相同区域的各评论记录树之间的差异，确定差异最大的一组评论记录树是表征评论记录中评论内容区域的评论记录树；

[0034] 抽取出各个表征评论记录中评论内容区域的评论记录树作为包括评论内容的评论记录树。

[0035] 较佳的，所述确定出抽取出的包括评论内容的各个评论记录树中的对应节点，并根据每组对应节点所对应的网页内容的差异性，确定出对应的网页内容的差异性满足设定规则的至少一组对应节点，具体包括：

[0036] 将所述抽取出的包括评论内容的各个评论记录树对齐，得到对齐树；

[0037] 确定从所述对齐树的根节点到叶子节点，节点稳定性都在兄弟节点中最小的路径，所述节点稳定性根据所述对齐树中各个节点的重复次数以及该节点在所述评论内容所在的组中的各个评论记录树中的文本长度确定；

[0038] 在所述稳定性最小的路径中确定出与其父节点稳定性差的绝对值最大的节点；

[0039] 确定以所述与其父节点稳定性差的绝对值最大的节点为根节点的子树为评论内容树；

[0040] 确定所述评论内容树中的各个节点在抽取出的包括评论内容的各个评论记录树中对应的节点，为对应的网页内容的差异性满足设定规则的至少一组对应节点。

[0041] 一种从网页中抽取评论内容的装置，包括：

[0042] 用于根据评论页面建立文档对象模型 DOM 树的单元；

[0043] 用于根据预先设定的评论区抽取规则，抽取出 DOM 树中符合所述评论区抽取规则的子树作为评论区树的单元；

[0044] 用于在评论区树根节点的各子树中，抽取出评论记录树的单元；

[0045] 用于根据每条评论记录中评论内容的差异性，抽取出包括评论内容的评论记录树的单元；

[0046] 用于确定出抽取出的包括评论内容的各个评论记录树中的对应节点，并根据每组对应节点所对应的网页内容的差异性，确定出对应的网页内容的差异性满足设定规则的至少一组对应节点的单元；

[0047] 用于抽取所述对应的网页内容差异最大的一组对应节点在网页中对应的内容的单元。

[0048] 本发明实施例提供一种从网页中抽取评论内容的方法和装置，通过建立评论页面的 DOM(Document Object Model，文档对象模型）树，并选择符合评论区抽取规则的子树抽取出评论区，再利用评论记录间的结构相似性，抽取出评论区中的评论记录，利用包含评论内容的子树的差异性，选择标准差最大的子树作为包含评论内容的子树，最后选取稳定性最小的一条路径中，稳定性差绝对值最大的相邻节点中的孩子节点作为根节点，这个子树就是要抽取的评论内容。由于利用了评论内容的无结构特性来进行抽取，而不是根据网页的模板进行抽取，所以网页的不同不会影响抽取的准确性，并且不需要根据网页的模板进

行过于复杂的配置，并通过计算去除了噪声信息，提高了从网页中抽取评论内容的效率和准确性。

附图说明

[0049] 图 1 为本发明实施例提供的从网页中抽取评论内容的方法流程图；

[0050] 图 2 为本发明实施例提供的根据评论页面建立的 DOM 树示例；

[0051] 图 3 为本发明实施例提供的从评论区抽取评论记录的示意图；

[0052] 图 4 为本发明实施例提供的获得对齐树的示意图；

[0053] 图 5 为本发明实施例提供的从评论记录中抽取评论内容的示意图。

具体实施方式

[0054] 本发明实施例提供一种从网页中抽取评论内容的方法和装置，根据由评论页面的 DOM 树，利用评论内容的格式、文本长度、字体变化性较大这一特性，通过相似度、标准差、稳定性等指标，来去除噪音信息，筛选出评论内容，进而进行抽取。

[0055] 由于本发明实施例提供的从网页中抽取评论内容的方法不需要依赖网页的模板，仅是利用了评论内容的无结构特性来判断出评论内容所在的位置，所以适用广泛，可以应用在不同的网页中，最多是根据网页的差异修改一下参数，省去了对不同网页的复杂配置，并且使用无结构特性来判断出评论内容所在的位置，准确性高，可以有效的去除噪音信息。

[0056] 本发明实施例提供的从网页中抽取评论内容的方法包括评论区识别、评论记录抽取、评论内容抽取三大部分，如图 1 所示，具体包括：

[0057] 步骤 S101、根据评论页面建立 DOM 树，一种 DOM 树的实例如图 2 所示；

[0058] 在建立 DOM 树后，进行评论区识别：

[0059] 步骤 S102、根据预先设定的评论区抽取规则，抽出 DOM 树中符合评论区抽取规则的子树作为评论区树；

[0060] 在识别出评论区后，在评论区树中进行评论记录的抽取：

[0061] 步骤 S103、确定评论区树的各子树中，每两个子树间的结构相似性；

[0062] 步骤 S104、根据每两个子树间的结构相似性在子树中抽取出评论记录树；

[0063] 在抽取出评论记录树后，再在各个评论记录树中抽取评论内容：

[0064] 首先根据每条评论记录中评论内容的差异性，抽取出包括评论内容的评论记录树：

[0065] 步骤 S105、确定各个评论记录树与和其结构相似性最大的评论记录树之间的间隔，并获得所确定的各个间隔的公约数；

[0066] 步骤 S106、将间隔为公约数的各个评论记录树作为一组，并确定组内各评论记录树差别最大的组作为评论内容所在的组；

[0067] 在确定出抽取出的包括评论内容的各个评论记录树中的对应节点，并根据每组对应节点所对应的网页内容的差异性，确定出对应的网页内容的差异性满足设定规则的至少一组对应节点：

[0068] 步骤 S107、将评论内容所在的组中的各个评论记录树对齐，得到对齐树；

[0069] 步骤 S108、确定从对齐树的根节点到叶子节点，节点稳定性都在兄弟节点中最小

的路径，节点稳定性根据对齐树中各个节点的重复次数以及该节点在评论内容所在的组中的各个评论记录树中的文本长度确定；

[0070] 步骤 S109、在稳定性最小的路径中确定出与其父节点稳定性差的绝对值最大的节点；

[0071] 步骤 S110、确定以与其父节点稳定性差的绝对值最大的节点为根节点的子树为评论内容树，确定评论内容树中的各个节点在抽取出的包括评论内容的各个评论记录树中对应的节点，为对应的网页内容的差异性满足设定规则的至少一组对应节点；

[0072] 步骤 S111、抽取对应的网页内容差异满足设定规则的至少一组对应节点在网页中对应的内容，即抽取评论内容树在网页中对应的内容。

[0073] 这样就筛选出了评论内容，实现了评论内容的抽取。

[0074] 在步骤 S102 中，可以先根据评论区抽取规则库中预先设定的评论区抽取规则来筛选 DOM 树中的节点，例如，评论区抽取规则为：1、节点面积与页面总面积之比大于 0.4；2、节点在页面中所占区域的上边界出现在屏幕内；3、节点在页面中所占区域的左边界出现在屏幕内；4、节点在页面中所占区域的长大于 270 单位；5、节点在页面中所占区域的宽大于 150 单位。这时，将 DOM 树中所有满足上述规则的节点筛选出来，在选择其中占用页面面积最小的节点作为评论区树的根节点，这样就从 DOM 树中抽取出了评论区树。

[0075] 当然根据页面的不同，可以对评论区抽取规则库中的评论区抽取规则进行修改，可以只利用其中的一部分规则进行抽取，也可以修改规则中的参数，比如可以筛选节点面积与页面总面积之比大于 0.5 的节点。

[0076] 在筛选出满足上述规则的节点后，由于评论区树的根节点及其上级节点都满足上述规则，所以需要选择其中占用页面面积最小的节点作为评论区树的根节点，由于评论区树的根节点的上级节点在包括评论区树外还包括其它子树，所以评论区树的根节点的上级节点占用的页面面积必然大于评论区树的根节点占用的页面面积，利用这一点，选择其中占用页面面积最小的节点作为评论区树的根节点，就成功筛选出了评论区树的根节点，进而从 DOM 树中抽取出评论区树。

[0077] 在步骤 S103 中，在确定评论区树的各子树中，每两个子树间的结构相似性时，可以通过确定评论区树的各子树中，每两个子树中匹配的节点个数为每两个子树间的结构相似性，其中，两个节点匹配可以定义为两个节点的标签相同和 / 或两个节点对应的文本的字体相同，同时两个节点的父节点也应该匹配或者这两个节点为根节点。

[0078] 由于每条评论记录中，都包含发表时间、发信人、回复按钮、删除按钮、编号以及评论内容等信息，除少部分评论内容可能采用非默认的字体外，每条评论记录中大部分信息所采用的字体都是一致的，而每个部分的标签必然也是相同的，所以可以使用两个节点的标签相同和 / 或两个节点对应的文本的字体是否相同来确定两个节点是否匹配。

[0079] 在步骤 S104 中，根据每两个子树间的结构相似性在子树中抽取出评论记录树，具体包括：

[0080] 首先确定每个子树的全局相似性，全局相似性是该子树和评论区树的各个子树中的结构相似性的最大值。例如，评论区树的根节点下有 3 棵子树 A、B、C，其中 A 和 B 的结构相似性为 4、A 和 C 的结构相似性为 3、B 和 C 的结构相似性为 5，那么它们的全局相似性分别为 A：4，B：5，C：5。

[0081] 确定每个子树的全局相似性后，再确定每个子树与其左相邻子树的全局相似性的差值，并确定出所得到的差值最大和最小的两棵子树，抽取差值最小的子树以及差值最大和最小的两棵子树之间的各子树作为评论记录树。

[0082] 例如，评论区树的根节点下有 6 棵子树，它们的全局相似性分别为 1、3、4、4、3、1，从第二棵子树开始，与左相邻子树的全局相似性的差为 -2、-1、0、1、2，其中，-2 最小，2 最大，则选取第 2 棵子树到第 5 棵子树为评论记录树。

[0083] 由于每条评论记录的格式都很相似，所以各个评论记录子树之间的结构相似性很大，确定了每个子树的全局相似性之后，各个评论记录子树的全局相似性值都比较大，而评论区树中的其它子树由于结构的差异，其全局相似性值比较小，所以所选择的评论记录子树是全局相似性值比较大的子树，并且由于评论记录树都是相邻的，所以评论记录树之间的差值在 0 左右，而最左边的评论记录树与其左相邻子树的差值为比较大的正值，最右边的评论记录子树的右相邻子树和它的差值很小，为绝对值很大的负值，所以计算出每个子树与其左相邻子树的全局相似性的差值后，选择抽取差值最小的子树以及差值最大和最小的两棵子树之间的各子树作为评论记录树。评论记录抽取示意图如图 3 所示。

[0084] 由于有些评论记录树是表征评论记录中的发表时间、发信人、回复按钮、删除按钮、编号等信息的，而所要抽取的仅仅是评论内容，所以需要超出各个评论记录树中包含评论内容的评论记录树。

[0085] 所以步骤 S105 中确定各个评论记录树与和其结构相似性最大的评论记录树之间的间隔，并获得所确定的各个间隔的公约数，因为表征评论记录中的发表时间、发信人、回复按钮、删除按钮、编号等信息的一棵评论记录树间必然与其它表征评论记录中的发表时间、发信人、回复、删除、编号等信息的评论记录树的结构相似性最大，包含评论内容的评论记录树也必然与其它包含评论记录树的结构相似性最大，所以，确定出各个评论记录树与和其结构相似性最大的评论记录树之间的间隔，并获得所确定的各个间隔的公约数后，即可确知几棵评论记录树构成一条完整的评论记录了。

[0086] 例如，一共有 9 个评论记录树，其中 ：评论记录树 1 和评论记录树 7 结构相似性最大，间隔为 6 ；评论记录树 2 和评论记录树 5 结构相似性最大，间隔为 3 ；评论记录树 3 和评论记录树 9 结构相似性最大，间隔为 6 ；评论记录树 4 和评论记录树 7 结构相似性最大，间隔为 3 ；评论记录树 5 和评论记录树 2 结构相似性最大，间隔为 3 ；评论记录树 6 和评论记录树 9 结构相似性最大，间隔为 3 ；评论记录树 7 和评论记录树 1 结构相似性最大，间隔为 6 ；评论记录树 8 和评论记录树 5 结构相似性最大，间隔为 3 ；评论记录树 9 和评论记录树 6 结构相似性最大，间隔为 3。各个间隔分别为 ：6、3、6、3、3、3、6、3、3，其最大公约数为 3，所以每三棵评论记录树构成一条完整的评论记录。

[0087] 构成一条评论记录的这三棵评论记录树中，有一棵是包含评论内容的评论记录树，步骤 S106 为确定包含评论内容的评论记录树的步骤。

[0088] 由于每条评论记录中发表时间、发信人、回复按钮、删除按钮、编号等信息的格式和文本长度基本是相同的，或者差别很小，所以每条评论记录中表征这些信息的评论记录树的结构和文本长度差别很小，而评论内容的差异性很大，所以将间隔为公约数的各个评论记录树作为一组，组内的各个评论记录数差别最大的组就是评论内容所在的组了。

[0089] 例如，一共有 9 个评论记录树，每 3 棵评论记录树构成一条完整的评论记录，那么

将评论记录树 1、评论记录树 4 和评论记录树 7 放在组一中，将评论记录树 2、评论记录树 5 和评论记录树 8 放在组二中，将评论记录树 3、评论记录树 6 和评论记录树 9 放在组三中，如果组二中的三棵评论记录树差别最大，那么组二中的评论记录树 2、评论记录树 5 和评论记录树 8 就是包括评论内容的评论记录树。

[0090] 确定组内各评论记录树差别的方式有很多，最直接的一种方式就是利用组内各棵树的结构特征做标准差，可以计算组内各棵树包括的节点数的标准差，也可以计算组内各棵树对应的文本长度的标准差，标准差最大的一组就是评论内容所在的组。

[0091] 例如一个组内有 3 棵评论记录树，分别为评论记录树 2、评论记录树 5 和评论记录树 8，评论记录树 2 中包括 5 个节点，评论记录树 5 中包括 4 个节点，评论记录树 8 中包括 9 个节点，平均值为 6，那么按照节点数量计算时，这个组的标准差为：$\sqrt{\frac{(5-6)^2+(4-6)^2+(9-6)^2}{3}}=\sqrt{\frac{14}{3}}$。

[0092] 这几棵评论记录树中也可能包含有格式信息等非评论内容的信息，接下来就需要在评论记录树中的各子树中找出评论内容树。

[0093] 首先在步骤 S107 中，将评论内容所在的组中的各个评论记录树对齐，得到对齐树，即对于一个组内的 3 棵评论记录树，将这三棵评论记录子树中格式相同的节点合并，并记录该节点在对齐树中出现的次数。

[0094] 一种合并为对齐树的示意图如图 4 所示，两棵待合并的评论记录树种，节点 a 和节点 a 的孩子节点 b、d 可以合并，其它的不能合并的节点就直接添加到对齐树中，合并后，节点 a 和节点 b、d 都在对齐树中出现了两次。

[0095] 由于评论内容的文本长度是各不相同的，而其它信息的文本长度通常是相同的，所以可以根据对齐树中各个节点的重复次数以及该节点在评论内容所在的组中的各个评论记录树中的文本长度，可以确定对齐树中各个节点的稳定性，具体的计算方式是：节点 a 的稳定性为$S(a)=\sum_{i=1}^{m}-\frac{|Wi|}{|W|}\ln\frac{|Wi|}{|W|}$，其中，m 是节点 a 在对齐树中出现的次数，W 是节点 a 在构成对齐树的所有评论记录树中对应的文本长度的总和，Wi 是节点 a 在构成对齐树的评论记录树 i 中对应的文本长度。节点 a 在各个评论记录树中的文本长度越均匀，稳定性就大。

[0096] 而由于评论内容的文本长度最不均匀，所以评论内容对应的节点的稳定性很差，确定从对齐树的根节点到叶子节点，节点稳定性都在兄弟节点中最小的路径，并在该路径中确定出与其父节点稳定性差的绝对值最大的节点，确定评论内容树是以与其父节点稳定性差的绝对值最大的节点为根节点的子树，直接抽取评论内容树在网页中对应的内容即可获得所需要的评论内容了。

[0097] 仍以图 4 中的对齐树为例，计算根节点 a 三个孩子节点的稳定性，如果节点 c 最小，则继续计算节点 c 的孩子节点的稳定性，得到路径 acb，计算节点 a 和节点 c 的稳定性的差的绝对值，以及节点 c 和节点 b 的稳定性的差的绝对值，如果前者较大，则节点 c 为评论内容树的根节点，评论内容树就是由节点 c 和节点 b 构成的树。评论内容抽取的示意图如图 5 所示。

[0098] 本发明实施例提供一种从网页中抽取评论内容的方法和装置，通过建立评论页面的 DOM 树，并选择符合评论区抽取规则的子树抽取出评论区，再利用评论记录间的结构相

似性，抽取出评论区中的评论记录，利用包含评论内容的子树的差异性，选择标准差最大的子树作为包含评论内容的子树，最后选取稳定性最小的一条路径中，稳定性差绝对值最大的相邻节点中的孩子节点作为根节点，这个子树就是要抽取的评论内容。由于利用了评论内容的无结构特性来进行抽取，而不是根据网页的模板进行抽取，所以网页的不同不会影响抽取的准确性，并且不需要根据网页的模板进行过于复杂的配置，并通过计算去除了噪声信息，提高了从网页中抽取评论内容的效率和准确性。

[0099] 显然，本领域的技术人员可以对本发明实施例进行各种改动和变型而不脱离本发明的精神和范围。这样，倘若本发明的这些修改和变型属于本发明权利要求及其等同技术的范围之内，则本发明也意图包含这些改动和变型在内。

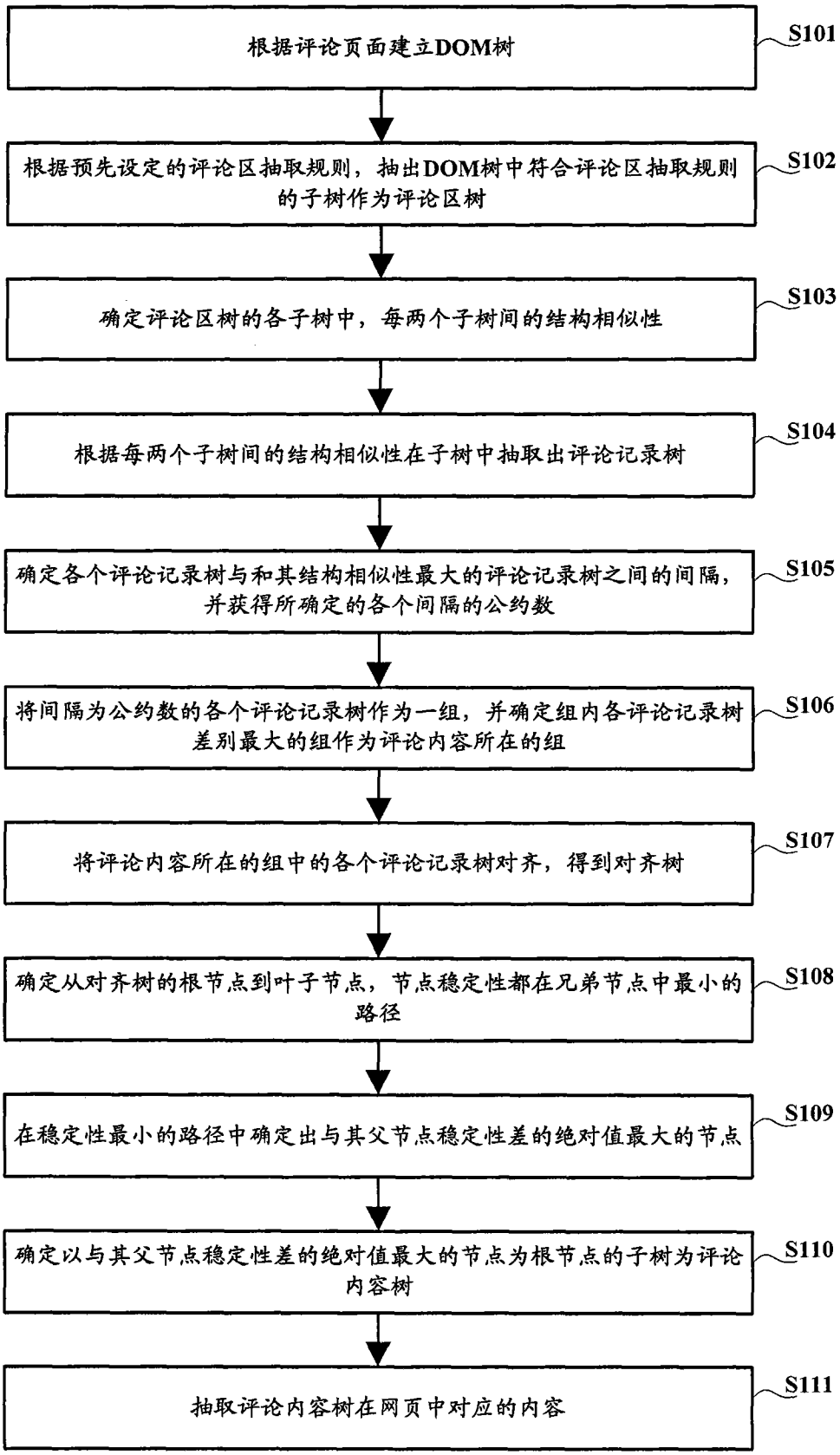

根据评论页面建立DOM树
S101
根据预先设定的评论区抽取规则，抽出DOM树中符合评论区抽取规则的子树作为评论区树
S102
确定评论区树的各子树中，每两个子树间的结构相似性
S103
根据每两个子树间的结构相似性在子树中抽取出评论记录树
S104
确定各个评论记录树与和其结构相似性最大的评论记录树之间的间隔，并获得所确定的各个间隔的公约数
S105
将间隔为公约数的各个评论记录树作为一组，并确定组内各评论记录树差别最大的组作为评论内容所在的组
S106
将评论内容所在的组中的各个评论记录树对齐，得到对齐树
S107
确定从对齐树的根节点到叶子节点，节点稳定性都在兄弟节点中最小的路径
S108
在稳定性最小的路径中确定出与其父节点稳定性差的绝对值最大的节点
S109
确定以与其父节点稳定性差的绝对值最大的节点为根节点的子树为评论内容树
S110
抽取评论内容树在网页中对应的内容
S111

图 1

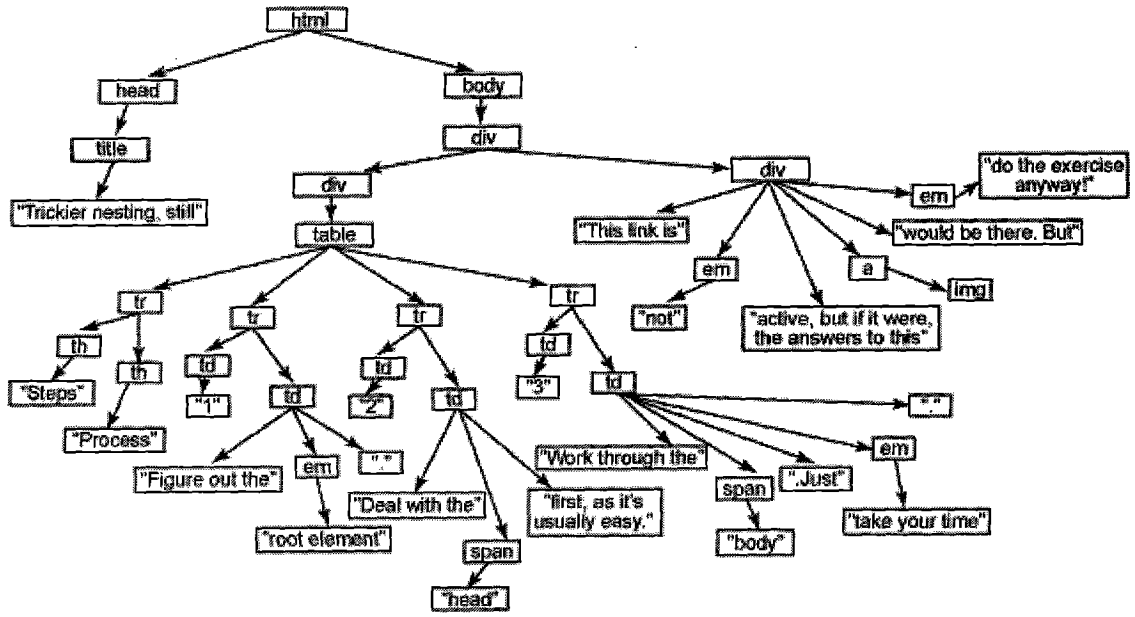

html
head
body
title
div
div
div
"Trickier nesting, still"
table
"This link is"
em
"do the exercise anyway!"
"would be there. But"
em
a
img
"not"
"active, but if it were, the answers to this"
tr
tr
tr
tr
th
th
td
td
td
td
"Steps"
"Process"
"1"
"2"
"3"
"."
"Figure out the"
em
"."
"Deal with the"
"root element"
span
"head"
"first, as it's usually easy."
"Work through the"
span
"body"
".Just"
em
"take your time"

图 2

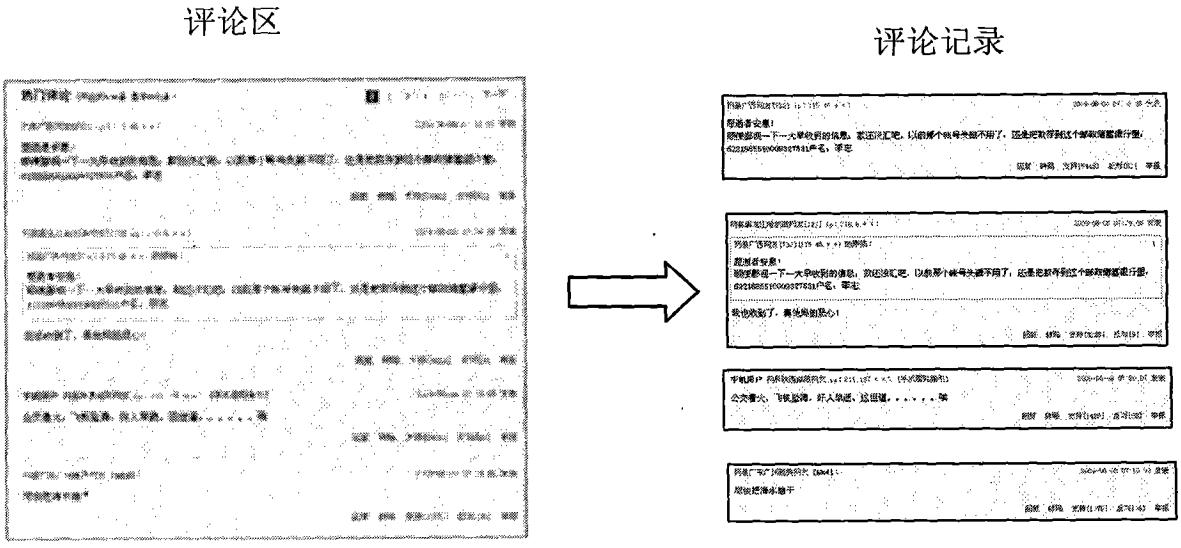

评论区
评论记录

图 3

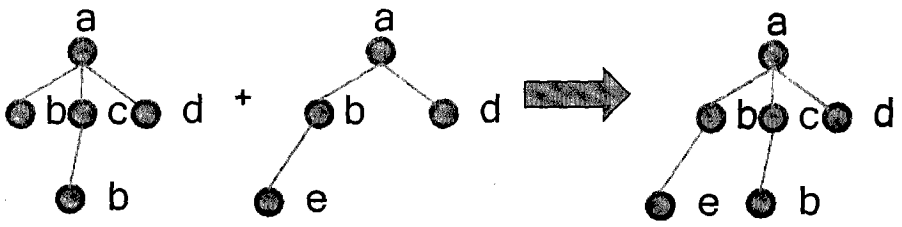

a
b
c
d
b
+
a
b
d
e
a
b
c
d
e
b

图 4

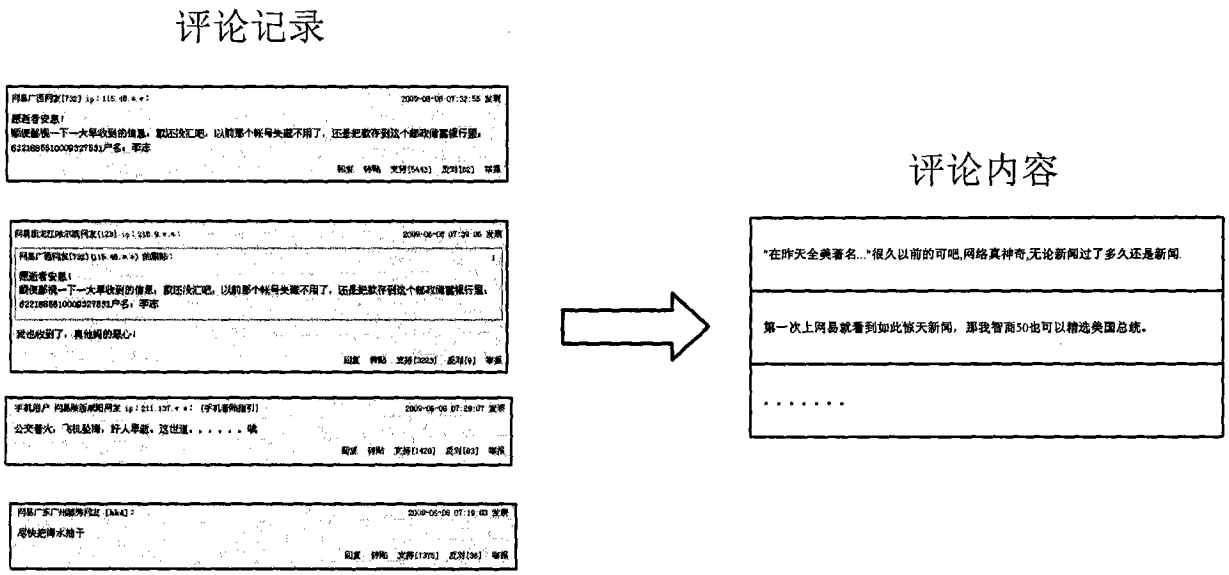

评论记录
评论内容
"在昨天全美著名..."很久以前的可吧,网络真神奇,无论新闻过了多久还是新闻.
第一次上网易就看到如此惊天新闻，那我智商50也可以精选美国总统。
. . . . . .

图 5